

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции
«Диалог» (2022)

Выпуск 21

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2022)

Issue 21

УДК 81'322(063)
ББК 81.1я431
К63

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

К63 **Компьютерная лингвистика** и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 21. Москва: РГГУ, 2022. С. I–614.

ISBN 978-5-7281-3205-9

Сборник включает 54 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2022», представляющих широкий спектр теоретических и прикладных исследований в области компьютерного моделирования естественного языка и создания новых лингвистических технологий.

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

УДК 81'322(063)
ББК 81.1я431

ISBN 978-5-7281-3204-2
ISBN 978-5-7281-3205-9 (осн. том)

© Оформление. Российский государственный гуманитарный университет, 2022

Предисловие

21-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 28-й международной конференции «Диалог». В 2022 году для публикации в ежегоднике редколлегией был отобран 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 конференции:

- Компьютерные лингвистические ресурсы;
- Компьютерный анализ документов (классификация, перевод, поиск, саммаризация, генерация, анализ тональности и аргументации и т. д.);
- Глубокое обучение в компьютерной лингвистике (методики применения, содержательная лингвистическая интерпретация);
- Компьютерный анализ Social Media;
- Корпусная лингвистика и корпусометрия (методики создания, использования и оценки корпусов и датасетов);
- Компьютерная семантика (от аналитических до дистрибуционных моделей);
- Лингвистические онтологии и автоматическое извлечение знаний;
- Мультимодальные подходы к анализу языка (на стыке NLP и Computer Vision);
- Мультиканальная коммуникация (включая лингвистический анализ речи);
- Лингвистические модели общения и диалоговые агенты;
- Лингвистический анализ текста (морфология, синтаксис, семантика);
- Компьютерная лексикография;
- Полевая компьютерная лингвистика (методы создания и переноса языковых моделей для малоресурсных языков).

В соответствии с традициями «Диалога», отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с содержательным 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:

- RUSSE Detox: Соревнование по автоматической детоксификации текстов;
- RuATD: Соревнование по распознаванию сгенерированных текстов;
- RuArg: Соревнование по анализу аргументации;
- RuNNE: Соревнование по извлечению вложенных именованных сущностей в *few-shot* режиме.

В сборник включены наиболее оригинальные работы участников этих тестирований.

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

- как правило, доклады подаются на английском языке. Это расширяет их аудиторию и позволяет привлечь к рецензированию международных экспертов;
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, могут подаваться 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 данный сборник индексируется Scopus.

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

Организаторы

Ежегодная конференция «Диалог» проводится при организационной поддержке компании АBBYY. Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем передачи информации РАН
- Компания АBBYY
- Филологический факультет МГУ
- Школа прикладной математики и информатики МФТИ

Международный программный комитет

Богуславский Игорь Михайлович	ИППИ РАН, Россия; Мадридский политехнический университет, Испания
Буате Кристиан	Университет Джозефа Фурье — Гренобль 1, Франция
Гельбух Александр Феликсович	Национальный политехнический институт, Мексика
Иомдин Леонид Лейбович	ИППИ РАН им. А.А. Харкевича, Россия
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова, Россия
Корбетт Гревил	Университет Суррея, Великобритания
Кронгауз Максим Анисимович	НИУ «Высшая школа экономики», Россия
Лукашевич Наталья Валентиновна	НИВЦ МГУ им. М. В. Ломоносова, Россия
Маккарти Диана	Кембриджский университет, Великобритания
Мельчук Игорь Александрович	Монреальский университет, Канада
Нивре Йоаким	Упсальский университет, Швеция
Ниренбург Сергей	Университет Мэриленда, Балтимор, США
Райгородский Андрей Михайлович	МФТИ, Россия
Раскин Виктор	Университет Пердью, США
Селегей Владимир Павлович	Компания АBBYY, МФТИ, Россия
Хови Эдуард	Университет Карнеги — Меллон, США
Шаров Сергей Александрович	Университет Лидса, Великобритания

Организационный комитет

Селегей Владимир Павлович, <i>председатель</i>	Компания АBBYY, Россия
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН, Россия
Браславский Павел Исаакович	Уральский федеральный университет, Россия
Добров Борис Викторович	НИВЦ МГУ им. М. В. Ломоносова, Россия
Захаров Леонид Михайлович	МГУ им. М. В. Ломоносова, Россия
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова, Россия
Козеренко Елена Борисовна	Институт проблем информатики РАН, Россия
Кутузов Андрей Борисович	Университет Осло, Норвегия
Ляшевская Ольга Николаевна	Институт русского языка им. В. В. Виноградова РАН, Россия
Пиперски Александр Чедович	РГГУ, Россия
Толдова Светлана Юрьевна	НИУ «Высшая школа экономики», Россия
Шаврина Татьяна Олеговна	СберТех, Россия
Шаров Сергей Александрович	Университет Лидса, Великобритания

Секретариат

Родионова Ольга Игоревна, *координатор оргкомитета* Компания АBBYY, Россия

Ульянова Анна Вячеславовна, *секретарь оргкомитета* РГГУ, Россия

Рецензенты

Азарова Ирина Владимировна

Андрианов Андрей Иванович

Антонова Александра Александровна

Артемова Екатерина Леонидовна

Баранов Анатолий Николаевич

Беликов Владимир Иванович

Богданов Алексей Владимирович

Богданова-Бегларян Наталья Викторовна

Богуславский Игорь Михайлович

Бочаров Виктор Владиславович

Бурцев Михаил Сергеевич

Васильев Виталий Геннадьевич

Гусев Илья Олегович

Добров Борис Викторович

Добровольский Владимир Андреевич

Добровольский Дмитрий Олегович

Зализняк Анна Андреевна

Захаров Леонид Михайлович

Иванов Владимир Владимирович

Ивойлова Александра Михайловна

Иомдин Леонид Лейбович

Инькова Ольга Юрьевна

Катинская Анисья Юрьевна

Кибрик Андрей Александрович

Клышинский Эдуард Станиславович

Клячко Елена Леонидовна

Князев Сергей Владимирович

Кобозева Ирина Михайловна

Копотев Михаил Вячеславович

Кортаев Николай Алексеевич

Котельников Евгений Вячеславович

Котов Артемий Александрович

Куратов Юрий Михайлович

Кутузов Андрей Борисович

Лапошина Антонина Николаевна

Левонтина Ирина Борисовна

Лобанов Борис Мефодьевич

Лукашевич Наталья Валентиновна

Малафеев Алексей Юрьевич

Митрофанова Ольга Александровна

Мичурина Мария Александровна

Недолужко Анна

Новицкий Валерий Игоревич

Орлов Евгений Анатольевич

Переверзева Светлана Игоревна

Петрова Мария Владимировна

Пиперски Александр Чедович

Подлеская Вера Исааковна

Рыгаев Иван Петрович

Селегей Владимир Павлович

Слюсарь Наталия Анатольевна

Смирнов Иван Валентинович

Смулов Иван Михайлович

Татевосов Сергей Георгиевич

Толдова Светлана Юрьевна

Федорова Ольга Викторовна

Феногенова Алена Сергеевна

Хохлова Мария Владимировна

Циммерлинг Антон Владимирович

Шаврина Татьяна Олеговна

Шаров Сергей Александрович

Янко Татьяна Евгеньевна

Contents¹

Abrosimov K. I., Mosyagina A. G. Sodner for Russian nested named entity recognition	1
Alibaeva K., Loukachevitch N. Analyzing COVID-related Stance and Arguments using BERT-based Natural Language Inference	8
Апресян В., Шмелев А. Русские итеративные наречия: штрихи к лексиграфическому портрету	18
Artemova E. L., Zmeev M. V., Loukachevitch N. A., Rozhkov I. S., Batura T. V., Ivanov V. V., Tutubalina E. V. RuNNE-2022 Shared Task: Recognizing Nested Named Entities	33
Баранов А. Н. Корпусный эксперимент в лингвистической экспертизе	42
Богуславский И. М., Вилинбахова Е. Л. Имена собственные в сфере действия метаязыкового отрицания	50
Bolshakova E. I., Telegina A. D. Refining Criteria of Paronymy for Building Computer Dictionaries of Russian Paronyms	61
Bondarenko I. Contrastive fine-tuning to improve generalization in deep NER	70
Buyanov I., Sochenkov I. The dataset for presuicidal signals detection in text and its analysis	81
Chistova E., Smirnov I. Discourse-aware text classification for argument mining	93
Chuprina A. O. Russian Verbal Affixation in Mental Lexicon: Priming Study and Its Online Replication With True and Stem-Modified Relative Prime Verbs	106
Dementieva D., Nikishina I., Logacheva V., Fenogenova A., Dale D., Krotova I., Semenov N., Shavrina T., Panchenko A. RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora	114
Добровольский Д. О., Зализняк Анна А. Эвиденциальность и эпистемическая оценка в значении немецких глаголов sollen и wollen (по данным немецко-русского параллельного корпуса)	132
Dobrovolskii V. A., Michurina M. A., Ivoylova A. M. RuCoCo: a new Russian corpus with coreference annotation	141
Дурягин П. В. Просодия и многозначность в русских дискурсивных формулах	150
Evdokimova A., Nikolaeva Y., Budennaya E. Motion verbs in multimodal communication	159
Evseev D. A. Lightweight and accurate system for entity extraction and linking	176
Evseev D. A., Nagovitsin M. S., Kuznetsov D. P. Controllable Multi-attribute Dialog Generation with PALs and Grounding Knowledge	185

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

Федорова О. В. В прямом эфире «Фильм о грушах»: когнитивные особенности репортажа	203
Fishcheva I. N., Osadchiy D., Bochenina K. O., Kotelnikov E. V. Argumentative Text Generation in Economic Domain	211
Goloviznina V. S., Kotelnikov E. V. Automatic Summarization of Russian Texts: Comparison of Extractive and Abstractive Methods	223
Гончаров А. А., Кобозева И. М. Еще раз о существительном причина: конструкции с сентенциальным актантом, вводимым союзом что	236
Горбова Е. В., Чуйкова О. Ю. Суффиксальная имперфективация приставочных глаголов: рекордсмены и аутсайдеры (в словаре, корпусе и Рунете)	247
Gusev I. Russian Texts Detoxification with Levenshtein Editing	264
Inkova O. Y., Nuriev V. A., Popkova N. A. The Role of Paragraph in the Corpora of Annotated Texts	273
Князев С. В., Евстигнеева М. Ю. «Пословный» тональный контур в русских говорах: количественный аспект	284
Kolesnikova A., Kuratov Y., Kononov V., Burtsev M. Knowledge Distillation of Russian Language Models with Reduction of Vocabulary	295
Колмогорова А. В., Калинин А. А. Эмоциональный анализ постов ВКонтакте: классификатор или регрессор?	311
Korzun V., Gadecky D., Berzin V., Ilin A. Speaker-agnostic mouth blendshape prediction from speech	323
Kotelnikov E., Loukachevitch N., Nikishina I., Panchenko A. RuArg-2022: Argument Mining Evaluation	333
Кротова Е. Б., Цветаева Е. Н., Шарандин А. В., Добровольский Д. О. Вариативность грамматической нормы: возможности корпусного и количественного анализа (на материале немецкого предлога wegen)	349
Кустова Г. И. Сентенциальные актанты ментальных предикатов с союзом когда (по данным Национального корпуса русского языка)	358
Левонтина И. Б. Милый идеал	367
Li B., Weng Y., Song Q., Deng H. Artificial Text Detection with Multiple Training Strategies	375
Лобанов Б. М., Житко В. А. Метод и программная модель оценки статистических характеристик мелодики речи	382
Maloyan N., Nutfullin B., Ilyshin E. DIALOG-22 RuATD Generated Text Detection	396
Movsesyan A. A. Russian neural morphological tagging: do not merge tagsets	402

Orzhenovskii M. V. Detecting Auto-generated Texts with Language Model and Attacking the Detector	412
Пекелис О. Е. Русские временные клаузы на шкале семантико-синтаксической интеграции (на примере сочинительного союза когда)	420
Petrova M., Ponomareva M., Ivoylova A. The Pilot Corpus of the English Semantic Sketches	436
Pletenev S. Between Denoising and Translation: Experiments in Text Detoxification	447
Подлеская В. И. «Потому что больше никто не читает прозу»: грамматика и просодия автономных причинных придаточных по корпусным данным	456
Posokhov P. A., Skrylnikov S. S., Makhnytkina O. V. Artificial text detection in Russian language: a BERT-based Approach	470
Post M. Spoken corpora of spontaneous speech as a source to study polar question intonation in Russian dialects	477
Rozhkov I. S., Loukachevitch N. V. Machine Reading Comprehension Model in RuNNE Competition	488
Shamardina T., Mikhailov V., Chernianskii D., Fenogenova A., Saidov M., Valeeva A., Shavrina T., Smurov I., Tutubalina E., Artemova E. Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian	497
Шерстинова Т. Ю., Москвина А. Д., Кирина М. А., Карышева А. С., Колпащикова Е. О. Тематическое моделирование русского рассказа 1900–1930: наиболее частотные темы и их динамика	512
Татевосов С. Г., Киселева К. Л. Русский делимитатив: линейный порядок или движение к кульминации?	527
Trofimchuk D. Distilled Model for Russian News Clustering: much lighter and faster, still accurate	538
Voloshina E., Serikov O., Shavrina T. Is neural language acquisition similar to natural? A chronological probing study	550
Vychegzhanin S. V., Kotelnikov E. V. Collocation2Text: Controllable Text Generation from Guide Phrases in Russian	564
Янко Т. Е. Метод поиска просодических данных по ключевым словам	577
Zimmerling A. V. Historical Text Corpora and the Conclusiveness of Linguistic Analysis	586
Zinina A. A., Kotov A. A., Zaidelman L. Y., Arinkin N. A. Human Communicative Responses to Different Modes of Gaze Management by the Robot	594
Abstracts	603
Авторский указатель	612
Author Index	613

Sodner for Russian nested named entity recognition

Kirill Abrosimov

Seldon

Nizhny Novgorod, Russia

abrosimov.k@myseldon.com

Arina Mosyagina

Seldon

Nizhny Novgorod, Russia

mosyagina.a@myseldon.com

Abstract

The article describes the solution for Russian nested named entity recognition that we presented in the RuNNE competition. The solution is based on the Sodner model that predicts named entities in a text as a graph. During the competition we improved the training dataset and annotated the additional corpus that contains entities of the few-shot classes. After several experiments with different model parameters high macro F1 and few-shot F1 scores were obtained – 74.08 and 64.41 respectively.

Keywords: named entity recognition; nested named entity recognition; NER; few-shot NER; Sodner

DOI: 10.28995/2075-7182-2022-21-1-7

Соднер для извлечения вложенных именованных сущностей на русском языке

Абросимов К. И.

Seldon

Нижний Новгород, Россия

abrosimov.k@myseldon.com

Мосягина А. Г.

Seldon

Нижний Новгород, Россия

mosyagina.a@myseldon.com

Аннотация

В статье описано решение задачи извлечения вложенных именованных сущностей для русского языка, представленное в рамках соревнования RuNNE. Для решения была использована модель Sodner, представляющая именованные сущности в тексте в виде графа. В ходе соревнования нами был доработан обучающий датасет и размечен дополнительный корпус текстов, содержащих сущности few-shot классов. В результате нескольких экспериментов с различными параметрами модели были получены высокие значения макро F1 и few-shot F1 – 74.08 и 64.41 соответственно.

Ключевые слова: извлечение именованных сущностей; извлечение вложенных именованных сущностей; NER; few-shot NER; Sodner

1 Introduction

Named entity recognition (NER) is a task that aims at identifying objects of certain classes (e.g. Person, Organization, Location etc.) in a given unstructured text. Practical applications of this task include news and social media analytics, content recommendation systems and extracting data from medical texts. Different use cases require different classes of entities. Mostly, named entities are proper nouns, but they can also be common nouns or numbers (phone numbers, quantities, dates).

One of the NER’s subtasks is nested named entity recognition (NNE). The main difference of NNE is that a named entity in this case can be a part of a larger one. For instance, the text span “*Московский государственный университет имени М.В.Ломоносова*” contains two entities: the whole span is an organization and the “*М.В.Ломоносова*” part is a person’s name. The contrary to NNE is flat named entities, within which other objects cannot be identified by NER algorithms. Most of the existing solutions for the NER task only deal with flat named entities, while in real-world text data nested named

entities are more frequent. Many current methods, which are used for flat named entity recognition, treat NER as a sequence labelling task. However, sequence-based methods can also be used for NNE, as well as the methods based on hypergraphs. Nested named entity recognition in Russian is a novel task, for which the NEREL dataset was presented by the RuNNE shared task organizers [17].

NER models usually require large in-domain labeled datasets, although in many cases available amounts of data are very limited and annotating a training dataset might be too resource-consuming. For this reason, the few-shot NER was introduced. In a few-shot surrounding, a NER model is trained on just a few labeled examples of a certain class. The NEREL dataset includes 3 few-shot classes of named entities [4].

In the given article our team presents the solution for the RuNNE competition. The article is structured as follows: in section 2 the overview of the existing NNE and few-shot NER algorithms is provided; section 3 is devoted to the training data; section 4 describes our solution for the task; in section 5 the obtained results and analysis are presented, and the conclusion is made in section 6.

2 Related work

Most of the current solutions for the nested named recognition task use sequence-based methods. For instance, in [8] the authors propose a sequence-to-set model that predicts a set of named entities in a text at once, which is its main difference from typical sequence-to-sequence NER models that identify one entity at a time and are more suitable for flat named entities. Sequence-based methods can also be used while stacking model layers. In [3] the NNE task is solved by stacking flat NER layers of a neural network. Each flat NER layer comprises a BiLSTM layer and a CRF layer. In [10] a partly-layered network architecture is used for identifying nested and overlapping named entities.

Other methods include detecting boundaries of the entities in a text. A case in point is the HIT model that was proposed in [12]. This model leverages entity boundaries and connections between its internal tokens without considering their order. In [6] the authors present a two-stage NER identifier that was inspired by the task of object detection in computer vision. The process consists of locating named entities in the text by accurately defining their boundaries and labeling the located spans later. The authors of [7] use a span-based method that is enhanced with boundary detection to predict words that are boundaries of entities. Hypergraph-based methods are another type of methods for nested named entity recognition. In this case, segmental hypergraph representation is used for entity modeling [11].

Three methods of solving the few-shot NER task are described in [2]: prototype-based methods, self-training and noisy supervised pre-training. The authors of the article use a pre-trained RoBERTa model as a base model for the experiments. Prototype-based method, which they introduce, is similar to the previous methods in terms of using nearest neighbors for choosing an entity label. The difference of this method is that only prototypes, not separate tokens, are considered during comparison. The StructShot method proposed in [13] is based on structured neighbor learning. During training the method uses a supervised NER model, which is trained on a source domain, for feature extraction. The architectures, which the authors use, are BiLSTM and BERT-based model. For prediction, the given method uses a nearest neighbor classifier and a Viterbi decoder for modeling label dependencies. The MUCO model described in [9] utilizes O-class entities to induce new undefined classes of entities. Such classes are later jointly classified with predefined classes. For identification of undefined classes, a prototypical network is trained on predefined classes, and the O-class entities, which typically cluster during training, are classified according to clusters.

In [5] NER is treated as a language modeling task, which is another method for few-shot NER. In this case, a pretrained language model is fine-tuned to predict a class label word in an entity position in a given text. One of the current state-of-the-art methods for NER in a few-shot setting is CONTaiNER. This method uses contrastive learning to optimize inter-token distribution distance [1].

3 Data

Training data for the model was provided by the RuNNE competition organizers. The NEREL dataset comprises more than 900 Russian Wikinews articles and can be used for training not only named entity recognition, but also relation extraction models. NEREL includes 29 types of named entities in total, and training model to recognize 3 of them required a specific few-shot method [4].

During the analysis of the provided dataset some inconsistencies were identified, so the following corrections were made:

- all prepositions were included inside DATE objects' borders
- if a DATE object was followed by the word “года” (i.e. “17 июня 2018 года”), this word was also added inside the object's borders
- street suffixes such as “улица”, “проспект” and “площадь” were added into the FACILITY objects
- contexts “гражданин/гражданка/граждане + COUNTRY” were marked up as NATIONALITY with keeping the COUNTRY object as an internal entity
- occasional errors such as missed or intersecting entities were corrected
- punctuation marks were excluded from the entities

Some of the corrections were performed automatically and some manually.

One of the most frequent inaccuracies in the NEREL dataset was different number of nested entities within an object of the same class. A case in point is PROFESSION class: if an entity of this class had an embedded geopolitical entity, two different ways of annotation, which did not depend on the context, were observed. According to one of them, only the geopolitical object should be annotated as a nested entity: *[президент [США] COUNTRY] PROFESSION*. However, in the approximately same number of cases the same text span in the similar context can be annotated differently, namely the separate profession name is also an internal entity: *[[президент] PROFESSION [США] COUNTRY] PROFESSION*. To uniformize the training corpus, all of the embedded single profession names were annotated as PROFESSION entities. This principle of uniformity was also applied to the LAW class, e.g. for the entities such as *[[УК] LAW [РФ] COUNTRY] LAW*. During the experiments, our model was first trained on the original dataset and then on the cleared one. The results were compared, and the F1 score of the model trained on the cleared corpus increased by 1% compared to the original NEREL.

For the training model to identify objects of the few-shot classes (PENALTY, WORK_OF_ART and DISEASE) an additional training corpus was used. For this corpus 480 news texts were collected and manually annotated. The source of the texts is the company's news database, and certain SQL masks were used as a search criterion. During the annotation process only the fragments (from 1 to 3 sentences) that included at least one entity of any few-shot class were used.

4 Experimental setup

To solve this problem, the Sodner [14] model (A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition) was used. It was developed by researchers from Wuhan University and Tianjin University.

4.1 Sodner model

The model predicts named entities in the form of a graph, the vertices of which represent the entities, while the edges represent the relationships between entities (Entity Fragment Relation Graph). For the purposes of the competition the relations of nesting entities and the relations of discontinuity of the entity were used. The Sodner model consists of the following parts: word representation, graph convolutional network, span representation and joint decoding. A pre-trained Russian language BERT [15] by DeepPavlov was used for word representation, however Sodner can also work with such models as ELMO and Word2Vec with the embeddings being additionally transmitted to the Bi-LSTM network. Attention-guided graph convolutional network (AGGCN) [16] shows good results in processing syntactic dependency in text for entity extraction. This model consists of three parts: the multi-head attention mechanism is applied to the adjacency matrix to extract various patterns in the form of a matrix of weights of the connections of the vertices of the graphs. The results of each of the heads are processed by a Densely Connected Layer. The resulting matrices are fed to the linear combination to combine data into hidden representations using a fully connected network. As a result, the relevant information is extracted from the text parsing. Afterwards, a span enumeration is performed to consider each fragment of the text. The final decision is made in two stages: the first stage is the extraction of all named entities; the second stage is the pairwise classification of extracted entities to detect the dependencies between

them. Fully connected neural networks are used for each of the stages. As a result, the model uses text as input, and a graph representation of the text parsing in the format of a symmetric adjacency matrix. The model's output is a graph dependency of named entities. The natasha library was used for syntactic analysis during the competition.

Let's consider the following example from the test sample: "Говорят, что бывший президент Чувашии развел в министерстве кумовство", пишут журналисты интернет-издания. Figure 1 shows the adjacency matrix of parsing, which will be processed by AGGCN.

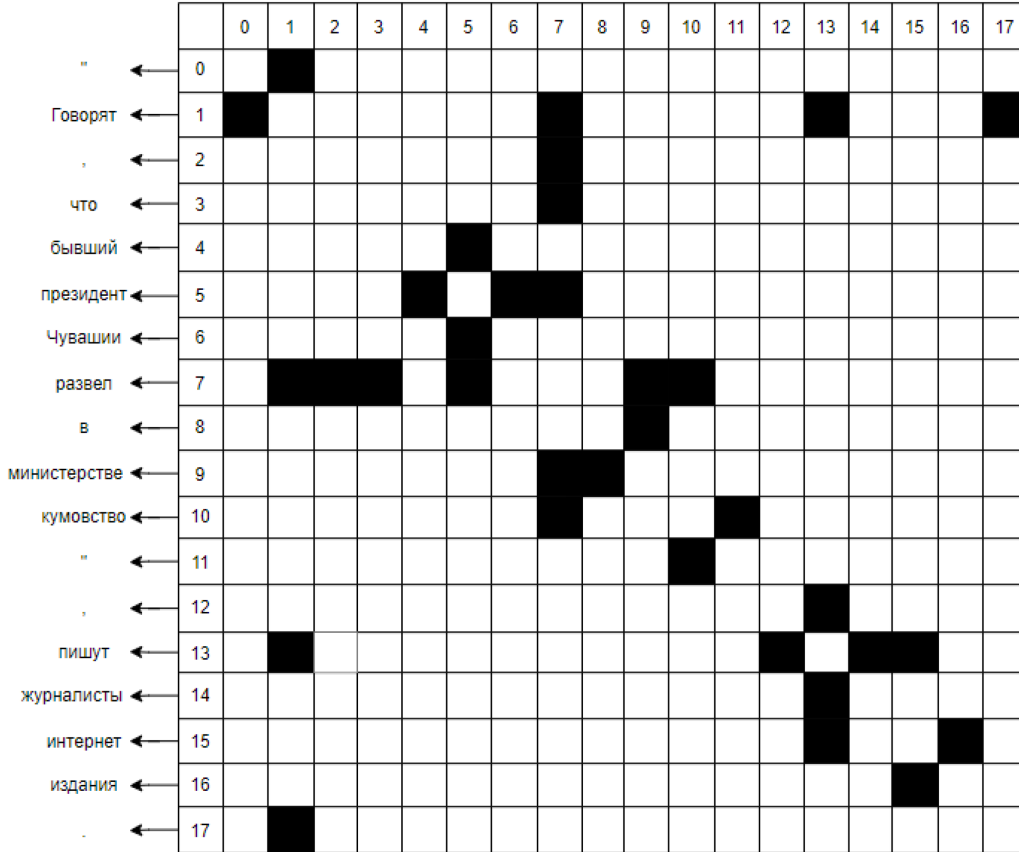


Figure 1: Adjacency matrix of sentence parsing

With the help of a pre-trained BERT, we obtain a word representation. After processing the received data, the Sodner model returns a graph shown in Figure 2.

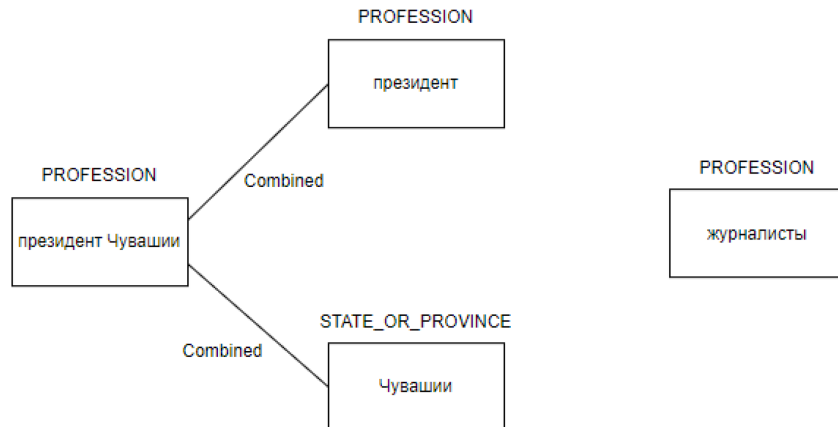


Figure 2: NER predicted by Sodner

4.2 Conducting experiments

A large number of experiments were conducted with various model configurations and datasets. The training of the Sodner model on GPU takes 10-12 hours on average. The main metrics in the competition were f1-scores: macro f1-score for regular NER, excluding few-shot classes, and macro f1-score for few-shot entities. All the metrics presented in the tables below are obtained from the final test sample.

5 Results and discussion

Next, we introduced a simple Sodner, which consists of 2 GCN layers with dropout 0.4, a vector representation of depth features dimension of 20, with 1 MLP layer of 150 neurons with dropout 0.4 and also uses an additional LSTM layer to represent words after the pre-trained BERT. First, we learned how much correcting errors in the layout of the dataset improves the prediction result of a simple Sodner and also how much the recognition of few-shot entities will improve thanks to additional marked-up data.

Data	Mention F1	Mention Recall	Mention Precision	Macro F1	Macro F1 few-shot
Original	81.88	80.37	83.44	72.17	54.56
Cleared	82.53	81.29	83.81	73.13	56.94
Augmented	81.57	79.50	83.76	72.39	60.87

Table 1: A simple Sodner trained on various data

Results in Table 1 allow us to conclude that the corrected version of the marked-up data has improved the metrics of the predictive model. However, there is always a possibility of incorrectly labeled named entities being present in the test sample. For example, the first result with extracted entities had an indexing error. For the phrase "*Российская Федерация в прошлом.*" indices [0, 20] were extracted instead of the true ones [0, 19]. However, even with this error, the result did not equal 0, which means there are errors in the markup of the test data. Additional data allowed the model to better extract few-shot entities. The next step consisted of configuring the model to obtain more usable results. There are a lot of settings in the model, some of which can help improve the metrics. The table below shows the main settings: the AGGCN block is represented by the number of layers, dropout, depth features dimension (dfd); the MLP block is the number of layers with the LSTM layer being used for additional information about words. The rest of the parameters are the recommended values specified in the article by the model developers. All experiments were carried out using an augmented cleared dataset.

GCN					Metrics				
layers	drop	dfd	MLP	LSTM	Mention F1	Mention Recall	Mention Precision	Macro F1	Macro F1 few-shot
2	0.4	20	1	+	81.57	79.50	83.76	72.39	60.87
1	0.2	64	2	+	82.23	80.63	83.89	73.78	59.35
2	0.2	64	2	–	83.01	81.14	84.97	74.12	61.85
3	0.2	64	3	–	82.77	81.62	83.95	72.56	64.41
3*	0.2*	64*	3*	–*	82.44*	81.69*	83.20*	72.45*	64.41*
4	0.2	64	4	–	82.79	81.52	84.10	72.30	64.29

Table 2: Sodner settings

Table 2 shows various model settings and the resulting metrics. The asterisk indicates the same model, but with additional named entities that were obtained using the NER model pre-trained on the natasha library. The models showed the biggest improvement when the vector representation of the GCN text was increased (from 20 to 64), as well as after disabling the LSTM layer. The resulting solution was an ensemble of the two best models that was created based on the following principle: all few-shot entities were taken from one model, and the remaining classes of named entities were taken from the other. In

total, the following metrics were obtained: macro f1 – 74.08, macro f1 few-shot – 64.41. Figure 3 shows the scheme that allowed us to obtain this result.

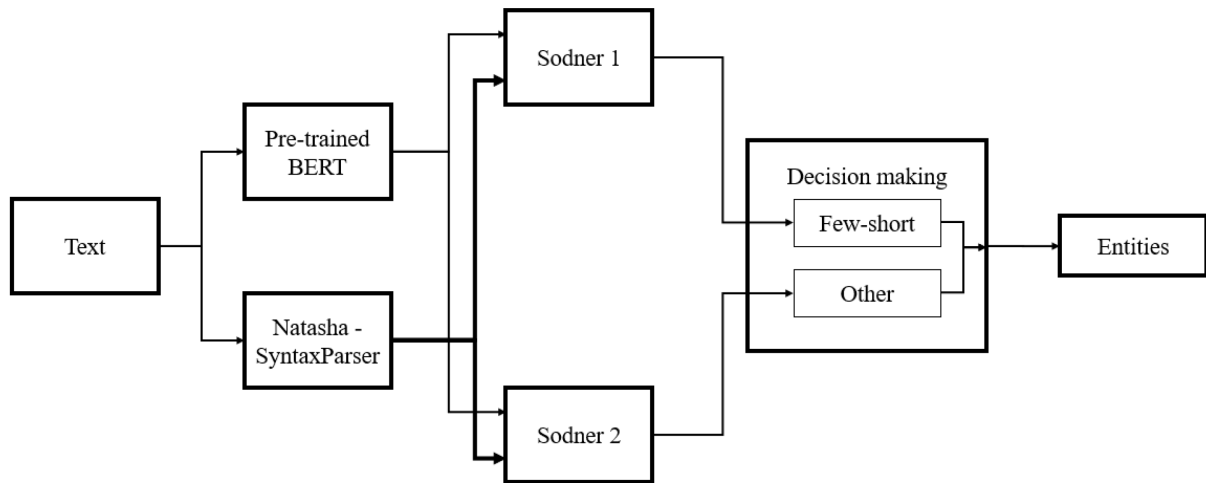


Figure 3: Our model that showed the best result

The presented solutions, despite very good results exceeding the baseline, have a number of difficulties. One of such difficulties is obtaining a parsing matrix, which means the quality of the nested entity extraction model itself depends on the quality of the resulting adjacency matrix. It is also worth noting that the model is not aimed at solving tasks with few-shot classes. Therefore, a small collection of examples of a certain class results in the model being undertrained on these classes and the extracted few-shot classes are labeled incorrectly. As part of further research on the Sodner model, it can be noted that experiments were conducted on the coefficients of importance of the problem – the extraction of the entity or extraction of the classes of connections between these entities, which are used in the loss function during the model training. One of the advantages of the model is the ability to extract not only nested entities, but also discontinuous ones. Such entities were present in the NEREL dataset, but they were not enough for high-quality extraction of such relations. Not all experiments with settings are presented in the final table, only the most significant ones. Not all the parameters were experimented with, which could also help improve the results.

6 Conclusion

As a result of participating in the RuNNE competition, our team was able to achieve excellent results, surpass the baseline model by editing and correcting inaccuracies in the NEREL dataset, adding additional marked-up texts with few-shot entities, as well as using the Sodner model with various parameters. We were able to get macro f1 – 74.08 and macro f1 few-shot – 64.41, which is 6.5% and 19.75% higher than the baseline metrics (macro f1 – 67.44, macro f1 few-shot – 44.66).

References

- [1] Das Sarkar Snigdha Sarathi, Katiyar Arzoo. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning — 2021. — Vol. arXiv:2109.07589. — version 1. Access mode: <https://arxiv.org/abs/2109.07589v1>.
- [2] Huang Jiaxin, Li Chunyuan et al. Few-Shot Named Entity Recognition: A Comprehensive Study — 2020. — Vol. arXiv:2012.14978. — version 1. Access mode: <https://arxiv.org/abs/2012.14978v1>.
- [3] Ju Meizhi, Miwa Makoto et al. A Neural Layered Model for Nested Named Entity Recognition // Proceedings of NAACL-HLT 2018. — New Orleans, Louisiana, USA, 2018. — P. 1446–1459.
- [4] Loukachevitch Natalia, Artemova Ekaterina et al. NEREL: A Russian Dataset with Nested Named Entities, Relations and Events — 2021. — Vol. arXiv:2108.13112. — version 2. Access mode: <https://arxiv.org/abs/2108.13112v2>.
- [5] Ma Ruotian, Zhou Xin et al. Template-free Prompt Tuning for Few-shot NER — 2021. — Vol. arXiv:2109.13532. — version 1. Access mode: <https://arxiv.org/abs/2109.13532v1>.
- [6] Shen Yongliang, Ma Xinyin et al. A Two-stage Identifier for Nested Named Entity Recognition — 2021. — Vol. arXiv:2105.06804. — version 2. Access mode: <https://arxiv.org/abs/2105.06804>.
- [7] Tan Chuanqi, Qiu Wei et al. Boundary Enhanced Neural Span Classification for Nested Named Entity Recognition // Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence. — New York, New York, USA, 2020. — P. 9016–9023.
- [8] Tan Zeqi, Shen Yongliang et al. A Sequence-to-Set Network for Nested Named Entity Recognition — 2021. — Vol. arXiv:2105.08901. — version 2. Access mode: <https://arxiv.org/abs/2105.08901>.
- [9] Tong Meihan, Wang Shuai et al. Learning from Miscellaneous Other-Class Words for Few-shot Named Entity Recognition — 2021. — Vol. arXiv:2106.15167. — version 1. Access mode: <https://arxiv.org/abs/2106.15167v1>.
- [10] Waldis Andreas, Mazzola Luca. Nested and Balanced Entity Recognition Using Multi-Task Learning — 2021. — Vol. arXiv:2106.06216. — version 1. Access mode: <https://arxiv.org/abs/2106.06216>.
- [11] Wang Bailin, Lu Wei. Neural Segmental Hypergraphs for Overlapping Mention Recognition — 2018. — Vol. arXiv:1810.01817. — version 1. Access mode: <https://arxiv.org/abs/1810.01817v1>.
- [12] Wang Yu, Li Yun et al. HIT: Nested Named Entity Recognition via Head-Tail Pair and Token Interaction // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. — Punta Cana, Dominican Republic, 2020. — P. 6027–6036.
- [13] Yang Yi, Katiyar Arzoo. Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning — 2020. — Vol. arXiv:2010.02405. Access mode: <https://arxiv.org/abs/2010.02405>.
- [14] Fei Li, Zhichao Lin, Meishan Zhang, Donghong Ji. A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition. — 2021. — Vol. arXiv: 2106.14373. — version 1. Access mode: <https://arxiv.org/pdf/2106.14373.pdf>.
- [15] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. — 2019. — Vol. arXiv: 1905.07213. — version 1. Access mode: <https://arxiv.org/pdf/1905.07213.pdf>.
- [16] Guo Zhijiang, Zhang Yan, Lu Wei. Attention Guided Graph Convolutional Networks for Relation Extraction. — 2020. — Vol. arXiv: 1906.07510. — version 8. Access mode: <https://arxiv.org/pdf/1906.07510.pdf>.
- [17] Artemova, Ekaterina and Zmeev, Maksim and Loukachevitch, Natalia and Rozhkov, Igor and Batura, Tatiana and Braslavski, Pavel and Ivanov, Vladimir and Tutubalina, Elena. RuNNE-2022 Shared Task: Recognizing Nested Named Entities // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog”. — 2022.

Analyzing COVID-related Stance and Arguments using BERT-based Natural Language Inference

Kamila Alibaeva

Faculty of Computational Mathematics
Lomonosov Moscow State University
Leninskie Gory, 1/4, Moscow, Russia
camalibi@yandex.ru

Natalia Loukachevitch

Research Computing Center
Lomonosov Moscow State University
Leninskie Gory, 1/4, Moscow, Russia
louk_nat@mail.ru

Abstract

In this paper we present our approach for stance detection and premise classification from COVID-related messages developed for the RuArg-2022 evaluation. The methods are based on so-called NLI-setting (natural language inference) of BERT-based text classification (Sun et al., 2019), when the input of a model includes two sentences: a target sentence and a conclusion (for example, *positive to masks*). We also use translating Russian messages to English, which allows us to leverage COVID-trained BERT model. Besides, we use additional marking techniques of targeted entities. Our approach achieved the best results on both RuArg-2022 tasks. We also studied the contribution of marking techniques across datasets, tasks, models and languages of RuArg evaluation. We found that "<A:ASPECT> keyword </A:ASPECT>" gave the highest average increase over corresponding basic methods.

Keywords: stance detection, premise classification, natural language inference, BERT, RuArg-2022

DOI: 10.28995/2075-7182-2022-21-8-17

Автоматический анализ авторской позиции и аргументов с использованием архитектуры BERT на основе подхода вывода по тексту

Алибаева К.

МГУ имени М.В. Ломоносова
Ленинские горы, 1/4, Москва
camalibi@yandex.ru

Лукашевич Н.

МГУ имени М.В. Ломоносова
Ленинские горы, 1/4, Москва
louk_nat@mail.ru

Аннотация

В этой статье представлен подход к определению позиции автора и классификации доводов из сообщений, связанных с ковидной инфекцией, разработанный для тестирования RuArg-2022. Предложенные методы основаны на так называемом NLI-варианте (natural language inference, вывод по тексту) использования модели BERT для классификации текстов. При этом подходе на вход модели поступают два предложения: целевое предложение и заключение (например, *позитивно к маскам*). Для классификации также используется перевод сообщений с русского языка на английский, что позволяет использовать специализированную англоязычную модель BERT, дообученную на текстах, посвященных тематике обсуждений ковидной инфекции. Кроме того, мы исследуем дополнительное выделение целевых объектов. Предложенный подход показал наилучшие результаты в обеих задачах RuArg-2022.

Ключевые слова: определение позиции автора, классификация доводов, текстовый вывод, модель BERT, RuArg-2022

1 Introduction

Opinion mining is an an important task in natural language processing (Pang et al., 2008; Liu, 2012). This task was started from general sentiment analysis over texts or text fragments such as users' reviews, which should extract overall authors' sentiment conveyed in texts. More detailed analysis of opinions can be achieved via so-called targeted sentiment analysis, which determines the author's sentiment towards

specific entities or topics, discussed in texts. Sentiment expressed in relation to specific targets can be different from general sentiment of the text.

Targeted sentiment analysis comprises such tasks as aspect-based sentiment analysis (ABSA) (Pontiki et al., 2016), which determines sentiment in relations to aspects (parts or characteristics) of an entity, reputation monitoring of companies and organizations (Amigó et al., 2013; Loukachevitch and Rubtsova, 2015), extraction of an attitude towards some topics, so-called stance detection (Mohammad et al., 2016), or sentiment relations between entities (Rusnachenko et al., 2019).

Another significant direction of opinion mining is argument mining (Lawrence and Reed, 2020). Argument mining tasks are quite diverse, minimal tasks are detection of arguments and classifying them to "for" and "against" classes (so-called premise classification).

For Russian, various tasks of sentiment analysis such as general sentiment analysis, aspect-based sentiment analysis, reputation monitoring have been studied. But only a few works were devoted to stance detection and argument mining. This paper is devoted to description of an approach proposed for the RuArg-2022 evaluation (Kotelnikov et al., 2022), which is devoted to stance detection and premise classification for COVID-related topics discussed in users' comments. We study methods of text classification based on so-called NLI-setting (natural language inference) of BERT-based text classification (Sun et al., 2019; Golubev and Loukachevitch, 2020), for which the input of a model includes two sentences: a target sentence and a conclusion (for example, *positive to masks*). Besides, we use additional marking of targeted entities. Our approach achieves the best results on both RuArg-2022 tasks. We also study the contribution of marking techniques across datasets, tasks and models of RuArg evaluation. We found that marking 4 (<A:ASPECT> keyword </A:ASPECT>) gave the highest average increase over corresponding basic methods.

2 Related Work

Intensive study of the stance detection task in social networks began in 2016, when Mohammad et al. (Mohammad et al., 2016) created the SemEval-2016 dataset containing five independent topics, such as *legalization of abortion* or *Hillary Clinton*. Each of the tweet/topic pairs selected for annotation was annotated via the CrowdFlower crowdsourcing system by at least eight annotators. Sobhani et al. (Sobhani et al., 2019) presented the problem of stance detection on several topics (Multi-target) and created a dataset that consists of three sets of tweets corresponding to target pairs (US presidential candidates): Donald Trump and Hillary Clinton, Donald Trump and Ted Cruz, Hillary Clinton and Bernie Sanders. Tweets with hashtags related to two politicians were extracted to form the dataset. The task was to determine the position of the author (for, against or otherwise) to each of the politicians mentioned in the tweet. Last stance-oriented studies are devoted to extraction of stance towards various aspects of COVID epidemic (Glandt et al., 2021; Miao et al., 2020).

The best results for stance detection in the SemEval-2016 experiments were obtained using the SVM-ngrams classifier, which used word and symbol n-grams (Mohammad et al., 2016) as features. In recent works, it has been found that the best results in stance detection are achieved by approaches based on the BERT (Devlin et al., 2018) neural network model. Gnosh et al. (Ghosh et al., 2019) compared previous approaches and found that the BERT model is the best model for stance detection in the SemEval2016 dataset. The work (Glandt et al., 2021) compares three groups of methods for stance detection regarding aspects of COVID: based on LSTM and CNN networks, and also based on the BERT model. The best results with a large margin are given by models based on BERT.

For Russian, in 2015-2016 a shared task on targeted sentiment analysis was organized (Loukachevitch and Rubtsova, 2015). The participants should extract sentiments towards banks or mobile operators from tweets. Later, the results on these datasets were greatly improved by using BERT-based classifiers (Devlin et al., 2018) and automatically annotated additional data (Golubev and Loukachevitch, 2020; Golubev and Loukachevitch, 2021; Smetanin and Komarov, 2021). The best results were achieved using Russian BERT RuBERT (Kuratov and Arkhipov, 2019) and a sentence-pair classification task (such as natural language inference (NLI)), when auxiliary sentences are added to initial sentences (Sun et al., 2019). In (Nugamanov et al., 2021), a new Russian dataset annotated with stance in relation to

four COVID aspects (masks, quarantine, vaccination, government actions), was presented. For stance detection, classical machine methods, several BERT-based classifiers were used. The best results were obtained with the NLI setting of the BERT model.

In (Vychegzhanin and Kotelnikov, 2017), the authors study stance towards children vaccination. The dataset consists of messages from the social network "VKontakte" classified to two classes: "for" and "against". The best results (84.3 F-measure) were achieved by the SVM classifier with rbf kernel. In subsequent work (Vychegzhanin and Kotelnikov, 2019) additional two topics were considered: "unified state exam" and "human cloning". The best results were obtained using by majority voting based on various classifiers (kNN, SVM, Naive Bayes, etc.)

Ethnicity-targeted sentiment analysis was considered in (Koltsova et al., 2020). The task was to determine hate-speech by classifying into three classes. The RuEthnoHate dataset containing 5,5K social media texts has been created. The best results were achieved by deep learning models despite a relatively small dataset size. The performance significantly benefit from a combination of linguistic and sentiment features with BERT pre-training and fine-tuning techniques.

3 Tasks and Data

RuArg-2022 evaluation (Kotelnikov et al., 2022) is devoted to analysis of COVID-related opinions and includes two tasks: stance detection and premise classification in relation to three topics: masks, vaccination, and quarantine during COVID epidemic. In the first task, it is required to determine the point of view (stance) of the text's author in relation to a given topic expressed in a given fragment. In the second task, it is necessary to determine if the text contains premises "for" or "against" to a given claim.

The dataset consists of single sentences. In total, 9,550 sentences were annotated by stance and premises for all three topics. Thus, each sentence has six labels. Each label can have one of the following values: "for", "against", "other", or "irrelevant". The difference between tasks can be explained as follows: the author of opinion can be positive to mask wearing, but does not explain why. In this case, stance to masks is "for", but argument for mask wearing is absent therefore the correct class for premise classification is "other".

Participating systems should automatically annotate each test sentence by stance and premises for each topics separately. In total, six labels (with one of four values) must be assigned to the sentence. For evaluating performance of systems, macro F-measure was used. It was calculated as averaging of F-measures of three relevant categories for each topic - $macroF_{rel}$ (Kotelnikov et al., 2022). In fact, for both tasks four-class classification is carried out, but irrelevant class is not significant for opinion mining therefore macro-averaging over three classes is performed.

4 BERT-based Natural Language Inference

The RuArg-2022 evaluation includes two tasks: stance detection and premise classification, which are both considered as four-class classification for each topic. Additionally, we can single out the relevance classification in both tasks, which separates irrelevant texts for each topic. Irrelevant texts are the same for both stance detection and premise classification. Therefore it is possible to consider a two-stage classification: extraction of relevant texts and then three-class classification for stance detection or premise classification.

Our approach to all tasks is based on NLI (Natural Language Inference) setting of the BERT model (Sun et al., 2019; Golubev and Loukachevitch, 2020). The NLI method adds to every input sentence an additional sentence, which can be an 'assumption' of the original sentence class. In the relevance classification task, the assumption sentence was the aspect itself ('Masks', 'Vaccination', 'Quarantine'). For other tasks, the assumption includes also a stance option (for the stance classification task) and a sentiment (for the premise classification task). In this way, the multi-label classification tasks are transformed into binary classifications for the model. Such a binary classification model for stance detection or premise classification can be trained and applied to all topics, it does not require training separate classifier for each topic. The NLI model was selected because it showed high performance in previous studies (Sun et al., 2019; Golubev and Loukachevitch, 2021; Nugamanov et al., 2021).

Table 1 shows examples of input for all tasks: relevance classification, stance detection, and premise classification.

We studied the following configurations based on the same NLI-BERT approach, where the RuBERT conversational model (Kuratov and Arkhipov, 2019) was used for Russian text representation:

- two-stage classification – three classifiers: relevance detection on the first stage, three-class classifiers for stance detection and premise classification applied to relevant sentences (**2stage3classf**),
- two four-class classifiers for stance detection and premise classification (**1stage2classf**),
- two-stage classification – two classifiers: relevance detection on the first stage, a single classifier for both tasks (stance detection and premise classification) (**2stage2classf**),
- three-stage processing – two classifiers using English translation: relevance detection on the first stage based on Russian texts, machine translation of Russian sentences into English, a single classifier for both tasks of translated texts using a specialized COVID-tuned BERT (**3stageEnglish2classf**).

For the relevance task, no additional text preprocessing was applied, the input sentence pairs were in Russian in all settings. The relevance classifier was trained using entire training part of the dataset.

For the 3stageEnglish2classf method, all texts having at least one relevant topic were translated into English using Helsinki-NLP/opus-ru-en model from HuggingFace Transformers library. Stance detection and premise classification tasks were combined into a single task via NLI method so every pair of the aspect and the current text that is relevant to this aspect was transformed into six input objects (three for each of the stance options and three for each of the sentiments for the premise classification). To run test cases, for every test text the relevant aspects set was obtained with the trained relevance classifier and then, as before, for each such aspect from the set six input examples were obtained. The classifier gives probabilities of answers "yes" (1) or "no" (0) for each label and topic in the stance detection and premise classification tasks. The final label is chosen according to the maximum soft-max BERT-classifier output value of the outputs for each label option.

The classifiers include the BERT models (different for each task) and the full-connected neural network. For all the above-described configurations, the same BERT parameters were used: learning rate = 0.000005, batch size = 16, epochs = 2. The parameters were not selected using validation with extra data because BERT learning is a very expensive and enduring process. The full-connected network consists of the dropout layer, the linear layer (BERT hidden size layer, 256) size, ReLU activation function, one more dropout layer and one more linear layer (256, number of classes) size, the softmax layer.

For the relevance task DeepPavlov/rubert-base-cased-conversational model was used that was trained on OpenSubtitles, Dirty, Pikabu, and a Social Media segment of Taiga corpus. For the stance detection and premise classification task specialized covid BERT model ¹ was used which was pretrained on a corpus of English messages from Twitter about COVID-19.

5 Additional Marking of Target Entities

Targeted opinion mining including stance detection and premise classification involves a target entity (topic). In this respect it is similar to relation extraction, which involves two entities. In previous works, several entity representation methods for relation extraction were proposed (Zhou and Chen, 2021), which we decided to compare in the RuArg evaluation. We evaluated the following entity representation techniques based on entity representations for relation extraction:

1. **Entity mask.** This technique introduces new special tokens [ASPECT] to mask the supposed entities in the original text, where ASPECT is substituted with one of the three topics studied in the evaluation; entity type,
2. **Entity marker.** This technique introduces a special tokens pair [E0], [/E0] to enclose the topic entity, therefore modifying the input text to the format of “[E0] keyword[/E0]”,
3. **Entity marker (punct).** This technique is a variant of the previous technique that encloses entity spans using punctuation. In our case, it modifies the input text to “* keyword*”. In contrast to the previous technique, this one does not introduce new special tokens into the model’s vocabulary,

¹digitalepidemiologylab/covid-twitter-bert-v2

Task	Sentence	Aspect	Tokenized input
Relevance	(‘I don’t get it. They said it was enough to wear a mask and gloves so they wouldn’t get infected when you left the street.’, ‘Masks’)	Masks	[‘[CLS]’, ‘i’, ‘don’, ‘ ’, ‘t’, ‘get’, ‘it’, ‘ ’, ‘they’, ‘said’, ‘it’, ‘was’, ‘enough’, ‘to’, ‘wear’, ‘a’, ‘mask’, ‘and’, ‘gloves’, ‘so’, ‘they’, ‘wouldn’, ‘ ’, ‘t’, ‘get’, ‘infected’, ‘when’, ‘you’, ‘left’, ‘the’, ‘street’, ‘ ’, ‘[SEP]’, ‘masks’, ‘[SEP]’]
Relevance	(‘At a time when the time is right to introduce quarantine, it seems too early.’, ‘Vaccination’)	Vaccination	[‘[CLS]’, ‘at’, ‘a’, ‘time’, ‘when’, ‘the’, ‘time’, ‘is’, ‘right’, ‘to’, ‘introduce’, ‘qu’, ‘aran’, ‘tine’, ‘ ’, ‘it’, ‘seems’, ‘too’, ‘early’, ‘ ’, ‘[SEP]’, ‘va’, ‘cci’, ‘nation’, ‘[SEP]’]
Stance Detection	(‘Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the quarantine.’, ‘Against Quarantine’)	Quarantine	[‘[CLS]’, ‘vacation’, ‘would’, ‘only’, ‘give’, ‘rise’, ‘to’, ‘the’, ‘spread’, ‘of’, ‘the’, ‘virus’, ‘ ’, ‘and’, ‘it’, ‘was’, ‘not’, ‘the’, ‘weekend’, ‘that’, ‘had’, ‘to’, ‘be’, ‘declared’, ‘but’, ‘the’, ‘qu’, ‘##aran’, ‘##tine’, ‘ ’, ‘[SEP]’, ‘against’, ‘qu’, ‘##aran’, ‘##tine’, ‘[SEP]’]
Stance Detection	(‘[USER], the virus is smaller than the mask cells, and the drops of water with which it flies are larger’, ‘None-stance Masks’)	Masks	[‘[CLS]’, ‘[’, ‘user’, ‘]’, ‘ ’, ‘the’, ‘virus’, ‘is’, ‘smaller’, ‘than’, ‘the’, ‘mask’, ‘cells’, ‘ ’, ‘and’, ‘the’, ‘drops’, ‘of’, ‘water’, ‘with’, ‘which’, ‘it’, ‘flies’, ‘are’, ‘larger’, ‘[SEP]’, ‘none’, ‘-’, ‘stance’, ‘masks’, ‘[SEP]’]
Premise Classification	(‘When I started buying groceries and started wearing a mask, everyone laughed.’, ‘Neutral to masks’)	Masks	[‘[CLS]’, ‘when’, ‘i’, ‘started’, ‘buying’, ‘groceries’, ‘and’, ‘started’, ‘wearing’, ‘a’, ‘mask’, ‘ ’, ‘everyone’, ‘laughed’, ‘ ’, ‘[SEP]’, ‘neutral’, ‘to’, ‘masks’, ‘[SEP]’]
Premise Classification	(‘China, without any vaccine, managed the infection, the method of self-isolation.’, ‘Positive to quarantine’)	Quarantine	[‘[CLS]’, ‘china’, ‘ ’, ‘without’, ‘any’, ‘vaccine’, ‘ ’, ‘managed’, ‘the’, ‘infection’, ‘ ’, ‘the’, ‘method’, ‘of’, ‘self’, ‘-’, ‘isolation’, ‘ ’, ‘[SEP]’, ‘positive’, ‘to’, ‘qu’, ‘##aran’, ‘##tine’, ‘[SEP]’]

Table 1: NLI-based BERT input representation.

4. **Typed entity marker.** This technique incorporates the stance topic types into entity markers. In our case, it introduces new special tokens “A:ASPECT“, “/A:ASPECT“, where ASPECT is the corresponding stance topic. The input text is accordingly modified to “<A:ASPECT> keyword </A:ASPECT>”,

5. **Typed entity marker (punct).** This variant marks the target span and target types without introducing new special tokens, which in our case looks as follows: * @ ASPECT @ keyword *.

Used keywords are shown in Table 2. We gathered keywords for marking in the following way. The RuArg evaluation concerns three topics: masks, quarantine, and vaccines. Thus, we selected these words (*mask*, *quarantine*, *vaccination*) as initial keywords for marking. Besides, we added synonyms of initial keywords and morphologically related words, known examples of vaccines. Each marking method replaces a word from the keyword list in case if the current text is relevant to the corresponding topic of the current word. All the techniques of marking are illustrated in Table 3.

All keywords were prepared in Russian. To use keywords for the stance detection and premise classification tasks in the English-based 3stageEnglish2classf approach, Russian keywords were translated using the same translation model as for the dataset’s texts. Original Russian lists are bigger than translated ones because some Russian words have the same translations into English. In Table 2 both original

Russian and auto-translated English variants are presented.

Aspect	Related words list (Russian)	Related words list (English)
Masks	'маска', 'масочный'	'mask'
Quarantine	'карантин', 'карантинный', 'локдаун'	'quarantine', 'lockdown'
Vaccination	'вакцина', 'вакцинный', 'вакцинация', 'иммунизация', 'вакцинировать', 'вакцинирование', 'прививка', 'прививать', 'прививочный', 'спутник', 'спутник v', 'модерна', 'pfizer', 'ковивак', 'эпиваккорона', 'astrazeneca'	'vaccine', 'immunization', 'vaccination', 'satellite', 'satellite v', 'moderna', 'pfizer', 'quivac', 'epivaccorone', 'astruseneca'

Table 2: Keyword lists used in marking methods.

No.	Marking rule	Text	Text with markers
1	[ASPECT]	'He certainly didn't make it, two weeks quarantine, and he went to work, healthy, not infected!'	'He certainly didn't make it, two weeks [QUARANTINE], and he went to work, healthy, not infected!'
2	[E0] keyword [/E0]	'After all, when a person wears a mask on his face and mouth, the perfect habitat for every microorgan appears.'	'After all, when a person wears a [E0] mask [/E0] on his face and mouth, the perfect habitat for every mibaselinecro-organ appears.'
3	* keyword *	'The normal decline must be after vaccination, at least the people will get less sick.'	'The normal decline must be after * vaccination *, at least the people will get less sick.'
4	<A:ASPECT> keyword </A:ASPECT>	'Now the finals are accepted only by your citizens, and ours either fly directly to you or wait for the quarantine to be cancelled.'	'Now the finals are accepted only by your citizens, and ours either fly directly to you or wait for the <A:QUARANTINE> quarantine </A:QUARANTINE> to be cancelled.'
5	* @ ASPECT @ keyword *	'Then the academic epidemiologist Gundarov said everything about death masks and panic.'	'Then the academic epidemiologist Gundarov said everything about death * @ MASKS @ masks * and panic.'

Table 3: Marking methods.

6 Results on the RuArg Dataset

Table 4 shows results obtained with all approaches described in Section 4 on the validation part of the RuArg dataset. Approaches 2stage2classf and 3stageEnglish2classf were also applied with all the marking methods. It can be seen that all models obtained much better results than the baseline model provided by RuArg-2022 organizers. The baseline used “bert-base-cased” model from Hugging Face². Three BERT models were trained separately for all three topics: “masks”, “vaccines”, “quarantine”.

²<https://huggingface.co/bert-large-cased>

In our proposed approach, three best methods (3stageEnglish2classf, 3stageEnglish2classf + marking4, 3stageEnglish2classf + marking5) were applied to the test part of the RuArg dataset. The final RuArg leaderboard is shown in Table 5. The best results (camalibi) in the competition were obtained with the proposed model 3stageEnglish2classf + marking5, its scheme is shown in 1.

Approach	Stance Detection	Premise Classification
Baseline	39.24	45.17
2stage3classf	59.76	54.25
1stage2classf	61.74	61.05
2stage2classf	60.85	57.70
2stage2classf + marking1	60.64	61.14
2stage2classf + marking2	62.23	63.26
2stage2classf + marking3	60.29	58.89
2stage2classf + marking4	62.57	62.93
2stage2classf + marking5	58.59	58.72
3stageEnglish2classf	69.81	67.81
3stageEnglish2classf + marking1	68.82	66.64
3stageEnglish2classf + marking2	67.33	67.06
3stageEnglish2classf + marking3	67.54	68.47
3stageEnglish2classf + marking4	71.30	67.37
3stageEnglish2classf + marking5	71.29	66.55

Table 4: Results of the proposed methods on the validation set.

Participant	Stance Detection	Premise Classification
camalibi	69.68	74.04
sevastyanm	68.15	72.35
iamdenay	66.76	65.55
ursdth	65.73	70.64
sopilnyak	56.03	43.38
kazzand	55.52	56.03
morty	53.53	54.53
invincible	52.86	54.28
dr	47.50	60.36
baseline	41.80	43.55

Table 5: Results on the RuArg test set. The best results (camalibi) in the competition were obtained with the proposed model 3stageEnglish2classf + marking5.

From Tables 4, 5, we can see that the contribution of marking methods may vary for different datasets and tasks. For example, marking 2 and 4 improve the basic method 2stage3classf in the stance detection task, marking techniques 1, 2, 4, 5 improve premise classification for the same basic method. The 3stageEnglish2classf basic method on the validation dataset can be improved using marking 4 and 5 for stance detection and marking 3 for premise classification.

To measure contribution of markers across tasks and datasets, we calculated the average improvement of marking techniques over a corresponding basic method for: two tasks of RuArg-2022, two datasets (validation and test) and two methods: Russian-based 2stage3classf and English-based 3stageEnglish2classf. The results of averaging are presented in Table 6. We can see that for stance detection the best techniques are markings 4, 5. For premise classification, the best marking technique is marking 4, and marking 5 is similar to a basic method on average. Thus, we can conclude that the marking method 4

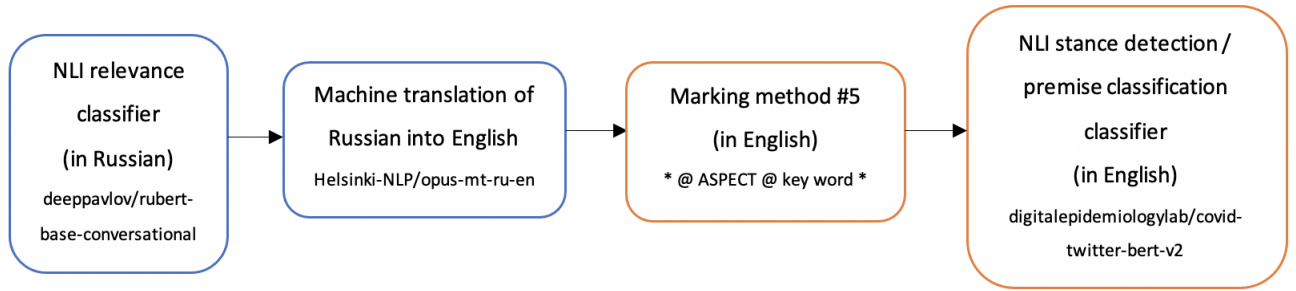


Figure 1: The pipeline scheme achieved the best results in the RuArg evaluation.

(<A:ASPECT> keyword </A:ASPECT>) was the best on average on the RuArg tasks. These results correlate with findings of (Zhou and Chen, 2021), which found that marking 4 (Typed entity markers) was best for BERT-based relation extraction.

Marking No.	Stance Detection	Premise Classification
1	0.46	-1.5
2	-0.71	-2.26
3	0.24	-0.61
4	0.79	1.94
5	1.65	-0.2

Table 6: Average score difference on marking methods.

7 Error Analysis

Tables 7 and 7 present confusion matrices for stance detection and premise classification. It can be seen that opposite labels are rarely mixed up. It is more difficult to distinguish between any polar opinion (stance or argument) and neutral one. Thus, the study should be continued to understand how best to find markers of difference between polar and neutral opinions.

Model prediction	Irrelevant	Against	Other	For
Irrelevant	2824	6	7	3
Against	0	144	46	5
Other	0	97	715	105
For	0	10	93	238

Table 7: Confusion matrix (stance classification).

Model prediction	Irrelevant	Against	Other	For
Irrelevant	2824	15	7	0
Against	0	72	54	4
Other	0	64	1087	47
For	0	7	37	81

Table 8: Confusion matrix (premise classification).

Tables 9 and 10 present examples that were misclassified by the best model.

Topic	Text	True label	Model prediction
vaccination	'Andrei will be first in line for the vaccine.'	For	Other
masks	'And only with masks do we want to stop this second wave?'	Other	For
quarantine	'It's funny– quarantine doesn't work, treatment doesn't work, prevention is funny– but!'	Other	Against

Table 9: Misclassification examples (stance detection)

Topic	Text	True label	Model prediction
vaccination	'Who will not be destroyed by the coronavirus, drugs and vaccines will kill'	Against	Other
masks	'In Japan and Korea and before the pandemic, the people wore masks, especially in transport.'	Other	For
quarantine	'[USER], quarantine is only a deterrent measure, not a neutralization of the virus.'	For	Against

Table 10: Misclassification examples (premise classification).

8 Conclusion

In this paper we presented our approach for stance detection and premise classification in argument mining from COVID-related messages developed for the RuArg-2022 evaluation. The proposed method is based on so-called NLI-setting (natural language inference) of BERT-based text classification, when the input of a model includes pair of sentences: a target sentence and a conclusion (for example, *positive to masks*) and should predict if a conclusion can be entailed from the target sentence. We also used translation of Russian messages to English, which allowed us to leverage a specialized BERT model pre-trained on a text collection of COVID-related tweets. Besides, we used additional marking of targeted entities. Our approach achieved the best results on both RuArg-2022 tasks.

We also studied the contribution of marking techniques across datasets, tasks and models of RuArg evaluation. We found that marking 4 (<A:ASPECT> keyword </A:ASPECT>) gave the highest average increase over corresponding basic methods. In the current evaluation, aspects for marking were very easy to determine. In future, we plan to integrate various techniques for aspect (topic) identification to use them for improving performance in opinion mining tasks.

Acknowledgements

The work is supported by the Russian Science Foundation, grant #21-71-30003.

References

- Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2013. Overview of replab 2013: Evaluating online reputation monitoring systems. // *International conference of the cross-language evaluation forum for european languages*, P 333–352. Springer.
- 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*.
- Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: a comparative study. // *International Conference of the Cross-Language Evaluation Forum for European Languages*, P 75–87. Springer.

- Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1.
- Anton Golubev and Natalia Loukachevitch. 2020. Improving results on russian sentiment datasets. // *Conference on Artificial Intelligence and Natural Language*, P 109–121. Springer.
- Anton Golubev and Natalia Loukachevitch. 2021. Multi-step transfer learning for sentiment analysis. // *International Conference on Applications of Natural Language to Information Systems*, P 209–217. Springer.
- Olessia Koltsova, Svetlana Alexeeva, Sergei Pashakhin, and Sergei Koltsov. 2020. Polsentilex: Sentiment detection in socio-political discussions on russian social media. // *Conference on Artificial Intelligence and Natural Language*, P 1–16. Springer.
- 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"*.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.
- John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.
- Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
- Natalia Loukachevitch and Yuliya Rubtsova. 2015. Entity-oriented sentiment analysis of tweets: results and problems. // *International Conference on Text, Speech, and Dialogue*, P 551–559. Springer.
- Lin Miao, Mark Last, and Marina Litvak. 2020. Twitter data augmentation for monitoring public opinion on covid-19 intervention measures. // *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. // *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, P 31–41.
- 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.
- Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.
- Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. *Proceedings of SemEval*, P 19–30.
- Nicolay Rusnachenko, Natalia Loukachevitch, and Elena Tutubalina. 2019. Distant supervision for sentiment attitude extraction. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, P 1022–1030.
- Sergey Smetanin and Mikhail Komarov. 2021. Deep transfer learning baselines for sentiment analysis in russian. *Information Processing & Management*, 58(3):102484.
- Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2019. Exploring deep neural networks for multitarget stance detection. *Computational Intelligence*, 35(1):82–97.
- Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. 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.
- Sergey V Vychezhnanin and Evgeny V Kotelnikov. 2017. Stance detection in russian: a feature selection and machine learning based approach. // *AIST (Supplement)*, P 166–177.
- Sergey V Vychezhnanin and Evgeny V Kotelnikov. 2019. Stance detection based on ensembles of classifiers. *Programming and Computer Software*, 45(5):228–240.
- Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *CoRR*, abs/2102.01373.

Russian adverbs of frequency: a lexicographic sketch

Valentina Apresjan

HSE University
Vinogradov Russian Language Institute
of the Russian Academy of Sciences /
Volkhonka 18/2, Moscow, Russia
valentina.apresjan@gmail.com

Alexei Shmelev

Vinogradov Russian Language Institute
of the Russian Academy of Sciences /
Volkhonka 18/2, Moscow, Russia
Moscow Pedagogical State University /
Moscow, Russia
shmelev.alexei@gmail.com

Abstract

The paper is a corpus study of Russian frequency adverbs *chasto* ‘frequently’, *zachastuju* ‘often’, *redko* ‘rarely’, *izredka* ‘rarely’, etc. In Russian lexicographic tradition, frequency adverbs either lack separate entries and are explained via references to their adjectival counterparts or are treated exclusively as denotations of intervals between events. As our study demonstrates, this covers only a small fraction of their actual corpus usage. Many frequency adverbs can quantify over subjects, and thus resemble classical quantifiers such as ‘many’ or ‘few’. Even when frequency adverbs quantify over predicates, they mostly refer not to intervals between events, but merely to their number. In some cases, they quantify over aspects of events, expressed by adjectives. There are also other important properties of Russian frequency adverbs missed by the dictionaries yet revealed by corpus analysis. Most frequency adverbs have a strong preference for topic or focus position, as motivated by their semantics. Some adverbs are preferable in generalized contexts, while others refer to specific events. Certain adverbs describe violations of the norm or undesirable events. Different adverbs quantify over different time periods: while some require a long time period, others may focus on very short stretches of time.

Keywords: frequency adverbs; quantification; topic; focus

DOI: 10.28995/2075-7182-2022-21-18-34

Русские итеративные наречия: штрихи к лексикографическому портрету

Валентина Апресян

Национальный исследовательский
университет «Высшая школа
экономики»
Институт русского языка
им. В.В. Виноградова РАН /
Волхонка 18/2, Москва, Россия
valentina.apresjan@gmail.com

Алексей Шмелев

Институт русского языка
им. В.В. Виноградова РАН /
Волхонка 18/2, Москва, Россия
Московский педагогический
государственный университет /
Москва, Россия
shmelev.alexei@gmail.com

1 Вступительные замечания

Статья посвящена русским лексическим единицам, дающим количественную оценку повторяющимся ситуациям, т. е. словам *часто* и *редко* и их неточным синонимам: *нередко*, *частенько*, *зачастую*, *сплошь и рядом*, *изредка*, *нечасто*, *эпизодически*. Делается попытка на основе корпусных данных (в первую очередь на материале основного подкорпуса Национального корпуса русского языка, далее — НКРЯ) проверить и подтвердить, опровергнуть или уточнить существующие описания этих единиц — как словарные, так и данные в специальных лингвистических работах (в частности, Булыгина, Шмелев 1988; 1989; Яковлева 1994).

В толковых словарях слова *часто* и *редко*, как правило, получают отсылочные описания через прилагательные *частый* и *редкий*. Так, в Малом академическом словаре (Евгеньева 1981–1984; далее — МАС) слово *часто* описано как «Нареч. к частый». При этом, по данным МАС, прилагательное *частый* многозначно, и первое, исходное значение не имеет отношения к повторяющимся событиям, а характеризует расположение объектов или их частей в пространстве, указывая на то, что объекты или части единого объекта расположены близко друг к другу (*частый гребень, частая изгородь*) или же предмет состоит из переплетений, образующих мелкие отверстия (*частый невод, частое сито*). Второе значение по МАС указывает на быструю смену однородных элементов ситуации (*частый пульс*), и лишь третье значение дает количественную оценку разным аспектам повторяющихся ситуаций (*частые выстрелы, частый посетитель*). При этом в толковании третьего значения (обоих «оттенков») МАС включает компонент ‘через короткие / небольшие промежутки времени’ (вопрос об адекватности таких толкований будет рассмотрен ниже).

Даже не прибегая к подсчетам на основе корпусных данных, можно констатировать, что для слова *часто* самым характерным оказывается именно употребление, соотнесенное с повторяющимися ситуациями. Хотя примеры в словарной статье слова *часто* в МАС подобраны так, чтобы иллюстрировать соотнесенность со всеми тремя значениями прилагательного *частый*, употребления, соотнесенные с первыми двумя значениями (*заросло как часто, не пролезешь; часто заморгали*), для современных текстов гораздо менее обычны — употребления типа *часто заросло* вообще представляют собою исключения, а употребления типа *часто моргали* обнаруживаются в существенно меньшем количестве по сравнению с употреблениями, соотнесенными с повторяющимися ситуациями.

Слово *редко* описано в МАС аналогичным образом: «Нареч. к редкий (в 1 и 2 знач.)». Необходимость уточнения в скобках вызвана тем, что набор значений прилагательного *редкий* не симметричен набору значений прилагательного *частый*, как эти наборы поданы в МАС. Первое значение прилагательного *редкий*, иллюстрируемое такими предложениями, как *редкие волосы, редкий лес, редкий ситец* (ср. *Тускло светились редкие фонари*), в общем соответствует первому значению прилагательного *частый*. Значение, аналогичное второму значению прилагательного *частый*, у прилагательного *редкий* в МАС не выделяется, хотя в текстах оно встречается. Ср. сочетание *редкий пульс* в примерах из НКРЯ (газетный подкорпус):

- (1) *Первые признаки: судороги, острая боль в животе, тошнота, рвота, понос, головокружение, тупая головная боль, редкий пульс. [Микробам от жары хорошо, 2001]*
- (2) *...сами пробы провоцируют аллергические реакции различной тяжести — редкий пульс, обморочное состояние, холодный пот. [Варенцова Лидия, 2002]*
- (3) *Могу назвать несколько явных противопоказаний: заболевания крови, выраженная брадикардия (слишком редкий пульс) и гипотония (пониженное давление). [Володина Светлана, 2003]*

Но, поскольку данное значение в МАС не выделено, второе значение прилагательного *редкий*, иллюстрируемое примерами *редкий гость, редкие одиночные выстрелы, редкие письма*, соответствует третьему значению прилагательного *частый* (по МАС).

Кроме того, у прилагательного *редкий* в МАС выделяются еще два значения, приблизительно толкуемых как ‘необычный’ (*редкое имя*) и ‘выдающийся по своим качествам’ (*редкие способности*), но, поскольку наречие *редко* с ними не соотнесено, можно их здесь не рассматривать.

Другие толковые словари дают нам сходную картину. В словаре С. И. Ожегова и Н. Ю. Шведовой (Ожегов, Шведова 1992, далее — ОШ) дается симметричное описание прилагательных *частый* и *редкий*. Первые два значения у обоих слов связаны с расположением в пространстве. Для *частый* они толкуются как ‘состоящий из близко расположенных друг к другу частей’ (*частая изгородь, частые заросли, частый гребень, частый невод*) и ‘расположенный на небольшом расстоянии друг от друга’ (*частые станции*), для *редкий* — как ‘состоящий из далеко расположенных друг от друга частей’ (*редкая борода, редкие зубы, редкий лес*) и ‘расположенный на большом расстоянии друг от друга’ (*редкие станции*). Следующие два значения у обоих слов

описаны как связанные с расположением во времени разных событий. Для *частый* они толкуются как ‘состоящий из быстро следующих одно за другим движений, моментов’ (примеры — *частый ритм, частый пульс, частый шаг, частая стрельба*) и ‘повторяющийся, появляющийся через короткие промежутки времени’ (примеры — *частые звонки, частый гость*; и здесь же дается пример с наречием *часто встречаться*), для *редкий* — как ‘состоящий из отдаленных друг от друга моментов’ (примеры — *редкие выстрелы, редкий пульс*; обратим внимание на то, что последнее сочетание, игнорируемое в МАС, в ОШ подано как совершенно нормативное) и ‘повторяющийся, появляющийся через большие промежутки времени’ (примеры — *редкий случай, редкий гость*; и дается пример с наречием *редко встречаться*). Включение наречий в словарные статьи соответствующих прилагательных полностью соответствует принципам построения словарных статей ОШ.

В переизданиях словаря С. И. Ожегова (последнее — Ожегов 2020) статей *редко* и *часто* нет вообще.

В «Большом толковом словаре» под редакцией С. А. Кузнецова (Кузнецов 2014; далее — БТС) для слова *часто* дается отсылочная статья (отсылка к прилагательному *частый*), а словарная статья *редко* (хотя бы отсылочная) вообще отсутствует. Чем обусловлено такое решение, принятое в БТС, остается не вполне ясным. В словарных статьях прилагательных *частый* и *редкий* рассматриваемые наречия упоминаются, и, как можно понять из примеров, предполагается соотнесение интересующих нас значений со значениями прилагательных, толкуемых как ‘происходящий, повторяющийся через короткие промежутки’ (для *частый*) и ‘бывающий, происходящий, повторяющийся через большие промежутки времени’ (для *редкий*).

В старом Большом академическом словаре (Словарь 1957–1961) слова *редко* и *часто* подавались при словах *редкий* и *частый* как подходы, и толкования им не давались и даже не указывались, с какими значениями прилагательных соотносятся эти наречия.

Лексикографические решения, связывающие значения наречий *часто* и *редко* с величиной интервалов между повторяющимися ситуациями, вполне объяснимы. Они позволяют проследить логику семантической деривации (от пространственных значений, связанных с расстояниями между объектами, к значениям повторения во времени). Однако они очевидным образом не покрывают целого ряда употреблений (тоже связанных с повторяющимися ситуациями), которые тем самым вообще не получают лексикографической фиксации. Так, во фразах *Голубоглазые кошки часто бывают глухими* или *Математики редко разбираются в лингвистике* ни о каком повторении через короткие или длительные промежутки времени речь не идет. Очевидная неадекватность существующих словарных описаний побуждает внимательнее проанализировать использование слов *часто* и *редко*, а также их неточных синонимов в реальной речевой практике с привлечением корпусных данных. Такой анализ может дать материал для создания набросков лексикографических портретов этих наречий.

2 Параметры анализа

Указание на промежутки времени нерелевантно для случаев, когда предикат, к которому относится наречие, обозначает нелокализованную во времени ситуацию. В этих случаях наречия *часто* и *редко* квантифицируют один из актантов данного предиката (как правило, субъект), за счет множественности которого ситуация и воспринимается как повторяющаяся. Предикаты ‘быть глухим’ или ‘быть пьяницей’ не локализованы во времени, поэтому фраза *Извозчики часто бывают пьяницами* естественным образом понимается как ‘Многие извозчики — пьяницы’, а фраза *Математики редко разбираются в лингвистике* — как ‘Мало кто из математиков разбирается в лингвистике’. Из сказанного вытекает, что, если при нелокализованном во времени предикате имеется единственный субъект, употребление слов *часто* и *редко* становится невозможным: аномально **Петя часто бывает пьяница* (Булыгина, Шмелев 1989: 55).

Напротив того, если предикат обозначает ситуацию, локализованную во времени, слова *часто* и *редко* в норме воспринимаются как квантифицирующие именно эту ситуацию: фраза *Петя часто бывает пьян* означает, что бывает много ситуаций, когда Петя пьян; фраза *Последний год она редко мне пишет* означает, что в течение последнего года она писала мало. Ассоциация с промежутками времени здесь вполне естественна: если какая-то ситуация повторяется через небольшие промежутки времени, то такая ситуация имеет место больше раз, а большие промежутки

времени, предполагают, что ситуация имеет место меньше раз. Однако жесткой детерминированности здесь нет.

Существенны коммуникативные свойства рассматриваемых слов и связанные с ними просодические возможности. Так, слово *редко* всегда представляет собою главную рему в клаузе и несет на себе фразовое ударение. А слово *часто* может нести или не нести фразовое ударение, и его коммуникативные характеристики не столь жестко закреплены.

Если мы включим в рассмотрение неточные синонимы слов *часто* и *редко*, то мы увидим, что целый ряд их характеристик не нашел должного отражения в словарных описаниях. Так, слова *часто* и *редко*, как мы видели, могут использоваться для количественной оценки актанта, тогда как слово *эпизодически* (неточный синоним *редко*) — не может (в силу сказанного *эпизодически* не сочетается с нелокализованными во времени предикатами: аномально **Математики эпизодически разбираются в лингвистике*). Слово *часто* может соотноситься с ремой и нести фразовое ударение (*Это происходило довольно часто*), а его неточный синоним *зачастую* — не может (аномально **Это происходило зачастую*)¹. Слово *редко* всегда представляет собою главную рему в клаузе и несет на себе фразовое ударение, а его неточный синоним *изредка* во многих случаях употребляется без фразового ударения.

Некоторые важные идиосинкратичные свойства рассматриваемых выражений нашли отражение в специальных лингвистических работах (см., в первую очередь, Яковлева 1994: 299–304). Так, в указанной работе отмечается, что слово *зачастую* задает «апостериорную» оценку, содержит элемент обобщения, тяготеет «к тематическому полюсу высказывания», выражает личную оценку автора высказывания. «Апостериорность» объясняет невозможность употребления в контексте модальных показателей (аномально **Вам придется зачастую его навещать*; **Надеюсь, что мы зачастую будем видеться*), а также, добавим, при формулировании планов на будущее (аномально **Летом я планирую зачастую ездить на дачу*). Элемент обобщения приводит к тому, что в контексте недостаточно продолжительного интервала, на протяжении которого возникает рассматриваемая ситуация, употребление *зачастую* становится аномальным или сомнительным. Притяжение к тематическому полюсу высказывания объясняет уже отмеченную невозможность фразового ударения на *зачастую*. Наличие в семантике *зачастую* элемента личной авторской оценки делает неуместным использование этого слова в энциклопедических словарях или «сухих» канцелярских документах. Отмечается также «наличие латентной отрицательной оценки» в семантике *частенько*, представляющей описываемую ситуацию «в некоем фамильярном тоне», что объясняет невозможность использования *частенько* в контекстах, исключающих фамильярность или пейоративность: аномально **Его доклады частенько просто блестящи* при правильном *Его доклады часто просто блестящи* (Яковлева 1994: 300).

Материал НКРЯ позволил не только проверить и уточнить некоторые из выделенных характеристик, но и обнаружить еще некоторые релевантные параметры. Так, для описания языкового поведения наречия существенно, может ли оно характеризовать не ситуацию в целом и не какой-либо из актантов, а лишь отдельный признак актанта (как правило, выраженный прилагательным). Для наречий *часто*, *нередко*, *зачастую* и *редко* это вполне обычно; ср.:

- (4) *Только через много лет я понял, что потому-то он и оказался не сломленным до конца своих дней, что обладал даром хороших крестьян и больших художников — извлекать удовольствие из самой работы, а не ждать её часто обманчивых плодов. [Фазиль Искандер, 1966]*
- (5) *Должен ли человек поступаться ответственностью за повседневную (пусть нередко иллюзорную) правду в пользу неколебимой Истины? [Ирина Новикова, 2003]*
- (6) *Сильное беспокойство, зачастую беспочвенное или по силе своего выражения не соответствующее причинам, его вызвавшим. [Вера Елгаева, 2003]*

¹ Это наблюдение отражено в «Активном словаре русского языка» (Апресян 2017, далее АС; автор словарной статьи *зачастую* — И. Б. Левонтина).

(7) *Главное — не за что зацепиться глазом, всё одинаково: глинобитные заборы, за ними аккуратные мазанки, редко белые, всё больше синие и зелёные (потом я узнал, что здесь в белила хозяйки добавляют купорос) ... [Ю. О. Домбровский, 1964]*

Ср. также пример из АС:

(8) *Чтение чужих рукописей — порой чрезвычайно обширных — тоже было его ежедневным занятием, равно как и переписка с обидчивыми и зачастую бездарными авторами этих увесистых опусов. [Корней Чуковский]*

Напротив того, для *частенько* и *изредка* такие употребления единичны, напр.:

(9) *Даже его заместитель, частенько похмельный Громышев, просто выходил из здания и возвращался. [В. М. Шапко, 2016]*

(10) *Вокруг стояли покосившиеся, одноэтажные, изредка двухэтажные каменные дома. [Юрий Петкевич. Явление ангела (2001)]*

Для наречий *нечасто* и *эпизодически*, а также наречного выражения *сплошь и рядом* они совсем не характерны.

В целом следует иметь в виду, что использование НКРЯ в качестве инструмента анализа рассматриваемых наречий сопряжено с трудностями двоякого рода. Во-первых, вычлененные параметры не всегда однозначно коррелируют с формальными признаками, которые можно использовать при поиске в НКРЯ. Скажем, наличие отрицательной оценки часто не заложено в каком-то из использованных слов, а определяется лишь общим контекстом (да и признак положительной или отрицательной оценки в НКРЯ автоматически ищется лишь у прилагательных). Во-вторых, вследствие общей несбалансированности НКРЯ во многих случаях не имеет смысла использование изолированного статистического анализа: могут возникнуть непредвиденные системные сдвиги. Однако, к счастью, для многих параметров достаточно показательным оказывается самая грубая количественная оценка, производимая «невооруженным глазом».

Так, ингерентная тематичность или рематичность рассматриваемых наречий определенным образом коррелирует с их позицией в клаузе: тематичные наречия охотно употребляются в самом начале клаузы, а рематичные — в самом конце (при этом самая распространенная позиция для почти всех из них — непосредственно перед главным предикатом). Данные Основного корпуса НКРЯ (по состоянию на 10 февраля 2022) можно представить в виде следующей таблицы:

Позиция	В начале	Перед предикатом (после подлежащего)	В конце
<i>Часто</i>	6050	62516	2302
<i>Редко</i>	905	12321	3358
<i>Нередко</i>	2100	10763	170
<i>Нечасто</i>	47	211	232
<i>Изредка</i>	1730	6269	174
<i>Зачастую</i>	445	2360	18
<i>Частенько</i>	191	1911	69
<i>сплошь и рядом</i>	146	554	120
<i>Эпизодически</i>	12	57	28

Таблица 1: Позиция в предложении

Обращает на себя внимание тот факт, что для большинства наречий самая распространенная позиция — после подлежащего перед глаголом. Единственное исключение — *нечасто*, для которого на первом месте оказывается конечная позиция. Можно предположить, что слову *нечасто* свойственна ингерентная рематичность, как и его неточному синониму *редко*. И действительно,

как показывает материал НКРЯ, даже в тех случаях, когда *нечасто* находится в начале или середине предложения, на него падает фразовое ударение. Ср.:

- (11) *Нечасто ты меня балуешь своими звонками. [Евгений Сухов, 2007]*
- (12) *На Западе такое нечасто встретишь, даже считается неприличным. [Сати Спивакова, 2002]*

Для слова *редко*, как и следовало ожидать, более характерна конечная позиция; при этом в начальной позиции *редко* несет на себе фразовое ударение. Это вполне соответствует ингерентной рематичности *редко*. Впрочем, в большинстве случаев *редко* находится в позиции после подлежащего перед предикатом, но и в этом случае фразовое ударение падает именно на него. Третий случай, когда число употреблений в позиции конца предложения превышает число употреблений в начале, — это наречие *эпизодически*. На первый взгляд это может показаться странным, потому что *эпизодически* вовсе не относится к ингерентно рематичным и всегда несущим на себе фразовое ударение. Но если посмотреть на примеры, в которых *эпизодически* находится на последнем месте в предложении, то все становится на свои места. Выясняется, что почти во всех таких примерах *эпизодически* находится в контексте рематизирующих частиц (напр., *только, лишь*), усилительных слов (*сугубо*), используется в уступительных конструкциях или в контексте противопоставления. Ср.:

- (13) *Увы, пока аккумуляторы, несмотря на многолетнюю и даже многовековую известность, применяются лишь эпизодически. [Нурбей Гулиа, 1974]*
- (14) *О полной внешней независимости рязанских князей не может быть и речи почти в продолжение целой истории княжества; такая независимость встречалась только эпизодически. [Д. И. Иловайский, 1858]*
- (15) *В нем нет программных рассуждений, раздумий о судьбах народа и страны, а главные события и действующие лица эпохи появляются сугубо эпизодически. [Наталья Самутина, Борис Степанов, 2009]*
- (16) *Сразу оговорюсь, что Одесса как город если и появляется здесь, то эпизодически. [С. И. Трунев, 2012]*
- (17) *Потом я устал от них и смотреть перестал – перестал систематически, смотря только эпизодически. [Д. С. Лихачев, 1985]*

С другой стороны, обращает на себя внимание чрезвычайная малочисленность примеров, в которых в конце предложения находится слово *зачастую*. Это вполне вытекает из его ингерентной тематичности. Любопытно посмотреть на те примеры, в которых *зачастую* все же находится в конце. Сразу бросается в глаза, что в них при естественном произнесении *зачастую* не несет на себе фразового ударения (акцентоносителем чаще всего оказывается предшествующее слово), так что можно считать, что мы имеем дело с инверсией. Кроме того, половина из этих примеров относятся к позапрошлому веку, так что мы можем заключить, что для современного языка такая инверсия не характерна.

Следует также иметь в виду, что русский язык неоднороден в отношении рассматриваемых наречий не только с диахронической, но и с регистровой точки зрения. В тех случаях, когда стилистические и временные различия носят ярко выраженный характер (т.е. можно говорить о сильных количественных предпочтениях в ту или иную сторону), мы будем их комментировать. Например, сравнительная частота слов *изредка* и *нередко* в художественных и нехудожественных текстах коренным образом различается. *Изредка* доминирует в художественных текстах (53 ipm для *изредка* vs. 29 ipm для *нередко*), в то время как в нехудожественных текстах соотношение обратное (86 ipm для *нередко* vs. 25 ipm для *изредка*). Иногда различия более дробные: так, в

нехудожественных текстах официально-деловые и научно-учебные противопоставлены бытовым.

В целом стоит отметить, что все рассматриваемые слова крайне неоднородны с точки зрения частотности (ср. их абсолютную и относительную частотность в Основном корпусе НКРЯ ниже). Это связано в первую очередь с их семантическими и прагматическими особенностями, которые отчасти определяют их стилистическую принадлежность. Интересно, что сама идея частоты выражается в языке намного чаще, чем идея редкости. Кроме того, различные виды оценки, которые добавляются как наречиям, выражающим смысл ‘часто’, так и к наречиям, выражающим смысл ‘редко’, сильно влияют на их употребительность. Конкретные стилистические характеристики, в том случае, если они необходимы, приводятся ниже, в индивидуальных портретах наречий.

<i>часто</i> 126593	<i>нередко</i> 20162	<i>зачастую</i> 5951	<i>частько</i> 2909	<i>сплошь и рядом</i> 1255	<i>редко</i> 30657	<i>изредка</i> 12459	<i>нечасто</i> 817	<i>эпизодически</i> 159
375.6 ipm	59.8 ipm	17.6 ipm	8.6 ipm	3.7 ipm	90.9 ipm	37 ipm	2.4 ipm	0.47 ipm

Таблица 2: Частотность итеративных наречий по Основному корпусу НКРЯ

3 Штрихи к лексикографическим портретам

Некоторые свойства рассматриваемых наречий прямо вытекают из их морфологического строения или из их семантики. Так, только наречия *часто* и *редко* имеют сравнительную степень (*чаще* и *реже*), у других наречий сравнительной степени нет. Наречия *изредка* и *эпизодически* употребляются в контексте уступительных или ограничительных частиц, для наречия *редко* это не характерно, а для прочих наречий практически невозможно. Это связано с их семантикой: *изредка* и *эпизодически* указывают на то, что, с одной стороны, некоторая ситуация имеет место, а с другой — что это происходит не очень много раз (‘пусть нечасто, но бывает’). Напротив того, наречия *часто*, *нередко*, *частько*, *зачастую*, *сплошь и рядом* на то, что ситуация имеет место много раз, так что употребление уступительных или ограничительных частиц оказывается неуместно, а наречия *редко* и *нечасто* сообщают, что ситуация, уже находящаяся в поле зрения, не воспроизводится много раз, так что базы для употребления уступительных или ограничительных частиц здесь опять-таки нет.

В дальнейшем мы, как правило, не будем специально останавливаться на характеристиках такого рода, сосредоточившись на менее тривиальных свойствах рассматриваемых единиц.

3.1 *Часто*

Наречие *часто* почти не имеет ограничений на свое функционирование. Оно может квантифицировать и множественный субъект, и какую-то характеристику субъекта, и локализованный во времени предикат, быть и темой, и ремой, описывать общее устройство мира или конкретное положение дел. Наряду с описанием ситуаций, имеющих место на протяжении относительно длительного интервала времени, *часто* используется для указания на движения, быстро следующие одно за другим в некоторый заданный момент. Ср.:

- (18) *Увеличить огонь и готовить около 20 минут, часто помешивая, пока кости не станут равномерно коричневыми со всех сторон. [Рецепты национальных кухонь: Франция, 2000–2005]*

Особенно характерны употребления, в которых говорится о часто бьющемся сердце, напр.:

- (19) *Юрий Андреевич обезумел от радости. Сердце часто забилося у него. [Борис Пастернак]*

Такого рода примеры довольно многочисленны в НКРЯ, напр.:

- (20) *Мысли стали путаться у него, лицо горело, и сердце часто и громко стучало...*
[Ф. Д. Крюков, 1896]
- (21) *Сердце часто заколотилось, но Юрий поспешно завертел штурвальчик.*
[Л. С. Соболев, 1932]
- (22) *Лицо его пылало, сердце часто билось.* [Ю. П. Герман, 1952]

3.2 *Редко*

Как уже говорилось, *редко* ингерентно рематично. Для *редко* характерны редупликация и сочетаемость с *к*-местоимением (особенно когда *редко* выносится в начало предложения):

- (23) *Серые ящерики греются на солнцепеке, пленюю закрыв глаза. Редко-редко встретишь змею.* [Борис Екимов, 1999]
- (24) *Редко какие фильмы пересматриваю (если пересматриваю, то раз в 3–10 лет).*
[коллективный, 2008–2011]

Обратим внимание на то, что сочетание с *к*-местоимением невозможно для *часто*.

Впрочем, позиция начала предложения для *редко* не характерна. Помимо случаев редупликации и сочетания с *к*-местоимением, она обычно встречается при наличии противопоставления *редко, но*, в конструкциях типа *редко бывает, редко можно* и т. п., а также при инверсии (напр., *Редко являлся фельдшер*).

Употребления в конце предложения для *редко* более характерны, но в целом тоже не очень многочисленны. Они преимущественно встречаются в форме компаратива либо в сочетании с модификатором типа *крайне, относительно, чрезвычайно, очень* и пр.

Как видно из вышеприведенной таблицы, самая обычная позиция для *редко* — после подлежащего перед предикатом.

Редко может квантифицировать и множественный субъект, и какую-то характеристику субъекта, и локализованный во времени предикат, описывать общее устройство мира или конкретное положение дел. Кроме того, *редко* иногда используется для указания на движения, быстро следующие одно за другим в некоторый заданный момент, хотя такие употребления не очень характерны:

- (25) *Пустые прозрачные глаза, редко помаргивая, бессмысленно шарили по комнате.*
[Аркадий Стругацкий, Борис Стругацкий, 1964]

Как уже говорилось, *редко* в некоторых случаях может характеризовать не ситуацию в целом и не какой-либо из актантов, а лишь отдельный признак актанта. Особенно характерны такие употребления для сравнительной степени, т. е. формы *реже*:

- (26) *Аконит восточный (*A. orientale*) — цветки белые или желтоватые, реже бледно-фиолетовые...* [С. А. Курганская, 2003]

3.3 *Нередко и нечасто*

Рассматриваемые слова не симметричны в разных отношениях.

Как уже говорилось, *нередко* может быть и темой, и ремой, а *нечасто* — всегда рема. В силу этого *нечасто* может писаться и слитно, и раздельно², причем оба написания в большинстве случаев практически эквивалентны, а *нередко* в соответствии с орфографическими нормами пишется слитно (раздельное написание *не редко* возможно только в полемических контекстах и представляет собою не реализацию единицы *нередко*, а единицу *редко* с отрицанием). При этом следует иметь в виду, что в реальных текстах, в том числе вошедших в НКРЯ, могут встретиться

² Раздельное написание становится невозможно в таких сочетаниях, как *весьма нечасто*.

ошибочные написания (слитное вместо раздельного или, наоборот, раздельное вместо слитного), напр.:

(27) *Полина Ивановна забежала к Рите между делами, нечасто и нередко, раз в несколько дней. [И. Грекова, 1980]*

(28) *У нас с Александрой не редко возникали споры о прочитанном. [Н. В. Нестерова, 2013]*

Это, разумеется, затрудняет корпусный анализ.

Нередко: в целом употребление в качестве главной ремы для слова *нередко* совершенно не характерно. В значительной части случаев оно оказывается уместным, когда в предшествующем тексте введено указание на квантифицируемую ситуацию, а при помощи слова *нередко* сообщается, что эта ситуация вполне обычна. Ср.:

(29) *Поэтому ледники очень опасны: трещины в них очень глубокие, в них можно провалиться, и такие случаи бывали нередко. [Владимир Котляков, 2012]*

(30) *Драки случаются, и нередко. [Мария Цырулева, 2015]*

Нередко стилистически отмечено, для него характерна (квази)научность, оно используется для передачи как бы объективных наблюдений над миром, поэтому странны некоторые примеры использования слова *нередко* в бытовой речи. Мы уже отмечали, что частота *нередко* в нехудожественных текстах значительно выше, чем в художественных, при этом этот перевес достигается за счет учебно-научных текстов, где частота употребления достигает 144 ipm; в обиходно-бытовых частота *нередко* всего лишь 26 ipm, а в официально-деловых – 32 ipm.

Это слово может описывать общее устройство мира или конкретное положение дел, квантифицировать и множественный субъект, и локализованный во времени предикат. Для него вполне обычны примеры, когда в сфере действия оказывается какой-то признак.

(31) *...негативное отношение значительной части мигрантов к самой процедуре регистрации как запретительной, нередко дорогостоящей и требующей много времени [Снижение масштабов внутренней миграции населения в России, 2004]*

(32) *Должен ли человек поступаться ответственностью за повседневную (пусть нередко иллюзорную) правду в пользу неколебимой Истины? [Ирина Новикова, 2003]*

Нечасто: Будучи неточным синонимом наречия *редко*, слово *нечасто* всегда несет на себе фразовое ударение. При этом оно более свободно, нежели *редко*, выносится в начало предложения, сохраняя на себе фразовое ударение, напр.:

(33) *Не часто русского инженера англичане и французы приглашали руководить строительством. [Даниил Гранин, 1987]*

По сравнению с *редко*, наречие *нечасто* лучше ставится в конец предложения; такие употребления весьма многочисленны.

С точки зрения текстовой функции слово *нечасто* иногда обнаруживает некоторую эвфемистичность или смягчение: возникает впечатление, что оно использовано чтобы не говорить *редко*, как в примере (11).

Весьма характерны для *нечасто* противительные конструкции, в них выражена своего рода уступительность, т. е. идея, что все-таки сколько-то. Ср.:

(34) *Брал он нечасто, но уж когда это случалось, могучая рыба шла с сопротивлением... [Алексей Варламов, 2000]*

(35) *Сейчас встречаться нам удаётся нечасто. Но если уж встречаемся, то отдыхаем на полную катушку. [Эльвира Савкина, 2002]*

(36) *Нечасто, но виделась в ЦДЛ, несколько раз я приезжал к нему в Переделкино... [И. Н. Вирабов, 2015]*

Нечасто может квантифицировать и множественный субъект, и локализованный во времени предикат (при этом, в отличие от *редко*, по-видимому, невозможна отнесенность к отдельному признаку субъекта), описывать общее устройство мира или конкретное положение дел.

Стилистически *нечасто* не является нейтральным словом. Его частотность в Основном корпусе – 2.4 ipm, при этом это наречие практически не встречается в официально-деловых текстах (1 вхождение из частного письма 18 в.), редко встречается в научных текстах (менее, чем 1 ipm) и наиболее частотно в публицистических текстах (более 3 ipm) и в устных выступлениях (2.8 ipm в Устном корпусе). Таким образом, кажется, что *нечасто* отражает интенцию говорить «красиво».

3.4 *Изредка*

Изредка обычно не несет на себе фразового ударения, которое может ставиться на нем в специальных условиях: в сочетании с модификаторами, контрастом, уступительностью:

(37) *Я старался не думать о Клэр, но лишь изредка мне это удавалось. [Г. А. Газданов, 1930]*

(38) *Однако на чердаке я бывал только изредка. [Ю. О. Домбровский, 1964]*

(39) *Стричься надо регулярно — а массаж изредка. [коллективный, 2013]*

(40) *Правда, говорила ему «ты». И то — изредка. [В. М. Шапко, 2016]*

Для *изредка* не характерна квантификация субъекта, такие примеры единичны:

(41) *Ей шел уже пятидесятый год, но она была еще хороша собой, при этом прекрасно образована, умна и добра душой, какими изредка бывают светские дамы. [Е. Ахматова, 1873]*

Совсем невозможна отнесенность к какому-либо признаку субъекта.

Обычно *изредка* квантифицирует предикат, напр.:

(42) *Маньчжуры изредка совершали на них набеги с целью захвата пленных и дани. [Александр Храмчихин 2012]*

При этом обнаруживается немало примеров, в которых описывается движение наблюдателя вдоль редко расположенных объектов:

(43) *Мы ехали по дороге, изредка попадались освещенные дома. [Дмитрий Глуховский, 2005]*

(44) *Машина движется в абсолютной гладкой тьме, дорогу освещают только фары, изредка лучи света выхватывают стоящие у края шоссе заснеженные синюшные деревья, а потом деревенские домики. [Майя Кучерская, 2012]*

Встречаются примеры, в которых описываются движения, быстро следующие одно за другим в некоторый заданный момент:

(45) *Он изредка моргал... [А. Е. Рекемчук, 2006]*

Изредка охотно сочетается с глаголами в прошедшем времени, для него характерна референция к конкретным ситуациям, локализованным во времени, и, напротив того, не характерны генерические высказывания. Иными словами, *изредка* плохо сочетается с генерализованными предикатами, не локализованными во времени.

Изредка способно сочетаться с глаголами прерывисто-смягчительного способа действия и итеративными глаголами (*посматривать, захаживать*). Это связано с тем, что глаголы этого типа обычно выступают как ремы и акцентоносители — поэтому сочетания с ними невозможны для *редко*, которое само выполняет функции главной ремы и акцентоносителя.

3.5 Зачастую

Как уже говорилось, слово *зачастую* не может нести фразового ударения.

Может квантифицировать множественный субъект, признак или предикат, но при этом не используется для квантификации ситуаций, повторяющихся в течение относительно небольшого промежутка времени (Яковлева 1994: 300).

Примеры квантификации субъекта:

(46) *Зачастую молодые музыканты думают: слава, деньги и сопутствующая суета и есть путь артиста. [Сати Спивакова, 2002]*

(47) *Именно поэтому архивы зачастую хранят три комплекта документов – оригиналы, матрицы (копии хранения), копии для использования. [В. Б. Прозорова, 2017]*

Примеры квантификации признака:

(48) *...её бесконечные выходки, зачастую скандальные, вкупе с колючими текстами стали основанием для причисления её к панковской плеяде. [Вячеслав Борисов, 2002]*

(49) *Но действовали они, основываясь на порывах страсти — зачастую бездумных... [Анастасия Завозова, 2017]*

Пример квантификации предиката:

(50) *Фрадкис зачастую действовал некорректными методами и никогда не отвечал за те обещания, которые давал. [И. Э. Кио, 1995–1999]*

В ОШ указывается, что *зачастую* говорится «обычно о чем-н. отрицательном». Это не вполне точно: для *зачастую* вполне обычно использование, когда формулируется статистическая тенденция, и в этом случае отрицательной оценки, как правило, нет. Ср.:

(51) *Зачастую целесообразнее купить готовое технологическое решение, чем пытаться придумать заново давно изобретенный велосипед. [Алексей Репик, 2015]*

Стоит отметить, что *зачастую* не является стилистически нейтральным словом. В Основном корпусе его частотность 17.6 ipm; Яковлева (1994) отмечает неупотребительность этого наречия в «сухих официальных документах и энциклопедических статьях». Однако это трудно подтвердить количественными данными НКРЯ, в частности, в силу того, что к жанру деловых документов в НКРЯ по формальным параметрам отнесены приказы времен Великой отечественной войны, в то время как по языковым особенностям эти приказы скорее относятся к публицистическому жанру (Шмелев 2015: 290). Соответственно, при поиске в сфере функционирования «Официально-деловая» в типе текстов «Деловые документы» частотность *зачастую* оказывается значительно выше, чем в Основном корпусе, а именно, 32.8 ipm, т.е. даже выше, чем в сфере функционирования «Публицистика» (27.6 ipm).

3.6 Частенько

Для *частенько* употребление в качестве ремы не характерно. Примечательно, что даже в тех случаях, когда *частенько* находится в конце предложения, мы имеем дело либо с инверсией (52), либо с контрастной ремой (53):

(52) *Придумал ревновать, хоть учение бросай. Выпивать стал частенько.*
[М. Л. Халфина, 1972]

(53) *Повесть я ещё не скоро буду перечитывать (пока не тянет), а фильм смотрю частенько.* [коллективный, 2001–2011]

Слово *частенько* не используется для квантификации актантов, в основном оно описывает обобщения, сделанные на основании личного опыта и плохо применимо к конкретным ситуациям (ср. *Она частенько смеется во время лекций*, но не **Во время этой лекции она частенько смеялась*).

Как уже говорилось, *частенько* практически не употребляется в модальных контекстах и в контекстах, исключающих фамильярность. Это наречие ни разу не встречается в официально-деловых текстах, очень редко в учебно-научных (1.36 ipm); в обиходно-бытовой сфере его частотность составляет 5.3 ipm, в время как в публицистике – 8.9 ipm, что близко к его встречаемости в Основном корпусе.

Для него, как и для *изредка*, характерно использование в контексте вторичных имперфективов прерывисто-смягчительного и итеративного способов действия, напр.:

(54) *Дед мой частенько наведывался к Замараеву.* [Сергей Довлатов, 1983]

(55) *В Вене Гастев частенько захаживал к профессору, автору безумной теории о том, что вся женщина — от макушки до пят — всего лишь чудовищный нарост на детородном органе...* [Анатолий Азольский, 1997]

(56) *Однажды ей удалось вытащить его на какой-то муторный спектакль, где публика между тем частенько похихатывала, и громче всех это делала его Татьяна...*
[А. Н. Бузулукский, 2011]

3.7 Сплошь и рядом

Для этого наречного выражения характерно употребление по отношению к отрицательно оцениваемым или нарушающим какие-то нормы ситуациям, относительно которых могли иметься противоположные ожидания. Поэтому во многих высказываниях с этим выражением ощущается некоторая полемичность (*Неужели такое бывает? — Да сплошь и рядом*). В силу идеи нормы и кванторных свойств *сплошь и рядом* обычно сочетается с глаголами в настоящем времени.

Для этого выражения можно выделить два типа употреблений.

Во-первых, это употребление в качестве акцентоносителя ремы, которой предшествует экзистенциальный предикат (*бывает, случается, происходит*) или глагол в узуальном значении, а иногда анафорический элемент, вводящий некую данную информацию, напр.:

(57) *Однако в школах детские психические эпидемии случаются сплошь и рядом.*
[В. М. Бехтерев, 1911]

(58) *В ходе наших проверок непродуктивные и непервоочередные расходы обнаруживаются сплошь и рядом.* [Наталья Ратиани, 2003]

Во-вторых, это употребления, за которыми следует акцентоноситель ремы, в роли которого обычно выступает предикат с отрицательной оценкой, напр.:

(59) ...на место сочного слова приходило случайное, бледное, а главное — сплошь и рядом ломался ритм. [Аркадий Мильчин, 2001]

(60) Партия власти сплошь и рядом нарушает обязательства. [Г. А. Зюганов, 2003]

Сплошь и рядом охотно квантифицирует множественный субъект, а также локализованный во времени предикат, но не признак субъекта; при этом *сплошь и рядом* не употребляется для указания на регулярно повторяющиеся через малые промежутки времени события (сомнительно *Он в прошлом году сплошь и рядом звонил мне по ночам).vs.

Стилистически *сплошь и рядом* тяготеет к разговорности, о чем свидетельствует его большая относительная частотность в Устном корпусе (4.7 ipm vs. 3.7 в Основном корпусе) и неупотребительность в официально-деловой сфере. В официально-деловых текстах встретилось всего два примера на 5 миллионов словоупотреблений, причем оба примера фактически относятся к сфере публицистики (Сборник приказов Великой отечественной войны и речь депутата Государственной Думы Вишневецкого на заседании 23 февраля 1917 года).

3.8 Эпизодически

Как уже говорилось, *эпизодически* как правило соотносится с темой, а ремой может быть только в контексте модификаторов: рематизирующих частиц, усилительных слов (*сугубо*), в уступительных конструкциях или в контексте противопоставления:

(61) Его имя лишь эпизодически упоминается в вузовских учебниках и зачастую отсутствует даже в программах филологических факультетов университетов. [Е. Г. Трубецкова, 2012]

(62) В начале войны Курчатов только эпизодически возвращался к ядерно-физическим вопросам... [Раиса Кузнецова, Виктор Попов, 2012]

Однако единичные примеры рематического употребления *эпизодически* вне таких контекстов все же встречаются:

(63) Вскоре дочь возвратилась в родной дом, прежнего сожителя навещала эпизодически. [Прошу встать! Суд идет, 2003]

Как уже говорилось, *эпизодически* не используется для квантификации субъекта. Оно не сочетается с вневременными предикатами. Для него невозможна квантификация отдельного признака какого-либо актанта, хотя допустимо сочетание с причастием, представляющим собою свернутую пропозицию:

(64) Очень важно подчеркнуть чрезвычайно характерную для эпилептоидов склонность к эпизодически развивающимся расстройствам настроения... [П. Б. Ганнушкин, 1933]

Для этого слова характерна идея периодического повтора (*время от времени*, чаще, чем изредка):

(65) ...отдельные работники могут по распоряжению работодателя эпизодически привлекаться к выполнению своих трудовых функций за пределами нормальной продолжительности рабочего дня. [Кирилл Добродеев, 2002]

(66) Аналогичные проекты существуют и в отношении теплого сезонного поверхностного течения Эль-Ниньо, которое эпизодически образуется в восточной части Тихого океана. [В. П. Максаковский, 2003]

Эпизодически – достаточно редкое слово, оно встречается в Основном корпусе всего 159 раз. Интересно при этом, что основная масса вхождений приходится на нехудожественные тексты – 133 вхождения, а в художественных текстах это слово встречается в 10 раз реже (при том, что по объему нехудожественный подкорпус превышает художественный лишь в 1.2 раза). Любопытно также, что это слово – единственное из рассматриваемых, которое изменило свое значение в микродиакронии. Это слово начинает появляться в корпусе с середины 19 века, отчасти в современном значении ‘время от времени’, но иногда в ныне исчезнувшем буквальном значении ‘в некоторых эпизодах’, причем в этом последнем значении оно может сочетаться с глаголами совершенного вида, что в принципе невозможно для итеративных наречий. Ср. пример (67), в котором отражено современное употребление и пример (68) на устаревшее буквальное:

(67) *О полной внешней независимости рязанских князей не может быть и речи почти в продолжение целой истории княжества; такая независимость встречалась только эпизодически. [Д. И. Иловайский, 1858]*

(68) *Тут есть, впрочем, и четвертый, эпизодически введенный, но тоже не лишний господин, которого мы тоже будем считать. [Н. А. Добролюбов, 1860]*

3.9 Заключение

В заключение можно отметить, что, как показывает корпусное исследование, семантические, коммуникативные и сочетаемостные свойства русских итеративных наречий значительно богаче и разнообразнее, чем это отражено в существующей лексикографической традиции. Помимо выражения смысла повтора ситуации через большие или меньшие промежутки времени (который выражается не всеми наречиями), многие наречия могут использоваться для квантификации субъекта ситуации, какого-либо аспекта ситуации или количества ситуаций. Кроме того, они передают целый спектр дополнительных значений: обобщение касательно распространенности того или иного явления, описание частотности конкретных имевших место ситуаций, отношение говорящего к этим ситуациям, в том числе к их (не)нормальности и (не)желательности и т.п. Наконец, семантические свойства этих наречий обуславливают их предпочтительные коммуникативные функции в высказывании, а именно, тяготение к позиции темы, ремы или акцентоносителя ремы.

Финансирование

Исследование частично осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

References

- [1] Apresian Iu.D. (2017) Active Dictionary of Russian [Aktivnyi slovar' russkogo iazyka], vol. 3. Apresian V.Iu., Galaktionova I. V, Iomdin B. L. (eds.) Moscow; Saint Petersburg: Nestor-Istoriia Publ., 2017.
- [2] Bulygina T. V., Shmelev A. D. (1988) Mechanisms of quantification in Russian and the semantics of quantificational assessment [Mekhanizmy kvantifikatsii v russkom iazyke i semantika kolichestvennoi otsenki] // Reference and issues in text generation [Referentsiia i problemy tekstoobrazovaniia]. — Moscow, 1988. P. 5–18.
- [3] Bulygina T. V., Shmelev A. D. (1989) Spatial and temporal localization as a super-category in a sentence [Prostranstvenno-vremennaia lokalizatsiia kak superkategoriiia predlozheniia], Issues in Linguistics [Voprosy Iazykoznaniiia], 1989, 3, pp. 51–61.
- [4] Dictionary of Modern Standard Russian [Slovar' sovremennogo russkogo literaturnogo jazyka] (1957–1961). Moscow; Leningrad: Academy of Sciences Publishing House, 1957–1961.
- [5] Evgen'eva A. P. (ed.) (1981–1984) Dictionary of Russian [Slovar' russkogo iazyka]. — Moscow: Russkii iazyk Publ., 1981–1984.
- [6] Iakovleva E. S. (1994) Fragments of the Russian linguistic worldview (models of space, time, and perception) [Fragmenty russkoi iazykovoi kartiny mira (modeli prostranstva, vremeni i vospriiatii)], Moscow: Gnozis Publ., 1994.

- [7] Kuznetsov S. A. (ed.) (2014), Great Explanatory Dictionary of Russian [Bol'shoi tolkovyi slovar' russkogo iazyka]. — 1st edition: Saint Petersburg.: Norint Publ., 1998. — The author's version of 2014: <http://gramota.ru/slovari/info/bts/>.
- [8] Ozhegov S. I. (2020) Explanatory Dictionary of Russian [Tolkovyi slovar' russkogo iazyka]. — Moscow: Mir i Obrazovanie, 2020.
- [9] Ozhegov S. I., Shvedova N. Iu. (1992) Explanatory Dictionary of Russian [Tolkovyi slovar' russkogo iazyka]. — Moscow: Az' Ltd Publ., 1992.
- [10] Shmelev A. D. (2015) Business Russian and contemporary speech etiquette: myths and reality [Delovoi russkii iazyk i sovremennyyi rechevoi etiket: mify i real'nost'], Proceedings of the 5th All-Russia Conference "Stratification of the national language in modern Russian society" (Saint-Petersburg, Nov. 28–31, 2015) [Sbornik statei Piatoi Vserossiiskoi konferentsii "Stratifikatsiia nacional'nogo iazyka v sovremennom rossiiskom obshchestve" (g. Sankt-Peterburg, 28–31 noiabria 2015 g.)]. — Saint-Petersburg: Zlato-ust Publ., 2015. — P. 287–292.

References

- [1] Апресян Ю.Д. (рук.) Активный словарь русского языка, т. 3. Под ред. Апресян В.Ю., Галактионовой И. В, Иомдина Б. Л. Москва; Санкт-Петербург: Нестор-История, 2017.
- [2] Булыгина Т. В., Шмелев А. Д. Механизмы квантификации в русском языке и семантика количественной оценки // Референция и проблемы текстообразования. — Москва, 1988. — С. 5–18.
- [3] Булыгина Т. В., Шмелев А. Д. Пространственно-временная локализация как суперкатегория предложения // Вопросы языкознания. — 1989. — № 3. С. 51–61.
- [4] Евгеньева А. П. (ред.) Словарь русского языка. — Москва: Русский язык, 1981–1984
- [5] Кузнецов С. А. (гл. ред.) Большой толковый словарь русского языка. — Первое издание: СПб.: Норинт, 1998. — Авторская редакция 2014 года: <http://gramota.ru/slovari/info/bts/>.
- [6] Ожегов С. И. Толковый словарь русского языка. — Москва: Мир и Образование, 2020.
- [7] Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка. — Москва: Азъ Ltd., 1992.
- [8] Словарь современного русского литературного языка. М.; Л.: Издательство Академии Наук, 1957–1961.
- [9] Шмелев А. Д. Деловой русский язык и современный речевой этикет: мифы и реальность // Сборник статей Пятой Всероссийской конференции «Стратификация национального языка в современном российском обществе» (г. Санкт-Петербург, 28–31 ноября 2015 г.). — СПб.: Златоуст, 2015. С. 287–292.
- [10] Яковлева Е. С. Фрагменты русской языковой картины мира (модели пространства, времени и восприятия). — Москва: Гнозис, 1994.

RuNNE-2022 Shared Task: Recognizing Nested Named Entities

**Ekaterina Artemova^{1,2}, Maxim Zmeev¹, Natalia Loukachevitch³, Igor Rozhkov³,
Tatiana Batura^{3,4,8}, Vladimir Ivanov⁵, Elena Tutubalina^{1,6,7}**

¹HSE University ²Huawei Noah’s Ark lab,

³Lomonosov Moscow State University, ⁴Novosibirsk State University

⁵Innopolis University, ⁶Kazan Federal University, ⁷Sber AI,

⁸Ershov Institute of Informatics Systems, Russia

Abstract

The RuNNE Shared Task approaches the problem of nested named entity recognition. The annotation schema is designed in such a way, that an entity may partially overlap or even be nested into another entity. This way, the named entity “The Yermolova Theatre” of type ORGANIZATION houses another entity “Yermolova” of type PERSON. We adopt the Russian NEREL dataset (Loukachevitch et al., 2021) for the RuNNE Shared Task. NEREL comprises news texts written in the Russian language and collected from the Wikinews portal. The annotation schema includes 29 entity types. The nestedness of named entities in NEREL reaches up to six levels. The RuNNE Shared Task explores two setups. (i) In the general setup all entities occur more or less with the same frequency. (ii) In the few-shot setup the majority of entity types occur often in the training set. However, some of the entity types are have lower frequency, being thus challenging to recognize. In the test set the frequency of all entity types is even.

This paper reports on the results of the RuNNE Shared Task. Overall the shared task has received 156 submissions from nine teams. Half of the submissions outperform a straightforward BERT-based baseline in both setups. This paper overviews the shared task setup and discusses the submitted systems, discovering meaning insights for the problem of nested NER. The links to the evaluation platform and the data from the shared task are available in our github repository.

Keywords: nested named entity recognition, few-shot setup, shared task

DOI: 10.28995/2075-7182-2022-21-33-41

Соревнование RuNNE-2022: извлечение вложенных именованных сущностей

**Артемова Е. Л.^{1,2}, Змеев М. В.¹, Лукашевич Н. В.³, Рожков И.С.³,
Батура Т. В.^{3,4,8}, Иванов В. В.⁵, Тутубалина Е. В.^{2,6,7}**

¹Национальный исследовательский университет Высшая школа экономики

²Huawei Noah’s Ark lab,

³Московский Государственный Университет им. М.В. Ломоносова

⁴Новосибирский государственный университет

⁵Иннопалис, ⁶Казанский федеральный университет, ⁷Sber AI,

⁸Институт систем информатики им. А.П. Ершова, Россия

Аннотация

Извлечение именованных сущностей – одна из самых востребованных на практике задач извлечения информации – предполагает поиск в тексте упоминаний имен, организаций, топонимов и других сущностей. Соревнование RuNNE посвящено задаче извлечения вложенных именованных сущностей. Разметка данных допускает следующие случаи: внутри одной именованной сущности находится другая именованная сущность. Так, например в сущность класса Organization “Московский драматический театр имени М. Н. Ермоловой” вложена сущность типа Person – “М. Н. Ермоловой”. Соревнование проводится на материале корпуса NEREL (Loukachevitch et al., 2021), собранного из новостных текстов WikiNews на русском языке. В корпусе NEREL представлено 29 классов различных сущностей, а глубина вложенности сущностей достигает 6 уровней разметки.

В рамках соревнования RuNNE мы предлагаем участникам рассмотреть few shot постановку задачи. Задача предполагает извлечение вложенных именованных сущностей, в обучающем множестве большая часть типов именованных сущностей встречается достаточно часто, а некоторое количество специально отобранных типов – встречается всего несколько раз. В тестовом множестве все типы сущностей представлены одинаково. В данной статье мы описываем соревнование RuNNE, подводим его итоги и проводим сравнение решений, полученных от участников.

Ключевые слова: извлечение вложенных именованных сущностей, соревнование

1 Introduction

Named entity recognition is one of the most popular tasks in natural language processing. It involves labeling mentions of personal names, organizations, toponyms and other entities in the text. In early works, only “flat” named entities, in which internal named entities within a longer entity are not allowed, were annotated and extracted from texts. However, the assumption that all entities are flat oversimplifies the task and leads to annotation collision. For example, consider such entities as the “Ministry of Education of the Russian Federation” or the “State Duma of the Federal Assembly of the Russian Federation”. In this case, there are two options of flat labeling: a) splitting into minimum spans (“Ministry of Education” and “Russian Federation”), or b) extracting maximal spans. The latter case would miss the mention of the *Russian Federation* as a separate entity. Both labeling approaches lead to losing meaningful information pieces, which, in turn, can be useful for further processing, e.g. relation extraction.

In recent years, recognition of nested named entities has received special attention. In case of nested NEs a longer named entity can contain internal named entities. In this research direction, multiple new datasets were published (Plank et al., 2020; Ringland et al., 2019; Ruokolainen et al., 2019). The best performing methods for nested NER utilize specialized neural network architectures (Straková et al., 2019; Yu et al., 2020; Jue et al., 2020). A recent dataset for nested NER in Russian, NEREL, is annotated with 29 entity types (Loukachevitch et al., 2021).

The RuNNE shared task aims to popularize the nested NER task. To this end, we setup two sub-tasks based on the NEREL dataset: (i) the general nested NER evaluation; entities of more or less the same frequency are used (ii) the few-shot nested NER evaluation, where low-frequency entities are used. This paper (i) introduces the NEREL corpus (Section 3), (ii) reports the results of the RuNNE Shared Task (Section 5.1) and (iii) compares submitted systems (Section 5.2).

2 Related NER Datasets and Shared Tasks

Several datasets for named entity recognition in the Russian language are available, e.g. the Gareev’s dataset (Gareev et al., 2013), Persons-1000 and Collection3 (Mozharova and Loukachevitch, 2016; Vlasova et al., 2014), FactRuEval (Starostin et al., 2016), the Russian subset of the BSNLP dataset (Piskorski et al., 2019), and RURED (Gordeev et al., 2020). In particular, FactRuEval and BSNLP datasets were used for corresponding shared tasks. FactRuEval published in 2016 has been the only Russian dataset annotated with nested named entities until recently. For example, a person object in FactRuEval could include name, surname, patronymic, and nickname spans. The dataset is rather small with approximately 250 news texts annotated with 11,700 spans and 7,700 objects. This limits its effectiveness in building large neural network models. Additional descriptions and statistics of Russian NER datasets can be found in (Loukachevitch et al., 2021).

In recent years, the NLP community has held multiple shared tasks on named entity recognition, tackling different aspects of the task:

- **Multilinguality** The series of SlavNER Shared Tasks (Piskorski et al., 2021) focuses on six Slavic languages; the MultiCoNER Shared Task (Malmasi et al., 2022) features 11 world languages. XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) data suites are collections of tasks and corresponding datasets for evaluation of zero-shot transfer capabilities of large multilingual models from English to other languages. XGLUE adopts CoNLL-2002 NER (Tjong Kim Sang, 2002) and CoNLL-2003 NER (Tjong Kim Sang and De Meulder, 2003) datasets covering four languages (English, German, Spanish, and Dutch) and four entity types (Person, Location, Organization, and Miscellaneous). XTREME uses Wikiann dataset (Pan et al., 2017), where Person, Location and Organization entities were automatically annotated in Wikipedia pages in 40 languages.

- **Complexity** The MultiCoNER Shared Task (Malmasi et al., 2022) was devoted to extraction of semantically ambiguous and complex entities, such as movie and book titles in short and low-context settings.
- **Applied domains** The WNUT initiative aims at developing NLP tools for noisy user-generated text. To this end, they run multiple shared tasks on NER and relation extraction from social media (Nguyen et al., 2020; Chen et al., 2020). PharmaCoNER is aimed at clinical and medical NER (Gonzalez-Agirre et al., 2019).
- **Languages, other than English** The Dialogue evaluation campaign has supported two shared tasks for NER in Russian. FactRuEval dataset comprised news texts, Wikipedia pages and posts on social media, annotated according to a two-level schema. The RuReBUS Shared task (Ivanin et al., 2020) approached joint NER and relation extraction in Russian business-related documents.

3 NEREL Dataset

The NEREL collection is developed for studying three levels of information extraction methods including named entity recognition, relation extraction and entity linking (Loukachevitch et al., 2021). Currently, NEREL is the largest Russian dataset annotated with entities and relations compared to the existing Russian datasets. NEREL comprises 29 named entity types and 49 relation types. At the time of writing, the dataset contains 56K named entities and 39K relations annotated in 900+ person-oriented news articles. NEREL is annotated with relations at three levels: within nested named entities, within and across sentences. Entity linking annotations leverage nested named entities, and each nested named entity can be linked to a separate Wikidata entity.

The NEREL corpus consists mainly of Russian Wikinews articles having size of 1–5 Kb, as such medium-sized texts are more convenient for annotation. The BRAT tool (Stenetorp et al., 2012) was used for annotation. Three levels of annotation – named entities, relations, Wikidata links, – were performed as independent subsequent passes. NEREL dataset is publicly available. It is subdivided into train, dev, and test parts, which were used in the RuNNE evaluation. The training part was specially reduced for few-shot NER evaluation. Evaluation was carried out on the NEREL’s dev and test sets. Table 1 contains examples of nested named entities in the NEREL dataset. Entity type statistics of the dataset used for the RuNNE competition are shown in Table 2. During the evaluation phase the NEREL github repository was temporarily closed.

4 RuNNE Task

The RuNNE competition offered two tasks: general nested named entity recognition and few-shot setting. In the training set, most named entity types occur quite often, and a certain number of specially selected types occur only a few times. In the test set, all entity types are represented equally.

As a quality metric in the RuNNE competition, macro averaging of the F1-measure is used in two versions: by classes of known entities (general formulation of the problem of extracting nested named entities) and by classes of new named entities (few-shot formulation). Three entity types annotated in the NEREL dataset were selected for the few-shot setting: DISEASE, WORK_OF_ART, PENALTY.

As a baseline we employed a RuBERT model (Kuratov and Arkhipov, 2019) with a fully connected output layer, predicting each of the classes in the data, from where only the flat most internal entities were taken. For named entity encoding the IOBES (Inside-Out-Begin-End-Single) scheme was used.

5 Evaluation

5.1 Participants’ submissions

We have received 156 submissions from nine teams. Table 3 presents with the final scores of the submitted systems. Four out of nine systems outperformed the baseline. Below, we give an overview of these approaches.

Team **Ksmith (Pullenti)** achieved the best results by using a rule-based approach. Customized rules were written for every entity type (Kozerenko et al., 2018). This team made the highest number of entries to the leaderboard.

Entity	Example	Annotation
General entities		
PROFESSION	<i>governor of California</i> <i>head of Gazprom</i>	[governor of [California] _{STATE_OR_PROVINCE}] _{PROFESSION} [head of [Gazprom] _{ORGANIZATION}] _{PROFESSION}
ORGANIZATION	<i>physics department of Lomonosov Moscow State University</i> <i>Russian Government</i>	[physics department of [[Lomonosov] _{PERSON} [Moscow] _{CITY} State University] _{ORG}] _{ORG} [[Russian] _{COUNTRY} government] _{ORG}
NATIONALITY	<i>citizen of Russia</i> <i>Russians</i> <i>Russian writer</i>	[citizen of [Russia] _{COUNTRY}] _{NATIONALITY} [Russians] _{NATIONALITY} [Russian] _{NATIONALITY} [writer] _{PROFESSION}
LAW	<i>Yarovaya law</i> <i>article 84 of the Constitution of Kyrgyzstan</i>	[[Yarovaya] _{PERSON} law] _{LAW} [article [84] _{ORDINAL} of the [[Constitution] _{LAW} COUNTRY] _{LAW}] _{LAW}
CRIME	<i>complicity in murder of VGTRK journalists</i> <i>armed attack on passers-by attempt to oust Erdogan</i>	[[complicity in murder] _{CRIME} of [[VGTRK] _{ORG} [journalists] _{PROFESSION}] _{CRIME} [[armed attack] _{CRIME} on passers-by] _{CRIME} [attempt to oust [Erdogan] _{PERSON}] _{CRIME}
PRODUCT	<i>Boeing-737 MAX</i> <i>Apple Watch</i>	[[Boeing] _{ORG} -[737] _{NUMBER} MAX] _{PRODUCT} [Apple] _{ORG} Watch] _{PRODUCT}
AWARD	<i>Merit for the Fatherland Order</i> <i>gold of the Olympic Games</i> <i>champion of the Olympic Games</i> <i>Miss Russia-2017</i> <i>gold medal</i>	[Merit for the Fatherland Order] _{AWARD} [gold of the [Olympic Games] _{EVENT}] _{AWARD} [champion of the [Olympic Games] _{EVENT}] _{AWARD} [Miss [Russia] _{COUNTRY} -[2017] _{DATE}] _{AWARD} [gold medal] _{AWARD}
EVENT	<i>40th Moscow International Film Festival</i> <i>UEFA champions league</i> <i>Sochi-2014</i>	[[40th] _{ORDINAL} [[Moscow] _{CITY} International Film Festival] _{EVENT}] _{EVENT} [UEFA] _{ORG} champions league] _{EVENT} [[Sochi] _{CITY} -[2014] _{DATE}] _{EVENT}
Few-shot entities		
PENALTY	<i>\$129 million fine</i> <i>imprisonment for 3 years</i>	[[[\$129 million] _{MONEY} [fine] _{PENALTY}] _{PENALTY} [[imprisonment] _{PENALTY} for [3 years] _{DATE}] _{PENALTY}
DISEASE	<i>died from Covid</i> <i>thyroid cancer</i>	[[died] _{EVENT} from [Covid] _{DISEASE}] _{EVENT} [thyroid [cancer] _{DISEASE}] _{DISEASE}
WORK_OF_ART	<i>the host of TV show “Vzglyad”</i>	[the host of TV show [“Vzglyad”] _{WORK_OF_ART}] _{PROFESSION}

Table 1: Examples of annotating nested named entities of different types

Team **Fulstock (MSU RCC)** applied the Machine reading comprehension model (MRC) (Li et al., 2020). The MRC model treats NER as a question-answering task. Entity types are translated into Russian. Next, their definitions are gathered from dictionaries in Russian and used as questions. The MRC model comprises three binary classifiers over the output of the last hidden layer from RuBERT. The first classifier determines the starting position of a named entity. The second classifier decides about the end position of a named entity (perhaps another) of the same class. The third classifier classifies, whether the start-end pairs are a single entity. These classifiers are trained for each of the entity type.

Team **Abrosimov_Kirill (Saldon)** used the span-based Sodner model (Li et al., 2021) that can recognize both overlapped and discontinuous entities jointly. The model is based on the graph convolutional network architecture. It includes two steps. First, entity fragments are recognized by traversing over all possible text spans. Second, relation classification is implemented to judge to detect overlapping or succession. The team used a pre-trained RuBERT model as contextualized encoder and the Natasha syntax parser.¹ Additionally, selected sentences for few-shot entity types were manually annotated.

¹<https://github.com/natasha/slovnet#syntax>

Entity	train	dev	test	Entity	train	dev	test
PROFESSION	4593	860	854	IDEOLOGY	300	36	43
PERSON	4518	947	961	LOCATION	272	64	62
ORGANIZATION	4059	616	675	PRODUCT	238	30	53
EVENT	2879	707	690	CRIME	181	64	35
COUNTRY	2521	355	456	MONEY	171	29	43
DATE	2276	527	523	TIME	154	29	47
CITY	1102	208	239	DISTRICT	98	18	25
NUMBER	1026	186	230	RELIGION	94	9	24
ORDINAL	565	102	107	PERCENT	82	9	7
AGE	554	137	138	LANGUAGE	43	7	8
NATIONALITY	394	59	66	DISEASE	32	117	57
LAW	392	77	62	PENALTY	32	57	18
FACILITY	371	84	63	WORK_OF_ART	30	104	93
STATE_OR_PROVINCE	343	99	112	FAMILY	17	7	14
AWARD	328	43	121	Total	27665	5587	5826

Table 2: NEREL dataset statistics for the RuNNE competition.

User	Team	# of runs	F1 _{full_set}	F1 _{few-shot}	System Summary
<i>Baseline</i>		-	0.674	0.447	RuBERT
<i>Participating Teams</i>					
ksmith	Pullenti	44	0.811	0.710	Rule-based
abrosimov_kirill	Saldon	20	<u>0.741</u>	<u>0.644</u>	Sodner model, labelling
fulstock	MSU-RCC	7	<u>0.749</u>	<u>0.604</u>	MRC model
svetlan		23	0.607	<u>0.572</u>	n/a
LIORI		6	0.653	0.433	n/a
botbot		8	0.460	0.414	n/a
bond005	SibNN	20	<u>0.743</u>	0.404	Siemese network, Viterbi alg.
Stud2022		24	0.477	0.395	n/a
mojesty		2	0.619	0.172	n/a

Table 3: Macro-averaged F-scores on the official test sets. The best results are in bold. Results above the baseline are underlined.

Team **Bond005 (SibNN)** fine-tuned a pre-trained BERT-based transformer two times. First, the contextualized encoder is fine-tuned as a Siamese neural network using the supervised contrastive learning loss. This step brings together embeddings of named entities of the same class and moves apart named entities of different classes. The second step fine-tunes the contextualized encoder for sequence labeling using a task-specific loss function. Finally, the team applied the Viterbi algorithm to smooth prediction probabilities. The team defined the transition probabilities manually.

5.2 Results

The results (Table 3) indicate that:

1. the conventional rule-based approach is capable of outperforming recent neural models. However, the rule-based approach requires careful configuration, leading to more attempts to submit to the leaderboard;

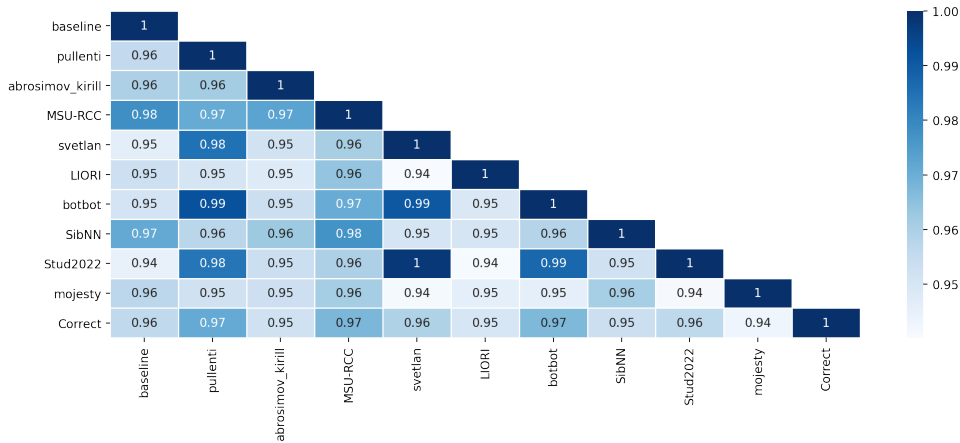


Figure 1: Agreement (Cohen’s κ) among participating systems in annotating entity types. The “Correct” row corresponds to the test dataset annotations.

2. the MRC approach benefits from additional textual descriptions and outperforms other learnable approaches;
3. the few-shot setup benefits from additional manual labelling and hand-crafted rules.

5.3 Detailed comparison of systems

Entity Type	baseln.	pullenti	abros.	MSU.	svetlan	LIORI	botbot	SibNN	Stud2022	mojesty
ORGANIZATION	23	21	28	21	20	18	17	29	14	24
COUNTRY	8	36	19	12	27	21	15	12	19	34
WORK_OF_ART	22	3	20	16	1	16	2	24	1	23
STATE_OR_PROV.	11	6	21	9	5	18	4	18	4	29
EVENT	10	9	17	8	20	7	5	13	15	8
LOCATION	9	8	12	8	6	18	8	5	6	10
FACILITY	6	9	11	11	5	7	2	10	3	10
NATIONALITY	8	2	6	7	8	10	2	9	5	7
PRODUCT	9	1	5	6	-	5	1	17	-	7
CITY	6	5	7	5	6	9	3	5	3	11
PERSON	12	2	11	4	5	7	1	6	2	8
FAMILY	9	-	10	6	2	7	1	8	1	7
DISTRICT	7	2	4	3	1	10	1	6	1	5
AWARD	8	1	9	3	2	3	1	4	1	5
DISEASE	2	3	1	3	5	3	2	6	4	7
PROFESSION	5	1	5	5	1	3	-	5	1	6
Total span matches	4401	4732	4948	4853	3511	4283	2363	4618	2361	4595

Table 4: Number of test instances with incorrectly recognized entity type. Top-16 entities cover approximately 85% of all mismatches. The last row shows the number of correctly recognized spans for each system. Symbol ‘-’ means that all system’s predictions were correct. All calculations were performed for the *full_set* test set.

We have compared solutions of participating systems in two aspects. First, we compare errors in entity type annotations given a span of a named entity was correctly detected. Second, we calculate pairwise agreements of systems’ answers, again only for the annotations that have exact match of spans. Both evaluations are done on the test set. In Table 4 we provide top entities for which systems give wrong entity type prediction. We found that around a half of errors come from six entity types: ORGANIZATION, COUNTRY, WORK_OF_ART, STATE_OR_PROVINCE, EVENT, and LOCATION. The most common mismatches are the following (correct entity is on the left of the ‘ \rightarrow ’ sign):

- ORGANIZATION → FACILITY / PRODUCT / CITY,
- COUNTRY → NATIONALITY / ORGANIZATION,
- WORK_OF_ART → ORGANIZATION / LAW / EVENT,
- STATE_OR_PROVINCE → CITY / COUNTRY / LOCATION,
- EVENT → ORGANIZATION / CRIME / DISEASE,
- LOCATION → STATE_OR_PROVINCE / CITY/ COUNTRY.

To assess pairwise agreements between systems, we build lists of spans that have exactly matching boundaries, but may have different type annotations. Then we calculate Cohen’s κ for each pair of systems (Fig. 1). Depending on a pair of systems, the number of matching spans varies between 1,294 and 5,826. Due to this difference, values in the Figure 1 slightly deviate from the ranking of the systems. Three teams, showing mediocre performance (**svetlan**, **botbot** and **Stud2022**) have highest agreement with each other. Also these three teams have the highest agreement with the **pullenti** team. This indicates that these systems learn named entity patterns, which can be described by rules, too. Other systems may gain more generalization abilities and thus are less correlated with the rule-based system. Finally, the main source of errors for all approaches may be attributed to ambiguous or complex cases.

6 Conclusion

In this paper we described the RuNNE Shared Task, aimed at extracting nested named entities from texts in Russian. We used NEREL (Loukachevitch et al., 2021) as a source dataset. RuNNE comprises two sub-tasks: (i) the general nested NER evaluation and (ii) the few-shot nested NER evaluation. Comparing submissions of the RuNNE participants, we found that the classic rule-based approach is capable to outperform recent neural models. However, it requires a special framework for writing rules and iterative improvement of rules, leading to more attempts in leaderboard submits. In both subtasks, the best model among machine learning approaches applied without additional manual data labeling was Machine Reading Comprehension model (Li et al., 2020).

Acknowledgments

The project is supported by the Russian Science Foundation, grant # 20-11-20166. The experiments were partially carried out on computational resources of HPC facilities at HSE University (Kostenetskiy et al., 2021) and the shared research facilities of HPC computing resources at Lomonosov Moscow State University. Ekaterina Artemova was supported by the framework of the HSE University Basic Research Program.

References

- Chacha Chen, Chieh-Yang Huang, Yaqi Hou, Yang Shi, Enyan Dai, and Jiaqi Wang. 2020. TEST_POSITIVE at W-NUT 2020 shared task-3: Cross-task modeling. // *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, P 499–504, Online, November. Association for Computational Linguistics.
- Rinat Gareev, Maksim Tkachenko, Valery Solovyev, Andrey Simanovsky, and Vladimir Ivanov. 2013. Introducing Baselines for Russian Named Entity Recognition. // *International Conference on Intelligent Text Processing and Computational Linguistics*, P 329–342.
- Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurre, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. // *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, P 1–10, Hong Kong, China, November. Association for Computational Linguistics.
- Denis Gordeev, Adis Davletov, Alexey Rey, Galiya Akzhigitova, and Georgiy Geymbukh. 2020. Relation extraction dataset for the Russian language. // *Computational Linguistics and Intellectual Technologies*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. // *International Conference on Machine Learning*, P 4411–4421.

- Vitaly Ivanin, Ekaterina Artemova, Tatiana Batura, Vladimir Ivanov, Veronika Sarkisyan, Elena Tutubalina, and Ivab Smurov. 2020. Rurebus-2020 shared task: Russian relation extraction for business. // *Computational Linguistics and Intellectual Technologies*, P 416–431.
- Wang Jue, Lidan Shou, Ke Chen, and Gang Chen. 2020. Pyramid: A Layered Model for Nested Named Entity Recognition. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 5918–5928.
- Pavle Kostenetskiy, Roman Chulkevich, and Viacheslav Kozyrev. 2021. HPC resources of the Higher School of Economics. // *Journal of Physics: Conference Series*, volume 1740, P 012050. IOP Publishing.
- Elena Kozerenko, Konstantin Kuznetsov, and Dmitrii Romanov. 2018. Semantic Processing of Unstructured Textual Data Based on the Linguistic Processor PullEnti. *Informatika i Ee Primeneniya [Informatics and its Applications]*, 12(3):91–98.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. // *Computational Linguistics and Intellectual Technologies*, P 333–339.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A Unified MRC Framework for Named Entity Recognition. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 5849–5859.
- Fei Li, ZhiChao Lin, Meishan Zhang, and Donghong Ji. 2021. A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition. // *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 4814–4828.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 6008–6018, Online, November. Association for Computational Linguistics.
- Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Iliia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. NEREL: A Russian dataset with nested named entities, relations and events. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, P 876–885, Held Online, September. INCOMA Ltd.
- Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). // *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.
- Valerie Mozharova and Natalia Loukachevitch. 2016. Two-stage Approach in Russian Named Entity Recognition. // *International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, P 1–6.
- Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 task 2: Identification of informative COVID-19 English tweets. // *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, P 314–318, Online, November. Association for Computational Linguistics.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. // *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1946–1958, Vancouver, Canada, July. Association for Computational Linguistics.
- Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The Second Cross-lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages. // *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, P 63–74.
- Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. // *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, P 122–133, Kiyv, Ukraine, April. Association for Computational Linguistics.

- Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DAN+: Danish Nested Named Entities and Lexical Normalization. // *Proceedings of the 28th International Conference on Computational Linguistics*, P 6649–6662.
- Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. NNE: A Dataset for Nested Named Entity Recognition in English Newswire. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 5176–5181.
- Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2019. A Finnish News Corpus for Named Entity Recognition. *Language Resources and Evaluation*, P 1–26.
- Anatoly Starostin, Victor Bocharov, Svetlana Alexeeva, Anastasiya Bodrova, Alexander Chuchunkov, Stanislav Dzhumaev, Irina Efimenko, Dmitry Granovsky, Viktor Khoroshevsky, Irina Krylova, et al. 2016. FactRuEval 2016: evaluation of named entity recognition and fact extraction systems for Russian. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog”*, P 702–720.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. BRAT: a web-based tool for nlp-assisted text annotation. // *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, P 102–107.
- Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 5326–5331.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. "introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". // *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, P 142–147.
- Erik F. Tjong Kim Sang. 2002. "introduction to the CoNLL-2002 shared task: Language-independent named entity recognition". // *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.
- Natalia Vlasova, Elena Suleymanova, and Igor Trofimov. 2014. Report on Russian corpus for personal name retrieval. // *Proceedings of TEL’2014 Conference on Computational and Cognitive Linguistics*, P 36–40.
- Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 6470–6476.

Corpus experiment in forensic linguistics

Baranov A. N.

Russian Language Institute of RAS,
Moscow, Russia
Baranov_anatoly@hotmail.com

Abstract

The report deals with current trends in cases concerning protection of honor, dignity and business reputation, which lead to the situations when courts use the category «abuse of law». Linguistic criteria are proposed that make it possible to identify linguistic signs of abuse of law in the text. One of the criteria is a corpus experiment. The proposed corpus experiment allows to compare the frequency of use of evaluation markers, opinions, assumptions, probabilities, etc. with the frequency of use of these forms in the representative corpus of the Russian language. The criterion of the corpus experiment is supplemented by the semantic criterion and the metatext criterion.

Keywords: forensic linguistics; abuse of law; linguistic markers of abuse of law; corpus linguistics

DOI: 10.28995/2075-7182-2022-21-42-49

Корпусный эксперимент в лингвистической экспертизе

Баранов А. Н.

Институт русского языка РАН,
Москва, Россия
Baranov_anatoly@hotmail.com

Аннотация

В докладе рассматриваются современные тенденции в рассмотрении дел по защите чести, достоинства и деловой репутации, которые приводят к использованию судами категории «злоупотребления правом». Предлагаются лингвистические критерии, которые позволяют выявлять в тексте лингвистические признаки злоупотребления правом. Одним из критериев является корпусный эксперимент, в процессе которого частота использования маркеров оценки, мнения, предположения, вероятности и под. сравниваются с частотой использования этих форм в представительном корпусе русского языка. Критерий корпусного эксперимента дополняется семантическим критерием и критерием метатекста.

Ключевые слова: корпусная лингвистика; лингвистическая экспертиза; злоупотребление правом; корпусный эксперимент

1 Сфера корпусного эксперимента в лингвистической экспертизе

Лингвистическую экспертизу (далее «ЛЭ») в самом общем смысле можно определить как приложение лингвистических знаний, требующее оптимизации используемого концептуального аппарата, разумного ограничения объекта исследования и специфической формулировки выводов, учитывающей (1) требования законодательства, (2) конфликтность ситуации и (3) претензии сторон, в том числе на знание русского языка. Иными словами, результат ЛЭ – это особый жанр делового документа, в основе которого лежит научное исследование, выполненное – в идеале – по самым строгим требованиям лингвистики как науки. Это объясняется тем, что ЛЭ в судебном процессе является доказательством, которое может существенно влиять на решение суда¹.

¹ Разумеется, следует иметь в виду, что по закону в судебном разбирательстве ни одно из доказательств не имеет заранее предустановленной силы. Впрочем, в делах по защите чести, достоинства и деловой репутации, в делах по оскорблению, установлению авторства, плагиату и некоторых других ЛЭ является важнейшим и часто единственным доказательством.

Правовые ограничения не позволяют эксперту самому собирать доказательства – это прерогатива уполномоченных инстанций – следствия и суда. Между тем, в целом ряде случаев лингвист вынужден обращаться к разнообразным источникам информации – как минимум, к словарям и грамматикам, хотя для наблюдателя извне (в частности юриста) такая деятельность сопоставима со сбором доказательств или чем-то напоминает такую деятельность. Впрочем, существующая практика такова, что использование таких источников, как словари и грамматики не рассматривается как сбор доказательств и допускается при проведении ЛЭ.

Существуют, однако, и другие виды деятельности лингвиста-эксперта, которые сами по себе являются научным исследованием, хотя функционально носят «справочный» характер и могут подпадать под ограничение на сбор доказательств. Характерным примером такой работы лингвиста-эксперта является корпусный эксперимент, то есть обработка корпусов текстов с помощью существующего инструментария корпусной лингвистики. Юридический статус корпусного эксперимента не вполне определен, хотя реально этот метод исследования уже используется в практике ЛЭ. Корпусный эксперимент позволяет в ЛЭ решать, как минимум, следующие задачи:

- подтверждение, проверка и корректировка словарных значений слов и семантики синтаксических конструкций;
- выявление лексических значений, не представленных в словарных источниках;
- проверка употребительности аббревиатур, слов и словосочетаний в экспертизах по товарным знакам;
- оценка тональности контекстов использования слов и словосочетаний;
- составление репрезентативных фоновых корпусов текстов в экспертизах по установлению авторства;
- составление словников текстов и корпусов текстов, а также оценка частоты употребления соответствующих форм в корпусе.

Перечисленные задачи вполне очевидны, однако они не исчерпывают всех возможностей использования корпусного эксперимента в ЛЭ. Рассмотрим довольно сложный случай обращения к корпусным технологиям, обусловленный особенностями современного российского законодательства и практикой правоприменения, с одной стороны, и функционированием публичного дискурса – с другой.

2 Свобода выражения мнений и злоупотребление правом

Согласно существующим российским законам, граждане вправе свободно выражать свои мнения в публичном дискурсе – ст. 29 Конституции РФ². Единственное ограничение состоит в том, чтобы право на выражения мнения не затрагивало права других граждан. Эта весьма тонкая грань между правами граждан регулируется ст. 10 ГК РФ, в которой злоупотребление правом связывается с реализацией гражданских прав «исключительно с намерением причинить вред другому лицу, действия в обход закона с противоправной целью, а также иное заведомо недобросовестное осуществление гражданских прав» (ст. 10 ГК РФ).

Комментаторы этой статьи отмечают, что реализация данного положения законодательства сопряжена с апелляцией к оценочным категориям справедливости, морали, честности и порядочности, которые не определены в законе, что создает определенные риски субъективности в принятии судебных решений. Тем не менее, данная норма действует, причем количество дел, в которых она используется, неуклонно возрастает.

В последнее время в связи с очевидным изменением судебной практики по ст. 152 ГК РФ, связанной с защитой нематериальных благ – чести, достоинства и деловой репутации, данная категория стала использоваться в судебных решениях по указанной статье. Так, в Решении Люблинского районного суда г. Москвы от 31 мая 2017 г. по делу «Усманов против Навального» (№ 2-4145/2017) указывается: «судом учитывается, что распространение порочащих сведений в отношении строго определенных, конкретных лиц посредством придания соответствующим утверждениям намеренно-общего характера либо изложение таких утверждений в виде якобы рассуждений, личного мнения, оценки и т.п., а равно сознательная разбивка такого единого по

² Речь в данном случае идет о праве, а не о существующей практике правоприменения.

своей сути утверждения на несколько фраз и положений, представляет собой умышленное злоупотребление правом (с. 10 ГК РФ), в связи с чем не может служить основанием для отказа в иске об опровержении сведений, не соответствующих действительности, порочащих честь, достоинство и деловую репутацию»³.

Размывание существующей судебной практики по ст. 152 ГК РФ налицо: если в известном постановлении Пленума ВС РФ № 3 от 24 февраля 2005 г., фиксировавшего и утверждавшего судебную практику по рассматриваемой статье, обращалось внимание на форму сообщения негативных сведений и, соответственно, мнения, предположения, оценочные суждения и т.п. не попадали по действию данной статьи, то в Решении Люблинского районного суда по делу «Усманов против Навального» попытки участников публичной коммуникации выразить свое мнение, предположение, оценочное суждение и т.д. как минимум потенциально подпадают под санкции по ст. 152 ГК РФ. Важно здесь то, что никаких ясных критериев отделения случаев «придания соответствующим утверждениям намеренно-общего характера либо изложение таких утверждений в виде якобы рассуждений, личного мнения, оценки» от собственно личного мнения, оценки, рассуждения и под. не указывается. Вообще говоря, отсутствие таких критериев прямо противоречит ст. 29 Конституции РФ, поскольку она гарантирует свободу мысли и слова («Каждому гарантируется свобода мысли и слова»).

Оставим в стороне собственно юридические аспекты законов, подзаконных актов и их интерпретаций правоприменителями и обратимся к собственно лингвистическим аспектам проблемы, которая действительно возникла и требует лингвистического анализа в рамках ЛЭ по делам о защите чести, достоинства и деловой репутации.

3 Лингвистический и дискурсивный аспекты проблемы

До последнего времени соглашение между участниками публичной коммуникации и правоприменителями относительно легитимных способов выражения мнения, если отвлечься от некоторых тонкостей, сводилось, в сущности, к тому, что **эксплицитное** введение в дискурс пропозиции лексическими маркерами мнения, предположения, оценки и т.п. признавалось правомерным и не подпадало под действие ст. 152 ГК РФ. Это обеспечивало необходимый баланс между положениями Конституции РФ о свободе мнения и слова и правом гражданина на защиту своего доброго имени – чести, достоинства, деловой репутации и пр. Иными словами, фраза *Петров ограбил банк* в существующей судебной практике будет признана утверждением, а фразы *Видимо / вероятно / возможно* [и т.п. – группа 1], *считаю / думаю* [и т.п. – группа 2], *Петров ограбил банк* – предположением (маркеры группы 1) и мнением (маркеры группы 2).

Поток публикаций негативного характера, естественно, вызвал обратную реакцию со стороны тех, кто так или иначе затрагивался в соответствующих текстах и становился объектом пропагандистских кампаний. Определенную роль в изменении отношения к формальным и неформальным соглашениям сыграло и изменение общественно-политической ситуации. Между тем, обвинения в «злоупотреблении правом» в связи с делами о защите чести и достоинства небесспорны. Действительно, некоторые тексты современных СМИ обнаруживают признаки того, что называется злоупотреблением языком по Дж. Локку (*abuse of language* – [2], см. также [3]) в попытке уйти от правовых санкций. Очевидно, что необходимы критерии, которые могли бы послужить ориентирами при выявлении случаев «злоупотребления правом» в рамках дел указанной направленности. Важный объективный критерий такого рода – корпусный эксперимент, позволяющий сравнить относительную частоту употребления соответствующих форм в публичном дискурсе с относительной частотой аналогичных форм в спорной публикации.

³ Здесь и далее воспроизводятся особенности орфографического и пунктуационного оформления оригинала.

4 Корпусный эксперимент: критерий частотности

Назовем для краткости статьи, которые, предположительно, содержат лингвистические признаки злоупотреблением правом в связи со ст. 152 ГК РФ «ангажированными» или А-статьями/публикациями, а соответствующий дискурс – «ангажированным», или А-дискурсом. Характерный признак публикаций указанного типа – злоупотребление эксплицитными маркерами выражения мнения, предположения, оценки, возможности и т.п. Разберем характерный пример.

В статье «Миллионы чиновников МИНОБРА в оффшорах?»⁴, бывшей объектом ЛЭ, использовались следующие лексические маркеры мнения, предположения, вероятности, оценки и т. п.: *сомнительно* 3, *наверное* 3, *видимо* 1, *предположительно* 1, *вероятно* 1, *можно сказать* 1, *может быть* 2, *кажется* – 5; всего – 17 вхождений. Поскольку в статье насчитывалось 998 словоупотреблений, то относительная частота указанных форм в совокупности составляет, приблизительно 0,017. Легко подсчитать среднюю частоту появления данных лексических единиц в представительном корпусе текстов. В «Корпусе текстов по современной публицистике» (60-е г. XX в. – начало XXI в.), созданном в отделе экспериментальной лексикографии Института русского языка РАН, насчитывается 29 225 404 словоупотреблений. Указанные единицы распределены в нем следующим образом: *сомнительно* 129, *наверное* 5147, *видимо* 5272, *предположительно* 273, *вероятно* 2346, *можно сказать* – 1399, *может быть* – 11272, *кажется* – 6811; всего – 32 649 вхождений. Тем самым, средняя частота маркеров мнения и предположения для корпуса публицистики оказывается равной 0,0011. Иными словами, в рассматриваемой статье относительная частота употребления слов или словосочетаний, эксплицитно указывающих на выражение мнения, предположения, возможности и пр. более чем 15 раз ($0,017/0,0011 = 15,45$) выше, чем в среднем по корпусу публицистики. Это можно считать серьезным аргументом в пользу отнесения анализируемой публикации к сфере А-дискурса.

В газетном корпусе Национального корпуса русского языка (НКРЯ) совокупный частотный ранг выражений указанного типа еще ниже: *сомнительно* 869, *наверное* 43054, *видимо* 34940, *предположительно* 12263, *вероятно* 21170, *можно сказать* – 12005, *может быть* – 103070, *кажется* – 52438; всего – 210458 вхождений⁵. Мощность газетного корпуса – 394 473 204 словоупотребления. Таким образом, средняя частота указанных маркеров мнения и предположения для газетного корпуса оказывается равной 0,00053.

Легко видеть, что в обсуждаемой статье относительная частота употребления указанных слов и словосочетаний, эксплицитно передающих мнение, предположение, возможность и пр., в 32 раза ($0,017/0,00053 = 32$) выше, чем в среднем по корпусу публицистики. Отнесенность спорной публикации к А-дискурсу в сопоставлении с данными цифрами не вызывает сомнений.

5 Семантический критерий и корпусный эксперимент

Критерий частотности для повышения достоверности результата должен быть дополнен другими критериями, указывающими – пусть и косвенно – на коммуникативное намерение автора текста любыми средствами донести негативное содержание до адресата и избежать правовых санкций. Авторы А-публикаций в стремлении обезопасить себя лишены чувства меры, что приводит к стилистическим и семантическим ошибкам в сочетаемости. Укажем некоторые типичные случаи.

Предположение в сфере действия модальности факта. В примерах, иллюстрирующий данное нарушение, в семантическую сферу действия выражения с семантикой факта попадает слово или словосочетание с семантикой предположения, ср. (1).

- (1) *Назовем вещи своими именами*⁶: сам Рыбкин *скорее всего* и передал своим юристам поддельные векселя для обращения в суд. А если это так, то именно Александру Рыбкину предстоит ответить за подделку документов.

⁴ Здесь и далее названия некоторых статей по понятным причинам изменены.

⁵ По состоянию газетного корпуса на середину марта 2022 г.

⁶ Здесь и далее выделение автора.

В приведенном примере выражение *называть вещи своими именами* передает семантику факта. Действительно, странно выглядит фраза: *Назовем вещи своими именами: он **возможно** / **вероятно** / **скорее всего** настоящий убийца* при норме *Назовем вещи своими именами: он и есть настоящий убийца*. Иными словами, введение в сферу действия модальности факта выражений, передающих семантику предположения, семантически некорректно.

Данные НКРЯ подтверждают этот вывод. Во всех 324 примерах употребления выражения *назвать вещи / всё своими именами*, найденных в корпусе в различных формах, в семантической сфере действия обсуждаемой идиомы обнаруживается пропозиция (и связанная с ней номинация), которая представляется говорящему истинной (подчеркнуто в примерах), а модальностей предположения, вероятности нет: *Ее заботило мое отношение к его, **назовем вещи своими именами**, **тунеядству**; **Агрессоры** (**назовем вещи своими именами**); **взяточдатели** (**назовем вещи своими именами**) неохотно распространяются о том, кому пришлось сунуть; Вкладчики не хотят мириться с очередным грабежом и **называют вещи своими именами**. Они открыто обвиняют ЦБ в рейдерском захвате своего банка под предлогом «оздоровления» банковской системы; Пришлось **называть вещи своими именами** и признаться в том, что администрация Варьете, в лице директора, финдиректора и администратора, пропала и находится неизвестно где, что конференсье после вчерашнего сеанса был отвезён в психиатрическую лечебницу <...>. [М. Булгаков. Мастер и Маргарита]. В последнем примере из романа «Мастер и Маргарита» использование глаголе *признаться* подчеркивает истинность вводимой пропозиции.*

Ср. аналогичный пример из текста выступления С. Доренко времен Лужкова:

- (2) Факты говорят о том, что, возможно, совершенно не случайно Ковалев оказывал Лужкову услугу, чтобы спрятать концы в воду с убийством Пола Тейтума.

Как минимум, фраза (2) стилистически не вполне корректна – в первую очередь из-за того, что в сферу действия факта (*Факты говорят о том*) попадает предположение (*возможно*), которое тут же сменяется подчеркиванием истинности: *совершенно не случайно*.⁷

Мнение в сфере действия нереализованного условия. Для примеров рассматриваемого типа характерна ситуация, когда используется условная конструкция с союзом *если* или аналогичной формой, а далее следует рассуждение, игнорирующее тот факт, что само условие не верифицировано и описывает гипотетическую ситуацию:

- (3) Ведь, *если предположить*, что хозяин алюминиевого комбината имеет компромат на судью, *следовательно и получается*, что суд принимает решение, в разы увеличивающее активы олигарха.

В примере (3) словоформы *следовательно и получается* используются для введения пропозиции ‘суд принимает решение, в разы увеличивающее активы олигарха’ как мнения (экспликация процесса рассуждения квалифицируется в ЛЭ по ст. 152 ГК РФ как мнение). Однако мнение оказывается в сфере действия нереализованного условия и предположения одновременно: *если предположить, что хозяин алюминиевого комбината имеет компромат на судью*, что хотя и не нарушает семантическую сочетаемость в сильном смысле, но стилистически не вполне удачно.

В данном случае косвенная проверка полученного вывода в корпусном эксперименте частично возможна. Словосочетание *следовательно и получается* в НКРЯ не обнаруживается⁸.

Чрезмерная детализация гипотетической ситуации. Злоупотребление маркерами модальностей может влечь нарушения прагматической природы. Например, обсуждение гипотетической ситуации в обычном случае не предполагает излишних подробностей, поскольку ситуация не реализовалась. Действительно, нереализованная ситуация не может быть известна говорящему в деталях. Между тем, формальная расстановка маркеров с целью ухода от правовых рисков приводит к нарушению этого естественного ограничения, ср. пример (4):

⁷ См. также разбор этого примера в [1: 134], однако с другой целью – как пример использования приема речевого воздействия.

⁸ Последовательность *следовательно, получается* представлена в трех примерах, однако она имеет другую семантику. Ср.: *Следовательно, получается сумма в размере 30 000 руб.*

- (4) **Вероятно**, Литвинов постарается донести до руководства «Росмостов» необходимость возврата Куропаткина, минуя совет директоров. Второй **сценарий**: назначить Куропаткина исполняющим обязанности, пока совет директоров не начнет задавать вопросы о временщиках в руководстве. Поэтому, **вероятно**, Литвинову **придется буквально вымолить у Буданова** место для Куропаткина для личного самосохранения. [«Правда и истина»]

Лексемы *вероятно* и *сценарий* указывают на гипотетические варианты развития ситуации, а грамматическая форма будущего времени (*постарается, начнет, придется*) вообще снимает вопрос о верификации описываемых событий – она невозможна, а неverified высказывания в соответствии с Постановлением Пленума Верховного Суда РФ № 3 от 2005 г. не являются утверждениями в смысле ст. 152 ГК РФ. Однако указание на то, что *Литвинову придется буквально вымолить у Буданова место для Куропаткина для личного самосохранения* семантически аномально, поскольку экстраполяция в будущее таких подробностей отражает стремление автора статьи представить нынешнюю ситуацию угрожающей для Литвинова. Корпусная проверка и подтверждение данного вывода на большом материале также затруднена, поскольку предполагает использование сложного семантического парсера, учитывающего к тому же и прагматические характеристики описываемой ситуации.

Ложная анти-эмпатия. К нарушениям семантического характера следует отнести также контексты, в которых при общем отрицательном отношении к обсуждаемому герою в публикации обнаруживается цитирование других лиц крайне негативного свойства, причем автор от цитаты навязчиво открещивается, ср. (5):

- (5) «Все это происходит при покровительстве службы безопасности под руководством господина Курьянова, который в угоду Литвинову **расправляется** с неугодными сотрудниками. Все, кто хоть немного оказывает сопротивление преступной деятельности, **подвергаются “наезду”** со стороны службы безопасности либо **увольняются под давлением**. На основании вышеперечисленного можно сделать выводы о том, что группа Литвинов–Куропаткин **активно занимается хищением денежных средств компании в особо крупных размерах**», – берутся рассуждать авторы письма.

В приводимом фрагменте цитируется письмо, в котором Литвинов характеризуется крайне отрицательно, что в полной мере соответствует направленности публикации, при этом автор статьи в словосочетании *берутся рассуждать* в явном виде отгораживается от цитаты и даже ставит ее под сомнение, что в сочетании с негативной тональностью статьи в целом смотрится довольно странно.

К сожалению, и в данном случае корпусный эксперимент на значительном материале в настоящее время вряд ли возможен из-за сложной комбинации характеристик тональности текста.

Использование будущего времени. В рамках ст. 152 ГК РФ к числу утверждений не относятся высказывания, которые не могут быть проверены на соответствие действительности. Тем самым, использование грамматического будущего времени исключает высказывание из сферы санкций по указанной статье. В некоторых случаях в А-статьях используется будущее время, причем события, предсказываются с исключительной подробностью, что сближает данный случай с «чрезмерной детализацией гипотетической ситуации», рассматривавшейся выше:

- (6) В то же время, отмечают эксперты, предприятие традиционно ссылается на нехватку средств, однако по информации источников, близких к «Россетям», «Екатеринбургская электросетевая компания» (ЕЭСК) по всей видимости не сможет освоить средства по инвестпрограмме в 2016 году, что может привести к ее значительному секвестру в 2017 году. Как следствие, вполне вероятно, что «компанию признают банкротом и продадут».

В приведенном примере подробно описывается цепочка «будущих» событий: компания не сможет освоить средства → компанию ждет секвестр → компанию признают банкротом и продадут. Такое ясновидение имеет совсем не волшебную природу и объясняется желанием сообщить негативные сведения о компании и одновременно избежать санкций по ст. 152 ГК РФ. Как пример

злоупотребления языком (abuse of language) следует рассматривать и использование кавычек во фразе «компанию признают банкротом и продадут», поскольку источник цитаты не указан, *эксперты* не раскрыты.

Корпусная проверка степени детализации описания ситуации в контексте будущего времени существенно затруднена.

Искусственная модификация иллокутивной семантики. Семантическая некорректность текста в А-статьях может быть связана с неудачными попытками преобразования грамматических утверждений в вопросы. Так, название статьи *Денег нет, но вы держитесь? Оффшорные миллионы чиновников Минобороны?* выглядит довольно странным из-за вопросов, которые нарушают условия успешности данного типа речевого акта. В статье указывается, что миллионы у чиновников есть и что их вполне достаточно – чиновникам «держаться» не надо. Заголовок статьи в виде общего вопроса в такой ситуации выглядит неестественно, но объяснимо, если иметь в виду попытку ухода от правовой ответственности.

Аналогичный заголовок статьи, разбиравшейся выше, также сформулирован в виде вопроса: *Миллионы чиновников МИНОБРА в оффшорах?* И в этом случае уместность такого заголовка вызывает серьезные сомнения.

По понятным причинам и в этом случае корпусный эксперимент затруднен.

6 Чрезмерная «защита» от правовых рисков

Как известно, семантическая информация распределяется в тексте по собственно тексту и метатексту [4] – различным комментариям автора, управляющим вниманием читателя (адресата). Во многих случаях А-статьи содержат эксплицитные метатекстовые элементы, формально указывающие на мнение, предположение и оценочное суждение автора, ср. (7):

- (7) И если так, то некоторым не очень бедным руководителям комбината придется вскоре изучать не инструкции министерства, а Уголовный Кодекс. Или объяснить загадочное происхождение миллионов евро. И очень сомнительно для меня, и *это мое журналистское мнение*, что это случайное совпадение.

Фраза *это мое журналистское мнение* носит метатекстовый характер и опять-таки выводит текст из-под санкций по ст. 152 ГК РФ. Злоупотребление в данном случае состоит в навязчивости эксплицитной формы метатекстового комментария *это мое журналистское мнение*, сочетающейся с будущим временем описываемых событий и нереализовавшемся условием (*и если это так*). Нагромождение различных речевых способов защиты от правовых санкций – чрезмерная защита – определенно указывает на злоупотребление правом.

Отметим, что метатекст как критерий не всегда дает однозначный результат. Так, в известном деле «Уткин-Советский спорт» спортивный журналист в явном виде рассказывает о том, что он консультировался у юристов, как писать о договорных матчах, чтобы не попасть под правовые санкции:

- (8) Я решил проконсультироваться с хорошим адвокатом и даже потратил на консультацию сумму, равную гонорарам за несколько заметок в «Советском спорте». Нисколько не жалею об этом. Я узнал, что высказывание Валерия Георгиевича неподсудно, поскольку представляет собой **МНЕНИЕ**. *Предполагать, размышлять, даже фантазировать* каждый человек волен как ему угодно.

И далее журналист продолжает:

- (9) Это наша победа, коллеги. Мы были влекомы каким-то непонятным политкорректным чувством, мы с вами, чтобы не употреблять слово «договорняк», придумывали разнообразные экивоки. Говорили «странные матчи», «двусмысленные», мы с вами невероятно изгалялись. Так вот! Мы теперь можем говорить и «договорные», и «проплаченные». *Если мы так думаем – имеем полное право.*

Рассматривать фрагменты (8) и (9) как случай злоупотребления правом вряд ли правомерно по той простой причине, что в этой публикации последовательно обсуждается проблема того, как журналист может выражать свое мнение о том, что он считает нарушением этических норм – пусть и недоказанными в судебном порядке.

Проверка случаев нагромождения лингвистических защит в корпусном эксперименте на данном этапе развития корпусной лингвистики недостижима, хотя сравнение частоты использования элементов метатекста в спорной статье и в представительном корпусе вполне возможно – см. выше раздел 5.

Сформулированные лингвистические признаки «злоупотребления правом» в делах по защите чести, достоинства и деловой репутации по большей части не являются однозначными показателями рассматриваемого феномена. Иными словами, необходима проверка исследуемого материала по комплексу лингвистических параметров злоупотребления. Корпусный эксперимент оказывается в этом случае важнейшим инструментом объективизации интуиции лингвиста-эксперта и весомым доказательством «злоупотребления» словами.

References

- [1] Baranov A.N. (2011), Forensic linguistics. Theoretical foundations and practice [Lingvisticheskaya ekspertiza teksta. Teoreticheskiye osnovaniya i praktika]. Moscow — 592 p.
- [2] Lock J. Of the abuse of words. — London: Penguin books, 2009 — 161 p.
- [3] Kress G., Hodge R. Language as ideology. — Routledge, 1979 — 182 p.
- [4] Wierzbicka A. (1978), Metatext in text [Metatekst v tekste], New in foreign linguistics [Novoye v zarubezhnoy lingvistike], Vol. 8, Moscow, pp. 402-421.

Proper names in the scope of metalinguistic negation

Igor M. Boguslavsky

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

Elena L. Vilinbakhova

St. Petersburg State University,
St. Petersburg, Russia
e.vilinbakhova@spbu.ru

Abstract

This paper analyses constructions with proper personal names negated metalinguistically, i.e. the objection is aimed not at the truth of the proposition (*ne Vasja, a Kolja*), but at the assertability of a linguistic expression (*ne Vasja, a Vasilij Pavlovič*). Our data from the Russian National Corpus shows, first, that metalinguistic negation constructions with proper names are used primarily to correct the representation of the referent which is associated with the particular form of naming, and the cases of correction of form only, such as linguistic mistakes, are by far less typical. Next, among the additional elements there is a prevalence of temporal markers, which indicates that metalinguistic negation constructions are commonly used to point to the change of the referent with time. Another observation is that the correction towards a more formal way of naming prevails over the opposite direction, showing that the lack of formality is regarded as a more serious mistake than the other way around. Finally, it turned out that in our sample the examples of self-correction prevail over the correction of others, which demonstrates a wider range of functions of metalinguistic negation than it was acknowledged in the previous studies.

Keywords: metalinguistic negation, proper names, Russian, corpus study

DOI: 10.28995/2075-7182-2022-21-50-60

Имена собственные в сфере действия метаязыкового отрицания¹

Богуславский И. М.

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

Вилинбахова Е. Л.

Санкт-Петербургский
государственный университет,
Санкт-Петербург, Россия
e.vilinbakhova@spbu.ru

Аннотация

Феномен метаязыкового отрицания характеризуется тем, что в его сфере действия находится не пропозиция, а способ ее языкового выражения. Мы рассматриваем конструкции, в которых в сфере действия метаязыкового отрицания оказывается имя собственное. На основании данных Национального корпуса русского языка можно сделать следующие выводы. Во-первых, основной функцией конструкции является коррекция информации, кодируемой в способе представления носителя имени, а не просто исправление ошибки в передаче этого имени. Во-вторых, распространенность временных маркеров в качестве дополнительных элементов конструкции указывает на использование метаязыковой конструкции для указания на изменение референта с течением времени. Далее, данные корпуса свидетельствуют, что говорящий использует конструкцию для коррекции в сторону более официальной формы имени чаще, чем наоборот, что, по-видимому, говорит о том, что недостаточный уровень формальности

¹ Исследование выполнено за счет гранта Российского научного фонда (проект № 19-78-10048).

осуждается в коммуникации больше, чем ее переизбыток. Также было выявлено, что случаи исправления говорящим самого себя преобладают над случаями коррекции другого, что позволяет говорить о более широком спектре употребления конструкции, чем утверждалось в предшествующих исследованиях. Наконец, было показано, что метаязыковой предикат названия может реализовываться не только лексически, но и входить в состав значения синтаксических конструкций.

Ключевые слова: метаязыковое отрицание, имена собственные, русский язык, корпусное исследование

1 Введение

Наше исследование посвящено проблеме метаязыкового отрицания в конструкциях с именами собственными, как в (1), где говорящий исправляет форму имени, выбранную для обращения к Василию Аксенову.

- (1) *Они много выпили, стали на «ты», но, когда Андрей назвал Аксенова Васей, сидевший рядом кто-то из пишущей братии по фамилии Холендро (фамилия необычная, поэтому и запомнил), строго сказал: — Не Вася, а Василий Павлович! — И Андрей сразу отрезвел².*

Феномен метаязыкового отрицания интересен тем, что в его сфере действия оказывается не пропозиция, а способ ее языкового выражения, см. (2): речь идет не о том, что имеет место не одно действие, а другое (как в случае *Не спим, а работаем*), а о том, что данное действие правильно называть не одним способом, а другим.

- (2) *Мы не пьем, а дезинфицируем душевные раны.* [Богуславский 2016: 126]

В настоящее время метаязыковое отрицание активно изучается в мировой лингвистике, однако большая часть работ выполнена на материале английского [Horn 1985, 1989; Burton-Roberts 1989; Carston 1996; Geurts 1998] и французского языков [Larrivée 2018; Moeschler 2018]. В работах на материале русского языка можно встретить анализ отдельных примеров, которые содержат метаязыковое отрицание, см. [Богуславский 1985; Падучева 2011; Escandell-Vidal, Vilinbakhova 2019], а также [Азаргаева 2018].

В литературе отмечается, что метаязыковое отрицание обладает особыми свойствами, которые отличают его от дескриптивного отрицания. К ним относится его особый интонационный контур, невозможность морфологического выражения или сочетания со словами с отрицательной поляризацией (NPI), употребление прежде всего в реактивных репликах, воспроизводящих предшествующее высказывание (1), см. [Horn 1989; van der Sandt 1991; Carston 1996].

Также Р. ван дер Сандт, Б. Гертс и Ж. Мешлер предлагают классификации различных типов метаязыкового отрицания, выделяя случаи, где оно направлено на форму, см. (1-2), выше; на импликацию (*не хорошо, а замечательно*); и на пресуппозицию (*жена не уйдет к другому, если у вас нет жены*) [van der Sandt 1991; Geurts 1998; Moeschler 2018]. Примеры, рассматриваемые в нашей работе, относятся к первой группе, в которой отрицание направлено на форму.

Следует отметить, что примеры с метаязыковым отрицанием являются частным случаем более широкого класса метаязыковых высказываний, см. (3а-б). Различие между ними состоит в том, что в (3а) метаязыковое употребление эксплицируется с помощью глагола речи, а в (3б) оно ничем не маркируется.

- (3) а. *Mary says [potaytou] and John says [potahto].* [Carston 1996: 327]
 б. *You like [potaytouz] and John likes [potahtoz].*

Метаязыковые высказывания о способах наименования включают метаязыковые предикаты названия (*звать, называть, именовать, обозначать* и др.), которые также могут быть выражены эксплицитно или имплицитно.

² Примеры взяты из Национального корпуса русского языка, если не указано иное. Из соображений краткости мы не приводим в тексте источники.

- (4) а. *Моего брата зовут Дмитрий.* [*Дмитрий*₁ = имя] [4а-в сконструированные примеры]
 б. *На пороге стоял Дмитрий.* [*Дмитрий*₂ = индивид, чье имя *Дмитрий*₁]
 в. *Моего брата зовут не Димочка, а Дмитрий.*
 г. *На пороге стоял не Димочка, а Дмитрий. От прежнего милого юноши ничего не осталось.*

Хотя примеры типа (4в) с эксплицитным метаязыковым предикатом не относятся к прототипическим случаям метаязыкового отрицания, такими как (4г), они разделяют с ними важные свойства: (а) отрицание направлено на языковое выражение, а не на внеязыковой объект (как в случае *Пришел не Петя, а Коля*), и (б) корректируемое и корректирующее выражение имеют один и тот же референт. Поэтому для наших целей удобно анализировать их совместно с чистыми случаями метаязыкового отрицания.

Мы рассматриваем (i) функции метаязыковых отрицательных конструкций с именами собственными; (ii) особенности конструкций с именами собственными в отличие от конструкций с другими синтаксическими группами; (iii) эксплицитное и имплицитное выражение метаязыковых предикатов называния.

Изложение построено следующим образом. В разделе 2 представлена методология исследования, в разделе 3 обсуждаются его результаты, в Заключение подводятся итоги работы.

2 Методология

Материал для нашего исследования собирался по основной коллекции Национального корпуса русского языка (далее – НКРЯ). Брались тексты от начала XX века и до сегодняшнего дня. В корпусе искали вхождения конструкции *не X, а Y*, где X – имя собственное, называющее лицо, а Y – произвольное существительное (часто – также имя собственное). После ручной фильтрации «мусора» осталось 318 фрагментов текста, которые и составили предмет нашего исследования. Никаких ограничений на жанр текста не накладывалось, но оказалось, что подавляющее большинство примеров приходится на художественную литературу (64,77%) и публицистику (32,07%). Остальные жанры представлены изолированными примерами.

Отобранные фрагменты были размечены по двум структурным и четырем содержательным параметрам: (i) заполнение слотов конструкции; (ii) дополнительные элементы; (iii) цель коррекции; (iv) изменение уровня формальности; (v) смена имени; (vi) исправляемый участник. Выбор параметров был обусловлен как наблюдениями в предшествующих работах о конструкциях с именами нарицательными (например, параметр «Исправляемый участник»), так и спецификой материала в лингвистическом и шире, социокультурном аспектах, включая разнообразие форм имени собственного как средств для выражения социальных отношений между участниками общения в русском языке (параметр «Изменение уровня формальности») или наличие традиций, связанных со сменой имени, в частности, при вступлении в брак (параметр «Смена имени»).

3 Описание результатов

3.1 Структурные свойства

Хотя мы изучаем только конструкцию *не X, а Y*, существует и другой вариант *Y, а не X*. Эти варианты демонстрируют легкое различие в коммуникативной перспективе, но в остальном полностью синонимичны. Близкая коммуникативная вариативность имеется и в английском языке, однако английское соответствие союза *a – but* – не допускает симметричного построения и выступает только в одном из двух вариантов [McCawley 1991]:

- (5) а. *John drank not coffee but tea.* [McCawley 1991: 190]
 б. *John drank tea, (*but) not coffee.*

Мы остановимся на двух типах структурных свойств конструкции *не X, а Y* – на том, чем именно могут заполняться позиции X и Y, и на том, какие дополнительные элементы могут в ней участвовать.

В Табл. 1 представлено распределение возможных способов заполнения позиций *X* и *Y*. Наиболее частотными являются построения, в которых обе позиции замещаются именами собственными (80,8%):

(6) *Тогда он был еще не Юрка, а Юрий Николаевич.*

Следом идут конструкции типа (7), в которых в позиции корректируемого элемента *X* находится имя собственное, а в позиции замещающего элемента *Y* – имя нарицательное (10,4%), а также конструкции типа (8), где корректирующий элемент включает как имя собственное, так и нарицательное (6%):

(7) *Не Валер, а товарищ старший сержант.*

(8) *Ты, видно, не в курсе, что я не просто Крячко, а полковник милиции Крячко.*

Другие варианты корректируемого и замещающего элементов встречаются заметно реже.

Заполнение слотов конструкции	Количество примеров	Относительная частотность (%)
<i>не N.Prop.n, а N.Com</i>	33	10,377
<i>не N.Prop.n, а N.Com N.Prop.n.</i>	19	5,974
<i>не N.Com N.Prop.n, а N.Com N.Prop.n.</i>	8	2,516
<i>не N.Com N.Prop.n, а N.Prop.n.</i>	1	0,314
<i>не N.Prop.n, а N.Prop.n.</i>	257	80,818
Всего	318	100

Таблица 1: Заполнение слотов конструкции.

Широкая представленность конструкций, в которых оба компонента выражены именами собственными, вполне предсказуема. Изучаемый массив образован конструкциями, в которых имена собственные вступают в контрастные отношения с другой номинацией, и естественно, что эти номинации семантически параллельны [Umbach 2004]. Менее ожидаемым оказалось то, что непараллельных примеров типа (7) в массиве тоже оказалось достаточно много. Тем не менее, это вполне объяснимо: коммуникативная задача говорящего состоит в том, чтобы сообщить, что выбор номинации в виде имени собственного неудачен, и альтернативой является более информативная номинация, указывающая на роль референта в ситуации, его статус, должность и т.п.

Помимо центральных компонентов конструкции – корректируемого и замещающего элементов – конструкция может содержать разного рода дополнительные элементы, вносящие темпоральные, модальные, эпистемические и некоторые другие компоненты значения. Здесь нас интересуют только такие элементы предложения, которые воздействуют на саму описываемую конструкцию или ее компоненты. Так, темпоральные элементы сообщают, что коррекция номинации мотивирована изменением ситуации: раньше была уместна номинация *X*, а теперь – номинация *Y* (*фамилия его уже не Гриншпан, а Громов*). Интенсификаторы, которые мы принимаем во внимание, воздействуют непосредственно на отрицание (*никакой не Иван, а Иоанн Суровый*). Фокальные частицы помещают в фокус корректирующий компонент (*не Елена, а именно Лена*). Свободный дательный, как будет показано ниже, также самым непосредственным образом связан со значением конструкции.

Дополнительные элементы можно распределить по следующим тематическим группам:

- **ТЕМПОРАЛЬНЫЕ МАРКЕРЫ** (*уже, еще, сейчас, теперь, тогда, отныне* и др.)

(9) *Вскоре он оказывается в Ленинграде, только зовут его теперь не Владимир Иосифович, а Владимир Осипович, и фамилия его уже не Гриншпан, а Громов.*

- ИНТЕНСИФИКАТОРЫ (*вовсе, никакой*)
- (10) *И мужик этот никакой не Иван, а Иоанн Суровый.*
- ФОКАЛЬНЫЕ ЧАСТИЦЫ (*именно, только, прямо, просто*)
- (11) *12 марта прилетела Лена Петрова (имя – по паспорту, не Елена, а именно Лена, в честь реки).*
- ВВОДНЫЕ ЭЛЕМЕНТЫ (*кажется, скажем*)
- (12) *Я знал одного Жеглова ... хотя тот, кажется, не Жеглов, а Жигалов... такая жалость.*
- РИТОРИЧЕСКИЙ ДАТЕЛЬНЫЙ (*тебе, вам*)
- (13) *Я вам не Коля, - грозно-равнодушно кричит он, - а Николай Иванович.*

Всего в нашем массиве дополнительные элементы встретились в 127 примерах, и их общее число составило 153. В Табл. 2 представлены их абсолютные частоты, вес относительно других элементов (где 100% составляют все 153 обнаруженных элемента) и распределение по массиву (где 100% составляют 318 проанализированных примеров). Мы включили в таблицу только те элементы, которые встретились больше трех раз.

Дополнительный элемент	Количество примеров	Доля среди всех доп. элементов (%)	Доля предложений с этим элементом (%)
УЖЕ	29	18,954	9,119
ВОВСЕ	18	11,764	5,660
ПРОСТО	14	9,150	4,402
ТЕПЕРЬ	12	7,843	3,773
ИМЕННО	7	4,575	2,201
НИКАКОЙ	7	4,575	2,201
ТЕБЕ/ВАМ	5	3,267	1,572
ДАЖЕ	4	2,614	1,257
УЖ	4	2,614	1,257

Таблица 2: Дополнительные элементы, встретившиеся больше трех раз.

Как видно из таблицы, среди дополнительных элементов с большим отрывом лидирует темпоральный маркер *уже* (18,9% случаев). Это неудивительно: употребление метаязыкового отрицания часто мотивировано тем, что с течением времени изменился сам носитель имени, а это в свою очередь привело к тому, что изменился и способ его уместной номинации. Ср., например, (14):

- (14) *И уже давно еврейский мальчик Эдик – не Эдик, а Эд, профессор Массачусетского Технологического.*

Дополнительные элементы, о которых мы говорили выше, употребляясь в нашей конструкции, не обнаруживают никакой специфики. Их значение, темпоральное, модальное, вводное и т.п., остается таким же, каким оно бывает и в других контекстах. Иначе обстоит дело с так называемым «свободным дательным» (*free, non-valence, extended, sentence dative* [Horn 2013]), на котором мы остановимся несколько подробнее.

Хорошо известно, что в языках мира представлен широкий спектр употреблений дательного падежа, далеко выходящих за рамки простой маркировки аргументов предикатов [Lambert 2010].

Сведения о таких употреблениях в разных языках можно найти, в частности, в [Hole 2006; Hole et al. 2006; Bosse et al. 2012; Pennington 2012]. Для русского языка детальный разбор подобных употреблений содержится в статье «Дательный падеж», подготовленной для Русской корпусной грамматики [Кустова 2012]. Для настоящей работы релевантно употребление, которое называют экспрессивный или «риторический» дательный [Кустова *ibid*: 2.3.4.3]. Он разделяет главное свойство всех случаев свободного дательного — маркирует объект, который лексически не предсказывается глаголом и не входит в его пропозициональное содержание [Gutzmann 2007]. Этот объект не включен в процесс, называемый глаголом, но тем не менее так или иначе им затронут [Lamigoу, Delbecque 1998]. Экспрессивный дательный выполняет в высказывании риторическую роль — он некоторым образом затягивает в ситуацию адресата. «Функция риторического дательного состоит в том, чтобы включить в предложение, описывающее некоторую ситуацию ..., некое лицо — обычно участника речевого акта, ... и тем самым сделать это лицо причастным к описываемой ситуации, к которой оно в действительности не имеет отношения в том смысле, что не является ее участником» [Кустова 2012: 2.3.4.3]. Ср. примеры (15)–(17) из [Кустова *ibid*]:

- (15) *Да, Вова, это тебе не Бухара!*
 (16) *Будет он тебе на «Жигулях» ездить!*
 (17) *Он тебе не Дима, запомни, а Дмитрий Вениаминович.*

Мы хотели бы показать, что примеры типа (17) стоят в этой группе несколько особняком. В (15-16) форма *тебе*, действительно, ничего не вносит в пропозициональное содержание предложений *Это не Бухара; Будет он на Жигулях ездить!* Адресат не является аргументом никакого предиката в этих предложениях и никаким образом лексически не предсказывается ими. Для (17) это не так. Значение этого предложения можно представить как ‘ты не должен называть его Дима; ты должен называть его Дмитрий Вениаминович’. Здесь видно, что адресат *ты* существенным образом участвует в пропозициональном содержании предложения. Он является полноценным субъектом пропозиций ‘ты называешь его Дима / Дмитрий Вениаминович’. Ср. еще один характерный пример, отчетливо проявляющий, что дативная ИГ является субъектом называния:

- (18) *Он неловко начал говорить — запинаясь, с повторами, — и вдруг я понял, что и впрямь они с Павлом давно были знакомы и дружны, и отчество он путал как раз потому, что Павел ему был именно что Павел или даже Пашиа, а вовсе не Павел Иванович, а сам он Павлу — никакой не Евгений Александрович, а просто Женька.*

Предложения (15) и (17) выглядят похожими, но в действительности между ними имеется существенная семантическая разница. В обоих предложениях отрицание воздействует на имя собственное, но по-разному. Если в (17) речь идет о том, какое именование уместно/неуместно, то (15) не означает ‘ты не должен называть это место Бухара’. Скорее, это предложение сообщает, что данное место не похоже на Бухару, и одновременно затягивает в эту ситуацию слушающего, хотя он и не имеет к ней непосредственного отношения.

Примечательной особенностью предложений типа (17-18) является то, что предикат именования не выражен лексически. Он входит в состав значения всей конструкции, включающей связочный глагол (часто в нулевой форме), ИГ в дательном падеже или с предлогом *для* и имя собственное.

- (19) *Она всем была Галина Ивановна, а мне просто Галя.* [сконструированный пример]

Стоит подчеркнуть, что предикат именования возникает в конструкции с дательным падежом, по-видимому, только в контексте имени собственного. Например, если конструкция содержит существительное, обозначающее отношение, то местоимение в дательном падеже обозначает аргумент этого отношения: *Ты мне не начальник* ‘ты не являешься моим начальником’ (но не ‘я не называю тебя начальником’).

Дополнительным свидетельством того, что дативная ИГ имеет разную семантику в металингвистических предложениях типа (17-19) и в чисто риторических предложениях типа (15-16), является то, что в первом случае тебе взаимозаменяемо с предложной группой *для тебя*, а во втором случае это исключено:

- (17') *Он для тебя не Дима, запомни, а Дмитрий Вениаминович.*
 (19') *Она для всех была Галина Ивановна, а для меня просто Галя.*
 (15') **Это для тебя не Бухара!*
 (16') **Будет он для тебя на «Жигулях» ездить!*

Как подчеркивалось в [Lambert 2010; Horn 2013] и других работах, употребления «свободного дательного» образуют континуум. Здесь мы столкнулись с другим проявлением этого обстоятельства. Как отмечалось выше, важным свойством «свободного дательного» является то, что он не входит в пропозициональное содержание предложения. С этой точки зрения, предложения типа (17) выглядят несколько парадоксально. Если исходить из того, что пропозициональное содержание предложения (или, его предикатно-аргументная структура) образована исключительно его словами, то тогда дативные ИГ в примерах типа (17-19) не входят в эту структуру, поскольку они не являются аргументами никакого слова предложения. С другой стороны, эти предложения содержат конструкцию, в значение которой входит предикат именованного, и дативная ИГ обозначает его субъект. Если принять во внимание то, что предикат именованного может быть выражен нелексически, то приходится признать, что дативная ИГ является полноценным участником предикатно-аргументной структуры предложения и тем самым сближается со стандартными употреблениями дательного падежа, используемого для маркировки аргументов предикатов, как, например, *давать, отвечать, помогать, вредить, соответствовать, верить, радоваться, поддаваться, равный, эквивалентный, синонимичный* и др.

3.2 Содержательные свойства

Обращаясь к содержательным свойствам конструкций, рассмотрим сначала параметр «Цель коррекции», который основан на классификации типов метаязыкового отрицания в [Geurts 1998]. В рамках данной классификации анализируемые нами примеры относятся к возражению против формы (англ. *form denial*), однако мы выделяем две более частные разновидности в зависимости от наличия / отсутствия дополнительного смыслового компонента, который указывает на определенный способ представления референта-носителя имени и принадлежит фонду общих фоновых знаний в области социальных отношений, региональных и культурных особенностей [Nyström 2016].

Примером конструкции, в которой коррекция направлена исключительно на формальные особенности языкового выражения, является (20), где речь идет о правильном написании фамилии героя.

- (20) — *Только не «ф», а «в», — сказал он. — Не Лифишиц, а Ливишиц.*

В (21), напротив, речь идет о коррекции способа представления героя в разные временные периоды и в разных обстоятельствах: если в тюрьме к нему обращались, используя кличку *Афоня*, то на свободе он ожидает обращения по имени и отчеству.

- (21) — *Не Афоня, а Степан Григорьевич. Это я в «чалкиной деревне» был Афоня Карельский, а здесь я уважаемый человек.*

На нашем материале случаи отрицания, направленного на способ представления референта, оказались более распространены (62%), чем примеры исправления исключительно формальных неточностей (38%), см. Табл. 3. Таким образом, метаязыковые отрицательные конструкции с именами собственными используются говорящим в первую очередь для указания на несоответствие корректируемой формы имени текущей коммуникативной ситуации, а не на коррекцию исключительно формальных ошибок.

Цель коррекции	Количество примеров	Относительная частотность (%)
Способ представления референта	198	62,264
Формальные свойства	120	37,736
Всего	318	100

Таблица 3: Цель коррекции.

Следующий параметр «Изменение уровня формальности» характеризует те примеры, в которых говорящий использует конструкцию для коррекции формы имени в сторону более или наоборот, менее официального варианта. К случаям первого типа относятся примеры (1) и (13) выше, в которых говорящий исправляет обращение по имени / кличке на официальную форму с именем и отчеством. Примером второго типа может служить (18) выше, где, напротив, полное имя и отчество *Евгений Александрович* исправляется говорящим на сокращенный вариант *Женька*.

Нам встретилось 63 примера, релевантных по данному параметру, из которых в 76% случаев коррекция была направлена на увеличение формальности: с одной стороны, при употреблении говорящим конструкции для указания на эволюцию референта (взросление, изменения статуса и т.д.) переход от менее к более формальному наименованию отражает естественный ход событий: с другой стороны, недостаточно формальное обращение расценивается как более серьезный и требующий исправления промах, чем обратная ситуация.

Изменение уровня формальности	Количество примеров	Относительная частотность (%)
Увеличение уровня формальности	48	15,094
Уменьшение уровня формальности	15	4,717
Не применимо	255	80,189
Всего	318	100

Таблица 4. Изменение уровня формальности.

Параметр «Смена имени» описывает два типа употреблений: случаи, когда в конструкции используются два разных имени, и примеры, когда речь идет о разных формах одного и того же имени. К иллюстрациям примеров первого типа относятся (9) выше и (22). В (22) действие происходит в 1914 году, во время начала Первой мировой войны, и персонажи — Борис и его отец — решают сменить немецкую фамилию *Мюллер* на русскую фамилию *Мельниковы*, чтобы избежать отрицательного отношения к себе из-за возможной ассоциации с Германией.

(22) *Они с отцом решили поменять фамилии. Теперь они не Мюллеры, а Мельниковы.*

На нашем материале встретилось 52 примера смены имени (16,4%), описывающие как ситуации следования традиции, например, взятие фамилии мужа или творческого псевдонима, так и индивидуальные обстоятельства героев, вызванные внешними условиями.

Последний параметр «Исправляемый участник» указывает, направлено ли исправление говорящего на форму, используемую ими самим или другим участником общения. В литературе отмечается, что метаязыковое отрицание используется в реактивных репликах, а корректируемый элемент является цитатным употреблением слов другого участника общения, и мы проверяли, насколько это наблюдение справедливо для употреблений конструкции в русском языке.

Примеры коррекции другого участника представлены в (1), (13), (21), а также (23). В (1), (13), (21) говорящий указывает, что считает неподходящим выбор формы имени другим участником и предлагает более уместную, на его взгляд, альтернативу. В (23) идет указание на ошибку — неверное имя и отчество, произнесенное персонажем из-за недостатка концентрации.

- (23) — *Мы ведь с вами... — Стоп! — командовал Эрмлер. — Федор Николаевич, дорогой, не Виталий Васильевич, а Василий Витальевич. Начали! — Здравствуйте, Викентий Васильевич! — встрепенулся Петров.*

К примерам исправления говорящим самого себя относятся (6), (9), (11), (12), (14) выше и (24). За исключением примеров типа (12), когда говорящий сам замечает и исправляет свою ошибку, корректируемый элемент, как правило, является более знакомым и ожидаемым для говорящего и адресата, а замещающий элемент преподносится как новая информация.

- (24) — *Пойдемте, Юрий Петрович, — сказала официантка Валя. Нет, уж не Валя, а Валентина Степановна. Внуки у нее росли.*

По нашим данным, исправление говорящим самого себя используется в 76,4% случаев, что не совпадает с утверждениями о реактивном характере метаязыкового отрицания в предшествующих работах. Тем не менее, нам кажется, что это наблюдение не противоречит, а расширяет спектр употреблений метаязыкового отрицания. Говорящий может использовать его не только для исправления ошибок, но и для усиления контраста, например, между двумя этапами жизни референта, уровнями формальности отношений с говорящим, позициями на социальной лестнице.

Исправляемый участник	Количество примеров	Относительная частотность (%)
Исправление самого себя	243	76,416
Исправление другого	75	23,585
Всего	318	100

Таблица 5. Исправляемый участник.

В целом, отвечая на исследовательские вопросы, поставленные во вводном разделе, мы можем сформулировать следующие наблюдения. На материале НКРЯ было выявлено, что метаязыковые отрицательные конструкции с именами собственными используются говорящим не столько для исправления ошибки в передаче имени, а скорее с целью коррекции информации, кодируемой в способе представления носителя имени. К нетривиальным свойствам конструкции с именами собственными, отличающим ее от метаязыковых конструкций с другими синтаксическими группами, описанными в литературе, относится, во-первых, тот факт, что метаязыковые поправки, направленные на самого говорящего, в НКРЯ более частотны, чем поправки, направленные на другое лицо³. Во-вторых, для данных конструкций случаи т.н. формального метаязыкового отрицания [Geurts 1998], оказываются более разнообразными, чем это отмечено для конструкций с именами нарицательными в предыдущих работах. Наконец, было показано, что «свободный дательный» может функционировать как актант лексически невыраженного предиката именованья, что демонстрирует более широкий, чем указывалось в литературе, спектр его реализации.

4 Заключение

Мы рассмотрели формальные и содержательные свойства метаязыковых отрицательных конструкций с именами собственными и получили следующие результаты. Во-первых, по нашим данным, основной функцией конструкции является коррекция способа представления носителя имени. Это отражается и в структурных характеристиках примеров: наряду с симметричным заполнением слотов именами собственными, говорящий использует в рамках замещающего элемента имена нарицательные, чтобы передать способ (верного) представления референта с большей точностью. Также распространенность временных маркеров в качестве дополнительных элементов конструкции указывает на ее употребление для указания на

³ Авторы благодарны за эту формулировку анонимному рецензенту.

изменения референта с течением времени. Далее, в наших примерах говорящий использует конструкцию для коррекции в сторону более официальной формы имени чаще, чем наоборот, и это, по-видимому, свидетельствует о том, что недостаточный уровень формальности осуждается больше, чем ее переизбыток. Кроме того, было выявлено, что случаи исправления говорящим самого себя преобладают над случаями коррекции другого, что позволяет говорить о более широком спектре употребления конструкции, чем утверждалось в предшествующих исследованиях. Наконец, было показано, что метаязыковой предикат называния может реализовываться не только лексически, но и входить в состав значения синтаксических конструкций.

Благодарности

Авторы благодарят за помощь в работе с материалом, подборе литературы и обсуждении исследования Б. Гертса, Л. Л. Июдина, С. А. Крылова, Ксению Лебедеву, а также трех анонимных рецензентов.

References

- [1] Azargaeva A. V. Metalinguistic negation in sentences with proper names: Evidence from Russian. A corpus-based study [Metazykovoe otritsanie v vyskazyvaniakh s imenami sobstvennymi: korpusnoe issledovanie na materiale russkogo yazyka]. Bachelor thesis / Azargaeva A. V.; St Petersburg State University, — St Petersburg, 2018. — Access mode: <http://hdl.handle.net/11701/13985>
- [2] Boguslavsky I. M. Studies in syntactic semantics: Scope of logical words [Issledovaniya po sintaksicheskoi semantike: sfery deistviya logicheskikh slov]. — Moscow: Nauka, 1985.
- [3] Boguslavsky I. M. Review of the book: E. V. Paducheva. Russian negative sentence [Russkoe otritsatel'noe predlozhenie]. Moscow: Yazyki slavyanskoi kul'tury, 2013 // Voprosy Jazykoznanija. — 2016. — Vol. 2. — P. 121–127.
- [4] Bosse S., Bruening B., Yamada M. Affected experiencers // *Natural Language and Linguistic Theory*. — 2012. — Vol. 30. — P. 1185–1230.
- [5] Burton-Roberts N. On Horn's dilemma: Presupposition and negation // *Journal of Linguistics*. — 1989. — Vol. 25. — P. 95–125.
- [6] Carston R. Metalinguistic negation and echoic use // *Journal of Pragmatics*. — 1996. — Vol. 25. — P. 309–330.
- [7] Datives and Other Cases: Between Argument Structure and Event Structure // Ed. by D. Hole, A. Meinunger, W. Abraham. — Amsterdam: John Benjamins, 2006.
- [8] Escandell-Vidal V., Vilinbakhova E. Negated Tautologies and Copular Contradictions: Interpretive Strategies // *International Review of Pragmatics*. — 2019. — Vol. 11, no. 2. — P. 153–199.
- [9] Geurts B. The Mechanisms of Denial // *Language*. — 1998. — Vol. 74 (2). — P. 274–307.
- [10] Gutzmann D., 2007. Eine Implikatur konventioneller Art: der Dativus Ethicus // *Linguistische Berichte*. — 2007. — Vol. 211. — P. 277–308.
- [11] Hole D. Extra argumentality – affectees, landmarks, and voice // *Linguistics*. — 2006. — Vol. 44, — P. 383–424.
- [12] Horn L. Metalinguistic negation and pragmatic ambiguity // *Language*. — 1985. — Vol. 61. — P. 121–174.
- [13] Horn L. A natural history of negation. — Chicago, IL: University of Chicago Press, 1989.
- [14] Horn L. I love me some datives: Expressive meaning, free datives, and F-implicature // Gutzmann D., Gärtner H.-M. (eds.) *Beyond Expressives: Explorations in Use-Conditional Meaning*. — Leiden: Brill, 2013. — P. 151–199.
- [15] Kustova G. I. Dative case [Datel'nyy padezh] // *Materials for the project of corpus description of Russian grammar*. — Manuscript. — Moscow, 2012. — Access mode: <http://rusgram.ru>
- [16] Lambert S. Beyond recipients. Towards a Typology of Dative Uses. Ph. D. thesis / S. Lambert; University of Buffalo. — Buffalo, 2010. — Access mode: <https://arts-sciences.buffalo.edu/content/dam/arts-sciences/linguistics/AlumniDissertations/Lambert%20dissertation.pdf>
- [17] Lamiroy B., Delbecque N. The possessive dative in Romance and Germanic languages // van Belle W., van Langendonck W. (eds.) *Casus and Grammatical Relations across Languages. The Dative*. — vol. II. — Amsterdam: John Benjamins, 1998. — P. 29–74.
- [18] Larrivé P. Metalinguistic negation from an informational perspective // *Glossa: a journal of general linguistics*. — 2018. — Vol 3(1): 56. — P. 1–22.
- [19] McCawley J. Contrastive Negation and Metalinguistic Negation // *The proceedings of the 27th Annual Meeting of the Chicago Linguistics Society (CLS 27)*. — Chicago, IL, 1991. — P. 189–206.

- [20] Moeschler J. 2018. A set of semantic and pragmatic criteria for descriptive vs. metalinguistic negation // *Glossa: a journal of general linguistics*. — 2018. — Vol 3(1): 58. — P. 1–30.
- [21] Nyström S. Names and Meaning // Hough C. (ed.) *The Oxford Handbook of Names and Naming*. — Oxford: Oxford University Press, 2016. — P. 39–51.
- [22] Paducheva E. *Russian Negative Sentence*. [Russkoe otritsatel'noe predlozhenie]. — Moscow: Yazyki slavyanskoi kul'tury, 2013.
- [23] Pennington J. The Bosnian-Croatian-Serbian adnominal possessive dative at the syntax-pragmatics interface // *Slavia Centralis*. — 2012. — Vol. 5 (1). — P. 104–121.
- [24] Russian National Corpus [Nacional'nyi korpus russkogo jazyka]. — 2003–2022. — Access mode: <http://www.ruscorpora.ru>
- [25] van der Sandt, R. 1991. Denial // *Chicago Linguistic Society*. —1991. — Vol. 27. — P. 331–344.
- [26] Umbach, C., 2004. On the Notion of Contrast in Information Structure and Discourse Structure // *Journal of Semantics*. — 2004.— Vol. 21 (2). — P. 155–175.

Refining Criteria of Paronymy for Building Computer Dictionaries of Russian Paronyms

Bolshakova E. I.

Lomonosov Moscow State University
HSE, Moscow, Russia
eibolshakova@gmail.com

Telegina A. D.

Lomonosov Moscow State University
Moscow, Russia
anya519@mail.ru

Abstract

Paronyms are words that have some similarity in sounding and spelling, but differ in meaning and usage (e.g., *sensitive – sensible, излишек – излишество*). In morphologically rich languages like Russian, paronymy is rather frequent phenomenon and one of the sources of speech difficulties. However, known dictionaries of Russian paronyms are not complete enough to help language learning or to support automatic correction of paronymy errors, and they do not provide precise definition of paronymy, which is necessary for constructing more extensive computer dictionaries. Aiming to clarify the concept of paronymy and to refine the previously proposed formal affix criterion of paronymy, we have performed a statistical study of paronyms taken from two printed dictionaries of Russian paronyms. Formal and semantic similarity of paronymy pairs were numerically estimated across various dimensions: proximity in affixes, in sounding, and in word meanings (the latter with the aid of neural models of distributive semantics and with an extensive base of Russian word combinations). Based on results of the study, refined criteria of paronymy and thresholds were proposed, which can be useful to automatically construct computer dictionaries of Russian paronyms, as well to replenish them by diagnostic contexts.

Keywords: paronymy conception; paronyms; criteria of paronymy; computer dictionary of paronyms

DOI: 10.28995/2075-7182-2022-21-61-69

Уточнение критериев паронимии для построения словарей паронимов русского языка

Большакова Е. И.

МГУ имени М. В. Ломоносова
НИУ ВШЭ, Москва, Россия
eibolshakova@gmail.com

Телегина А. Д.

МГУ имени М. В. Ломоносова
Москва, Россия
anya519@mail.ru

Аннотация

Паронимы – это слова, имеющие некоторое сходство в звучании и написании, но различающиеся по значению и употреблению (*sensitive – sensible, излишек – излишество*). В морфологически богатых языках, как русский, паронимия является довольно частым явлением и служит одним из источников речевых трудностей. Известные печатные словари русских паронимов недостаточно полны для помощи в изучении языка или для автоматизированного исправления паронимических ошибок, и они не дают точного определения паронимии, необходимого для построения более обширных компьютерных словарей. С целью уточнения понятия паронимии и ранее предложенного аффиксального критерия паронимии, нами было проведено статистическое исследование паронимов, взятых из двух печатных словарей русских паронимов. Формальное и семантическое сходство паронимических пар оценивалось численно по различным аспектам: близость в аффиксах, в звучании и в значениях слов (последнее с помощью нейронных моделей дистрибутивной семантики и обширной базы русских словосочетаний). По результатам исследования были предложены уточненные критерии паронимии и пороговые значения, полезные для автоматического построения компьютерных словарей русских паронимов, а также способ их пополнения диагностическими контекстами.

Ключевые слова: понятие паронимии; паронимы; критерий паронимии; компьютерный словарь паронимов

1 Introduction

Paronymy is linguistic phenomenon existing in many natural languages, as a relation between two or more words, which are similar in form (sounding and spelling), but differ in meaning and usage [7, 17], e.g.: Eng. *hare* – *hair*, Rus. *исправить* – *nonправильно*, Germ. *original* – *originell*. Such words-paronyms may be easily confused thus causing difficulties in speech understanding and creating [15, 16]. Mistakes, when one word is replaced by another word similar to it, but with different meaning, are called malapropisms in Western linguistics (e.g., *sensual news* instead of *sensational news*), while in Russian they are known as paronymy errors.

As a rule, such mistakes are typical for foreigners, but they also may appear in speech of native speakers. Several scientific works propose ways to automatically reveal and correct them [2,5], and for this purpose, appropriate computer dictionaries of paronyms are required, they are also obviously useful for teaching foreign languages. However, there are a few printed dictionaries of paronyms, for example, [1, 15], they have very limited size and rely on paronymy conception, which is not formal enough and varies significantly. Most dictionaries imply that paronyms differ only in affixes, i.e., prefixes and suffixes (e.g., *одеть* – *надеть*, *massive* – *mass*), the others indicate only similarity in sounding and spelling, with substantial difference in their meaning (*hare* – *hair*).

For Russian with its plenty of various of affixes, paronymy is a more vivid phenomenon, and there are three informative dictionaries of Russian paronyms [1, 10, 20], but the largest [10] contains only about 2,5 thou. words. Almost all paronyms in these dictionaries have the same root and part of speech (POS), thus having similar but somewhat different meaning. Differences in meanings of paronyms are explained and illustrated by diagnostic contexts. Meanwhile, the dictionaries do not give a precise definition of paronymy, which is necessary to build a more complete dictionary of paronyms.

The first computer dictionary of Russian paronyms constructed in [6] contains only word pairs differing in one or two letters (such as *комплекс* – *комплект*), whereas numerous word pairs that differ in several suffixes were not included. The work [4] describes a method to build computer dictionary of paronyms, which is intended to correct paronymy errors in Russian texts and based on the proposed formal affix criterion of paronymy. Only those pairs of words of the same root and POS are recognized as paronyms, for which the differences in affixes (suffixes and prefixes) are within the particular fixed limits. The volume of the built dictionary (about 135 thou. pairs of paronyms) is larger than all known dictionaries, but it also contains many word pairs that hardly be attributed to paronyms, in particular, dissimilar couples (such as *ходули* – *перерасходы*).

Thus, to build computer dictionaries more accurately, criteria of paronymy are to be further investigated. The difficulty of formalizing paronymy is related to several various aspects of similarity, as well as subjectivity of its perception. In order to refine criteria of similarity, we have performed a computational study of a representative set of paronyms (hereafter, etalon set) that were taken from two dictionaries of Russian paronyms [1, 10]. They were manually compiled by linguists and thus guarantee paronymy in its intuitive sense. In our work, only word pairs with the same root and POS were considered, and the following statistics for these pairs were evaluated:

- closeness in affixes, depending on POS (nouns, verbs, and adjectives, including participles);
- proximity in sounding by applying Soundex algorithm [9];
- difference in semantics, with the aid of neural distributive models [14] and CoSyCo corpus of syntactically related words [8].

Based on the revealed features of the paronymy pairs, we have formulated a refined affix criterion of paronymy that depends on POS, and also have proposed cut-off thresholds for proximity in sounding and in semantic similarity, which can be useful for building computer dictionaries.

We also have performed experiments to automatically build dictionaries of Russian paronyms and to estimate coverage of the etalon set. For the study and the experiments, we have exploited datasets¹ with Russian words split into morphs, open source programming tools², as well as our own tools³ developed for the task.

Thus, the contributions of this paper are the following:

¹ <https://github.com/cmc-msu-ai/NLPDatasets>

² <https://github.com/cmc-msu-ai/ParonStatistics>

³ <https://github.com/annatelegina/ParonymsAnalyzer>

- we present the results of the statistical study of paronyms from the etalon set, their closeness in affixes, sounding, semantics, and combinability;
- we propose refined criteria of paronymy that account for various aspects of word similarity and depend on preset thresholds;
- we report on the experiments for building dictionaries of Russian paronyms undertaken with the developed open-source tools and various combinations of the proposed criteria.

In the next section, we shortly describe and compare paronymy dictionaries [1, 10, 20], consider the affix criterion of paronymy used in [4], and based on this, clarify our understanding of paronymy. The results of the statistical study of formal and semantic proximity of paronyms from the etalon set are reported in sections 3 and 4, respectively. Section 5 describes experiments on building several computer dictionaries of Russian paronyms. Conclusions are presented in Section 6.

2 Paronymy Conception and Paronymy Dictionaries

Printed dictionaries of Russian paronyms [1, 10, 20] contain only nouns, verbs, adjectives (including certain participles), and most paronyms have the same root and the same part of speech (POS). Meanwhile, conception of paronymy somewhat vary: for example, in [20] the same gender for nouns and aspect for verbs, the same number of syllables and similar place of accent are indicated as additional features for true paronyms, thereby narrowing the conception. Paronyms are gathered either by pairs (~1000 pairs in [20]) or by so-called paronymy groups of 2–7 semantically close words (about 200 groups in [1] and 1100 groups in [10], such as *земельный – землистый – земляной – земной*). The relation between paronymy conception and semantics of the compared word roots remains unclear: in [1] similar words with different roots (e.g., *индейка – индианка*) called quasi-paronyms, in [10] words with homonymous roots (e.g., *платный – платьной*) are not considered as true paronyms, but this dictionary includes synonyms (*патетический – патетичный*), as well as words differing only in combinatorics (*туристский – туристический*). Nevertheless, the dictionaries present many diagnostic contexts for distinguishing meanings of particular paronyms (e.g., *игорный бизнес – игральный стол – игривый щенок – игристое вино – игровая зона*).

All in all, these limited-size dictionaries do not give a precise definition of paronymy needed to build a more complete computer dictionary of paronyms. For this purpose, the work [4] proposed a formal affix criterion of paronymy that takes into account statistics of affix proximity of paronyms from the largest dictionary [10]. The affix similarity of two words of the same root and POS is estimated separately for prefixes and suffixes, by a pair of integers (N_p, N_s), N_p is the number of different prefixes, i.e. the minimum number of elementary editing operations [11] transforming chain of prefixes of one word into prefixes of another word. The number N_s for suffixes is computed similarly. The affix criterion is written as $(N_p = 0) \ \& \ (N_s \leq 3) \vee (N_p = 1) \ \& \ (N_s \leq 2)$ – either the prefixes in the compared word pair are the same, and there are no more than three differences in suffixes, or words have one different prefix, and there are no more than two differences in suffixes, for example: *о-дар-ённ-ый – дар-овит-ый* ($N_p = 1, N_s = 1$), *такт-ик-а – такт-ичн-ость* ($N_p = 0, N_s = 2$). Though this affix criterion covers almost 99% of paronyms from the dictionary [10], and its application to words from [3] produced a volume computer dictionary of paronyms (~135 thou. pairs), the resulted dictionary has some drawbacks. It turned out that the criterion allows antonyms (*типичный – атипичный*) and near synonyms (*патетический – патетичный*), some outwardly dissimilar word pairs (*седловина – сиденье*), and pairs distinguished only by a diminutive suffix (*мел – мелок*). At the same time, some similar pairs (e.g., *мерзость – омерзительность*) were not recognized as paronyms.

Aiming to refine the formal criterion of paronymy, we have studied various aspects of similarity for the set of paronyms taken from [10] and enlarged by paronyms from the dictionary [1], additional features of the paronymy pairs were studied as well. We considered pairs of words with the same root and POS, not paronymy groups, because some words within a particular group may vary significantly in affixes and semantics. Since surprisingly many noun pairs within paronymy groups have different gender, unlike [1, 4], we do not exclude such pairs from consideration.

The preliminary analysis of paronyms from all the dictionaries [1, 10, 20] has showed that allomorphy of the roots (owing to alternating consonants and fluent vowels, e.g., *отчий – отецский*) is allowed, but there are no pairs distinguished by:

- antonymic prefixes (*не-, а-, анти-, контра-, против-,* etc.);
- foreign prefixes (e.g., *гуро, инфра*) and prefixoids (*меж-, само-,* etc.);
- diminutive and magnifying suffixes (*-ушк, -онок, -ёнок, -ица,* etc.);
- postfixes (*-ся, -сь*) or verb aspect (perfect and imperfect).

Therefore, such word pairs should not be included in paronymy dictionaries. In our opinion, synonyms also do not belong to true paronyms, although there are several pairs in [10] (e.g. *крохотный – крошечный*). At the same time, similar to the work [4] it is reasonable to admit quite similar pairs with homonymous roots (such as adjectives *бурный – буровой*). But unlike [4], in our study:

- we consider a more representative set of paronyms encompassing two dictionaries [1, 10];
- the words are split into morphs (prefixes, root, suffixes, ending) according to the derivational dictionary [19] and the work [13] (as splitting of suffixes used in [3, 4] is not conventional);
- statistics for word pairs were collected and evaluated separately for different POS;
- besides affix proximity of paronyms, their semantic similarity was estimated in order to exclude synonymous pairs and also pairs differing by ambiguous diminutive suffixes used not in diminutive sense (e.g., *цвет – цветок*).

In total, our etalon set of paronyms encompasses 2704 words, most of them belong to adjectives: 54% of adjectives (including participles), 33% of nouns and 13% of verbs. All paronymy pairs (2013) were estimated for their formal similarity and semantic proximity.

3 Formal Similarity

3.1 Affix Distances

To estimate affix proximity for pairs from the etalon set of paronyms, morpheme segmentation of words were taken from RuMorphs-Lemmas⁴ dataset (created from [19]) and were additionally modified according to [13], as this is more relevant for comparing word pairs separately within the considered POS: nouns, adjectives, verbs. In particular, the so-called thematic vowels of verbs *a, i, e* were attached to the verb suffixes and suffixes of participles, some adjacent suffixes were also concatenated (e.g., *норм-ирова-ть* instead of *норм-ир-ов-а-ть*). Statistics of affix distances for paronymy pairs were automatically computed, the results are given in Table 1. The first two columns show all discovered combinations of distances in prefixes (N_p) and suffixes (N_s), respectively. The other columns present the number of paronymy pairs for the particular POS, according to the corresponding combinations. The affix distance equals the minimal number of editing operations [12] on affixes (their deletion, insertion or substitution) that transform one word of the compared pair into the other (word endings are not taken into account).

N_p	N_s	Nouns	Adjectives	Verbs
0	0	12	11	4
0	1	273	544	9
1	0	89	30	331
1	1	8	26	0
0	2	206	357	4
2	0	0	3	1
1	2	3	11	0
0	3	34	57	0
2013		625	1039	349

Table 1: Statistics of affix distances for paronymy pairs

One can notice that the statistics differ for various POS: most of verb pairs (around 94%) are prefix paronyms with distance $N_p = 1$, only 4 pairs have suffix distance $N_s = 2$ (e.g. *хозяйничать – хозяй-*

⁴ <https://github.com/cmc-msu-ai/NLPDatasets>

ствовать, etc.) and only one pair has prefix distance $N_s = 2$ (изобразить – отразить). In contrast, for adjectives and nouns, most pairs are suffix paronyms with distances $N_s = 1$ and 2, and prefix paronyms with $N_p \geq 1$ are relatively rare or absent.

This statistics corresponds to the well-known fact that Russian derivation for verbs is predominantly prefixal, while for nouns and adjectives, it is suffixal. Thus, the criteria of affix similarity for paronyms should obviously depend on the particular part of speech.

Some features are common for all POS: most paronymy pairs differ only in one affix, and many pairs differ in two affixes. Some couples turned to be at the minimum distance (0, 0), when the words do not differ in affixes and have the same root morph, in particular, verbs with allomorphic roots (огородить – оградить), adjectives differing only by ending (временный – временной).

It seems acceptable do not account for rare pairs with distances (2, 0), so the refined affix criterion of paronymy can be set as follows, separately for various POS:

for nouns and adjectives: $(N_p = 0) \ \& \ (N_s \leq 3) \ \vee \ (N_p = 1) \ \& \ (N_s \leq 2)$
 for verbs: $(N_p = 0) \ \& \ (N_s \leq 2) \ \vee \ (N_p = 1) \ \& \ (N_s = 0)$.

This affix criterion cover 99.4% pairs from the etalon set. When applying the refined affix criterion for building a computer dictionary of paronyms, it is necessary to additionally exclude pairs that differ only i) by any antonymic prefix; ii) by any foreign prefix or prefixoid; iii) by postfix; iv) by a nonambiguous diminutive suffix (e.g., аргумент – контраргумент, сетевой – межсетевой, умывший – умывшийся, кот – котик, etc.). For this purpose, a pre-compiled list of such prefixes, prefixoids, and suffixes is used, but it does not contains ambiguous diminutive suffixes (-чик, -ик, etc.), because it is necessary to separate pairs with suffixes in diminutive sense (дом – домик,) from those without diminutiveness (перевод – переводчик, такт – тактика), and additional techniques are required for their recognition.

3.2 Proximity in Sounding

Another aspect of word similarity is their sound proximity and to estimate it, we use Soundex algorithm [9] and its open implementation⁵ for Russian. For a given word pair, this algorithm produces an integer that indicates the degree d of sound proximity: the smaller it is, the more similar the words sound. The sound proximity was evaluated for paronymic pairs from the etalon set, separately for various POS, the results are given in Figure 1. Though most paronymy pairs has d equal to 2 or 3, the situation for verbs is again different from that for nouns and adjectives: the most verb pairs have $d < 5$. The “tail” of statistical distribution corresponds to rare dissonant pairs (such as натуралистический – натурный with $d = 10$). If sound proximity is important, such pairs can be eliminated from a computer dictionary, with the aid of a preset threshold P_d : pairs with $d > P_d$ are to be excluded. In particular, $P_d = 5$ for verbs and $P_d = 7$ for nouns and adjectives exclude 19 pairs of adjectives, 10 pairs of nouns, and only 1 pair of verbs (сигнализировать – сигналить) from our etalon set.

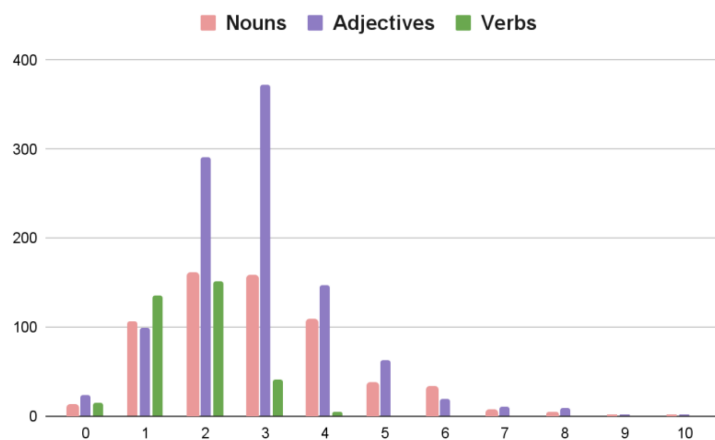


Figure 1: Statistics of sound proximity

⁵ <https://pypi.org/project/ru-soundex/>

4 Semantic Proximity

4.1 Word Embeddings

Neural models of distributive semantics are widely used in modern computational linguistics, so we used word-level models relevant for our task, Word2Vec [14], to estimate semantic proximity of paronyms. Words are represented as vectors (embedding) in the space formed by such models, and similarity (or difference) in words meaning are evaluated by cosine distance between their vectors: if it is large enough, then the words have similar meanings. Thus, the models are suitable to identify synonyms (in order to exclude them from a dictionary), and also to reveal diminutive sense of ambiguous suffixes (*-чик-*, *-ик-*, *-к-*, *-ок-*, *-ец-*, *-иц-*, *-ц-*, etc.) in particular words.

Among pre-trained Word2Vec models from RusVectors project⁶ [11], for our experiments we have chosen RusCorpora and Tayga – cf. Table 2, since they contain more words⁷ from the etalon set (the number of absent word are given in the third column of Table 2).

Text Collection For Training	#Words	Training Method	#Absent Paronyms
RusCorpora	278 million	CBOW	110 (171 pairs)
RusCorpora + Wikipedia	778 million	SkipGram	133 (202 pairs)
Tayga	5 billion	CBOW	106 (162 pairs)
Russian News	2.6 billion	SkipGram	400 (434 pairs)

Table 2: Word2Vec models from RusVectors

The statistical distribution of computed cosine distances ($dcos$) for pairs of paronyms is presented in Figure 2. Values $dcos$ vary greatly, the right "tails" correspond to pairs that are close in meaning and may be synonyms. In experiments we have found the threshold $P_s = 0.8$ (for Tayga model), such that pairs close in meaning ($dcos > P_s$) can be considered synonyms and so be excluded from the dictionary of paronyms: *крохотный – крошечный* ($dcos=0.98$), *завесить – занавесить* ($dcos=0.81$), *целебный – целительный* ($dcos = 0.81$). However, words that seem synonymous may differ by combinations with other words used in speech, and for word pairs with $dcos$ close to the cut-off threshold, it seems reasonable to additionally compare their combinability with other words.

To identify diminutive meaning of an ambiguous diminutive suffix, we have performed additional experiments with 450 arbitrary taken pairs differing by such suffixes, and have empirically chosen a threshold for $dcos$: $P_a = 0.55$. Word pairs with $dcos \geq P_a$ are excluded (e.g., *блокнот – блокнотик*, $dcos=0.75$), while pairs with $dcos < P_a$ are recognized as paronyms (*такт – тактика*, $dcos = 0.14$).

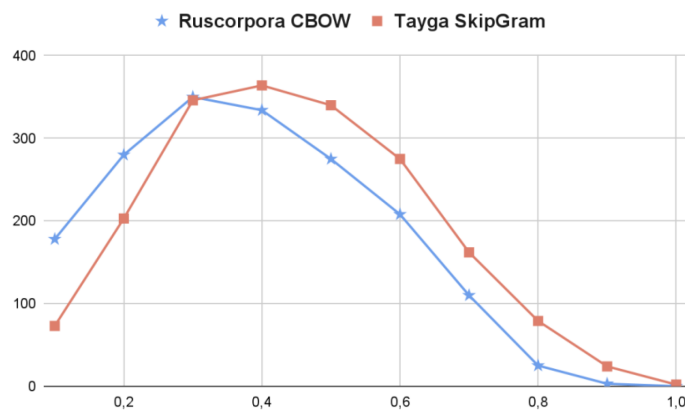


Figure 2: Distribution of semantic proximity $dcos$

⁶ <https://rusvectors.org/ru/>

⁷ Almost all such models do not include vectors for words that are either rare or not encountered in the text collection for training. FastText still constructs vectors for such word, but we cannot rely on their properties.

4.2 Word Combinability

Since paronyms are often distinguished by the words combined with them, another way to compare their semantics implies revealing differences in context words syntactically related with them. Corresponding diagnostics contexts are usually presented in paronymy dictionaries, such as *болотный цвет* but *болотистая местность* for paronyms *болотный* and *болотистая*.

Based on the assumption that the closer the words are in meaning, the greater the number of identical words combined with them, and vice versa, we evaluate their semantic similarity (or difference) by comparing their combinations with words of other POS, taking them from CoSyCo [8], a large corpus of syntactically related Russian words (~1.75 million combinations). The corpus was automatically constructed from large text collections and encompasses word combinations (bigrams) of several types, along with frequencies of occurrences for each particular combination and its words-components.

Since CoSyCo contains a plenty of word combinations (up to 2–3 thou.) for each particular word, and most of them are neither stable nor idiomatic, we need to reveal the most stable and typical among them. After extraction from CoSyCo all combinations: *Adjective + Noun* for each nouns or adjective and *Verb + Noun* for each verb, they are ranged by applying logDice association measure [18]:

$$\logDice = \log_2 \frac{2 * Freq(a,b)}{Freq(a) + Freq(b)},$$

where $Freq(a, b)$, $Freq(a)$, $Freq(b)$ are frequencies of the word combination and its components, respectively. This measure is well-known and well-performing for revealing stable combinations, and it is suitable for our task, because it does not require the size of source texts (on which statistics were collected), such information is absent in CoSyCo.

Having two ranged lists of word combinations for two given words W_1 and W_2 , we take N -tops of the lists (the most stable combinations), and compare sets S_1 and S_2 of words combined respectively with W_1 and with W_2 , by computing proportion of common words in them (more precise, cardinality of intersection of these sets, divided by N):

$$sim = \frac{|S_1 \cap S_2|}{N}$$

In such a way, semantic similarity sim of the words can be estimated, but what is more important, words that are not included in the intersection (i.e., are not common) distinguish the meaning of the compared words W_1 and W_2 and thus can be considered as diagnostic contexts. For example, if $N=10$, for paronyms *дружеский* and *дружественный* semantic similarity equals to 0.3 (they are somewhat similar), and their diagnostic contexts include: *дружеский* – *вечеринка, участие, пирушка, шарж* but *дружественный* – *интерфейс, страна, держава, государство*.

To find the number N that is appropriate for extracting diagnostic contexts for paronymy dictionary, we have evaluated sim values for paronyms from the etalon set, for various N (up to 80) and separately for various POS, the resulted mean values are shown in Figure 3.

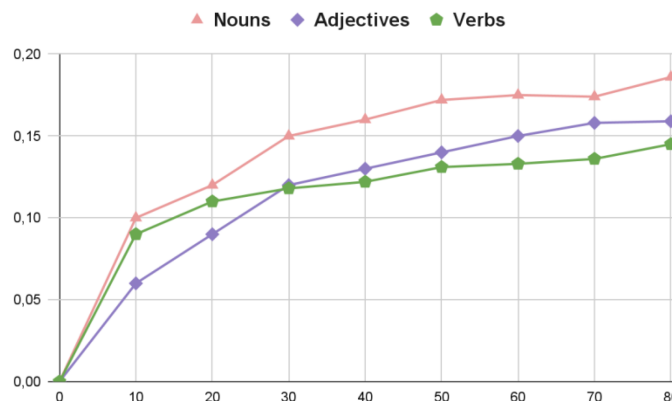


Figure 3: Distribution of similarity in word combinability

The growth of similarity values *sim* with the growth of *N* are explained by increase of common words for the considered paronyms, but at the same time the stability of the corresponding word combinations decreases – this means that distinguishing words (diagnostic contexts) are appeared more rarely and within less stable combinations. Our analysis of extracted word combinations for 90 paronymy pairs has showed that *N* = 20 is appropriate for revealing typical diagnostic contexts to be included them into a dictionary of paronyms being constructed.

5 Constructing Dictionaries with Refined Criteria

Due to ambiguity in understanding of paronymy and various purposes of paronymy dictionaries, it is reasonable to consider the above-proposed cut-off thresholds P_s, P_a, P_d as parameters of procedure for building a dictionary – this makes it possible to set more or less strict restrictions on paronymic pairs, depending on the task. In order to build different dictionaries, the developed programming tools⁸ provide API for setting needed thresholds and other parameters.

In our experiments to build dictionaries of Russian paronyms, as input data we have used Ru-Morphs-CrossLexica⁹ dataset (26 thou. words taken from [3]). To estimate built dictionaries (and thus the refined criteria), we have evaluated their volumes, as well as coverage of the etalon paronymy set, the coverage degree is calculated as follows:

$$C_{measure} = \frac{|D \cap E|}{|E|}$$

where *D* is the set of paronymic pairs in the built dictionary and *E* are pairs in the etalon set. Table 3 presents results obtained for different combinations of criteria and the proposed thresholds: $P_s = 0.8$, $P_a = 0.55$, $P_d = 5$ for verbs and $P_d = 7$ for nouns and adjectives. The first column corresponds to particular combinations of the criteria, while the second and the third show the coverage and volume (thou. paronymic pairs) for the built dictionary.

Criteria	$C_{measure}$	Volume
Refined Affix Criterion	99.7	100.7
Refined Affix Criterion + P_s	99.6	100.6
Refined Affix Criterion + P_a	99.5	101.0
Refined Affix Criterion + P_d	98.2	98.1
All Criteria	98.0	97.4

Table 3: Comparison of the built dictionaries

One can notice that coverage degree slightly decreases with the introduction of the thresholds, as does the volume of the resulted dictionary of paronyms. With all refined criteria, the volume of the dictionary is about 97.4 thou. paronymy pairs (instead of 135 thou. pairs and 99% coverage degree for the dictionary built in [4]). Decrease of the volume and the coverage does not mean that the refined criteria work worse, as only word pairs are excluded that are less typical for paronymy.

In addition, we manually reviewed some fragments of the resulted dictionary, to estimate the appearance of non-paronymy pairs in them. The most pairs correspond to our refined understanding of paronymy, however, there are some pairs of participles (such as *увиденный – увидевший*, passive and active forms), but whether such pairs should be excluded from paronyms is an open question that requires further research.

6 Conclusions

Based on the study of paronyms taken from the representative etalon set, their features of formal and semantic proximity, we have refined affix criterion of paronymy in Russian, as well as have proposed the ways to estimate proximity of potential words-paronyms in sounding and semantics. It turned out that similarity in affixes and in sounding differ for nouns and adjectives, on the one hand, and verbs,

⁸ <https://github.com/annatelegina/ParonymsAnalyzer>

⁹ <https://github.com/cmc-msu-ai/NLPDatasets>

on the other hand, thus resulting in the affix criterion depending on POS of words being compared. For estimating semantic proximity, two ways were considered: neural models of distributive semantics and the large database of word combinations, the latter enables to reveal diagnostic contexts (typical word combinations) that distinguish meaning of compared words. The proposed cut-off thresholds for proximity in sounding and semantic similarity can be changed according to particular tasks.

Clearly, it is hardly possible to fully formalize the concept of paronymy, since it is associated with diverse aspects of word similarity, as well as subjectivity of its perception. Nevertheless, dictionaries of Russian paronyms can be automatically built for various applied tasks, in particular, for teaching languages or automatic correction of paronymy errors. The dictionaries may differ in volume, degree of formal and semantic similarity of paronymy pairs, which can be achieved with the developed programming tools, by setting appropriate thresholds.

References

- [1] Belchikov Yu. A., Panjusheva M. S. (2004), Dictionary of Russian Paronyms [Slovar' paronimov russkogo yazyka], Russkij Jazyk, Moscow.
- [2] Bolshakov I.A., Gelbukh A. (2003) On Detection of Malapropisms by Multistage Collocation Testing. // A. Düsterhöft, B. Talheim (Eds.) Proc. Int. Conf. Applications of Natural Language to Information Systems NLDB'2003, June 2003, Burg, Germany, GI-Edition, LNI V. P-29, p. 28–41.
- [3] Bolshakov I.A. (2013), CrossLexica – Universum of links between Russian words [CrossLexica – universum svyazey mezshdu russkimi slovami], Business Informatics [Biznes Informatica], No 3 (25), pp. 12–19.
- [4] Bolshakova E. I., Bolshakov I. A. (2015), Affix criterion of paronymy for building a computer dictionary of Russian paronyms [Affiksaly'nyj kriterij paronimii dlya postroeniya komp'yuternogo slovarya paronimov russkogo yazyka], Scientific and technical information [Nauchno-tehnicheskaya informaciya], Ser. 2, № 11, pp. 28–35.
- [5] Costin-Gabriel C. (1998), Malapropisms detection and correction using a paronyms dictionary, a search engine and Wordnet, Search, Bucharest, pp. 364-373.
- [6] Gusev V. D., Salomatina N. V. (2001), Electronic dictionary of paronyms [Elektronnyj slovar' paronimov, ver. 2], Scientific and technical information [Nauchno-tehnicheskaya informaciya], Ser. 2, № 7, pp. 26–33.
- [7] Hartmann R. R. K., James G. (1998), Dictionary of Lexicography. London, Routledge.
- [8] Klyshinsky E., Lukashevich N., Kobozeva I. Creating a corpus of syntactic co-occurrences for Russian. In: Computational Linguistics and Intellectual Technologies: Proc. of the International Conference “Dialogue 2018”, Issue 17 (24). Moscow, 2018, pp. 311-324.
- [9] Knuth D. E. (1973). The Art of Computer Programming: Volume 3, Sorting and Searching. Addison-Wesley, pp. 391–392.
- [10] Krasnykh V.I. (2007), Explanatory Dictionary of Russian Paronyms [Tolkovyj slovar' paronimov russkogo yazyka], AST Astrel, Moscow.
- [11] Kutuzov A., Kuzmenko E. (2017), WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Analysis of Images, Social Networks and Texts. AIST 2016. CCIS, vol. 661, Springer, Cham.
- [12] Levenshtein V.I. (1966). “Binary codes capable of correcting deletions, insertions, and reversals”. Soviet Physics Doklady. 10 (8): 707–710.
- [13] Lopatin V. V., Ulukhanov I. S. (2016). The dictionary of word-forming affixes of the modern Russian language [Slovar' slovoobrazovatel'nyh affiksov sovremennogo russkogo yazyka], Azbukovnik, Moscow.
- [14] Mikolov T., Chen K., Corrado G. (2013), Dean J. Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR, Arizona.
- [15] Müller W. (1973), Easily confused words [Leicht verwechselbare Wörter], Duden Taschenwörterbücher, vol. 17, Mannheim, Bibliographisches Institut.
- [16] Péchoin D., Dauphin B. (2001) Dictionnaire des difficultés du français d'aujourd'hui. Larousse.
- [17] Rosenthal D. E., Telenkova M. A. (1976), Dictionary-handbook of linguistic terms [Slovar'-spravochnik lingvisticheskikh terminov], Prosveshchenie, Moscow.
- [18] Rychlý P. (2008) A Lexicographer-Friendly Association Score: RASLAN, Brno.
- [19] Tikhonov A. N. (1990), Word Formation Dictionary of Russian language [Slovoobrazovatel'nyj slovar' russkogo yazyka], Russkij jazyk, Moscow.
- [20] Vishnyakova O. V. (1984), Dictionary of Russian Paronyms [Slovar' paronimov russkogo yazyka] Russkij jazyk, Moscow.

Contrastive fine-tuning to improve generalization in deep NER

Ivan Bondarenko

Novosibirsk State University
Russia, Novosibirsk
i.bondarenko@g.nsu.ru

Abstract

A novel algorithm of two-stage fine-tuning of a BERT-based language model for more effective named entity recognition is proposed. The first stage is based on training BERT as a Siamese network using a special contrastive loss function, and the second stage consists of fine-tuning the NER as a "traditional" sequence tagger. Inclusion of the contrastive first stage makes it possible to construct a high-level feature space at the output of BERT with more compact representations of different named entity classes. Experiments have shown that this fine-tuning scheme improves the generalization ability of named entity recognition models fine-tuned from various pre-trained BERT models. The source code is available under an Apache 2.0 license and hosted on GitHub https://github.com/bond005/runne_contrastive_ner

Keywords: named entity recognition, contrastive learning, Siamese neural networks, BERT

DOI: 10.28995/2075-7182-2022-21-70-80

Сопоставительное дообучение для повышения обобщающей способности нейросетевого распознавателя именованных сущностей

Иван Бондаренко

Новосибирский государственный университет
Россия, Новосибирск
i.bondarenko@g.nsu.ru

Аннотация

Предложен новый алгоритм двухэтапного дообучения нейросетевой языковой модели BERT для более эффективного распознавания именованных сущностей. Первый этап представляет собой дообучение BERT как Сиамской нейронной сети с использованием специальной сопоставительной функции потерь, а второй этап связан с окончательным дообучением распознавателя именованных сущностей как "традиционного" классификатора элементов последовательности. Добавление первого этапа, основанного на методе сопоставительного обучения, обеспечивает построение высокоуровневого признакового пространства на выходе нейросетевой языковой модели BERT с более компактными представлениями разных классов именованных сущностей. Эксперименты показывают, что такая схема дообучения повышает обобщающую способность распознавателей именованных сущностей на базе целого ряда предобученных языковых моделей BERT. Исходный код доступен под лицензией Apache 2.0 и размещен на GitHub https://github.com/bond005/runne_contrastive_ner

Ключевые слова: распознавание именованных сущностей, сопоставительное обучение, Сиамские нейронные сети, BERT

1 Introduction to problem: why and what do we solve?

The NER (Named Entity Recognition) task has been known for a long time and is generally formulated as finding key elements, like names of people, places, brands, monetary values, and more, in a text. This is used in many software products, so there is a lot of research on this topic. Specifically, over the past few years, most NER solutions have been based on the Transformer architecture.

There are many different approaches to Transformer fine-tuning. First, there is a development direction dedicated to the modification of the loss function and a specific problem statement. For example, training

problem could be set as machine reading comprehension (question answering) instead of the standard sequence classification, or focal loss, dice loss and other things from other deep learning domains could be used instead of the standard cross-entropy loss function. Second, there are papers devoted to BERT extension, related to adding more input information from the knowledge graph, morpho-syntactic parsers and other things. Third, there is a group of algorithms associated with changing the learning procedure, such as metric learning (contrastive learning).

Each direction has its own advantages and disadvantages, but the metric learning seems the most promising to us. Because the goal of any training is not to overfit the training sample and not just to take the top of the leaderboard on a particular test sample from the general population, but to ensure the highest generalization ability on the general population as a whole. High generalization ability is associated with good separation in the feature space. A good separation is possible when objects of different classes form sufficiently compact regions in our space. And methods of contrastive learning achieve better separation.

Our goal is to test, on the basis of the RuNNE competition (Artemova et al., 2022), how true are these theoretical considerations in practice and how much will the use of comparative learning in BERT’s fine tuning allow us to build more compact high-level representations of different classes of named entities and, as a result, improve the quality of recognition of named entities.

2 Standing on the shoulders of giants: research of our predecessors

Let’s consider some important approaches illustrating the previously described directions. In one the approaches, described in the paper “A Unified MRC Framework for Named Entity Recognition”(Li et al., 2020a), the sequence tagging task transformed into question answering. The inputs of the model are the “question” in natural language (for example, “Find locations in the text”) and the text, for which the model must predict the indexes of the beginning and the end of the entity, which is the answer to the “question”. Questions are generated separately for each entity class and can be done manually or based on rules. The model trained in this way showed the best quality of results compared to other BERT-based models on various datasets, including those with nested entities.

A comparable result was achieved by the WCL-BBCD(Zhou et al., 2022) approach, where, instead of changing the task, a modified training procedure of the BERT-based model is used. Additionally, a priori information from knowledge graphs is also used. The idea of BERT training is representation learning - it is necessary to teach the model to separate representations of different classes in the feature space. To do this, the generation of “similar” sentences by translating into another language and then back-translating into the original language is used. The better the model determines whether a pair of sentences are similar or not, the more the classes of entities contained in these sentences are separated in the feature space, and the better the model will be fine-tuned for the NER task. This idea is similar to ours.

Contrastive learning also helps to get better results in the few-shot NER problem. For example, the authors of the CONTaiNER(Das et al., 2021) article successfully used contrastive learning to solve the few-shot problem and surpassed the results of previous models. In that model, unlike ours, Gaussian embeddings were used and there was no second stage of fine-tuning.

We looked at the approaches that, in one way or another, inspired us to create our model. There are also many other popular and interesting approaches to various formulations of the NER problem.

3 CoNER: proposed contrastive named entity recognizer

We propose **CoNER** - a **C**ontrastive **N**amed **E**ntity **R**ecognizer. It is based on a special two-stage fine-tuning of a pretrained BERT language model:

- The first stage is a fine-tuning of the pretrained BERT as a Siamese neural network to contrast semantics of different entities in text pairs
- The second stage is a fine-tuning of the resultant neural network as a standard NER (i.e. sequence classifier) with a BILOU tagging scheme

3.1 Contrastive fine-tuning

The first stage of the fine-tuning is based on working with BERT as with the Siamese neural network with a contrastive loss (see Figure 1).

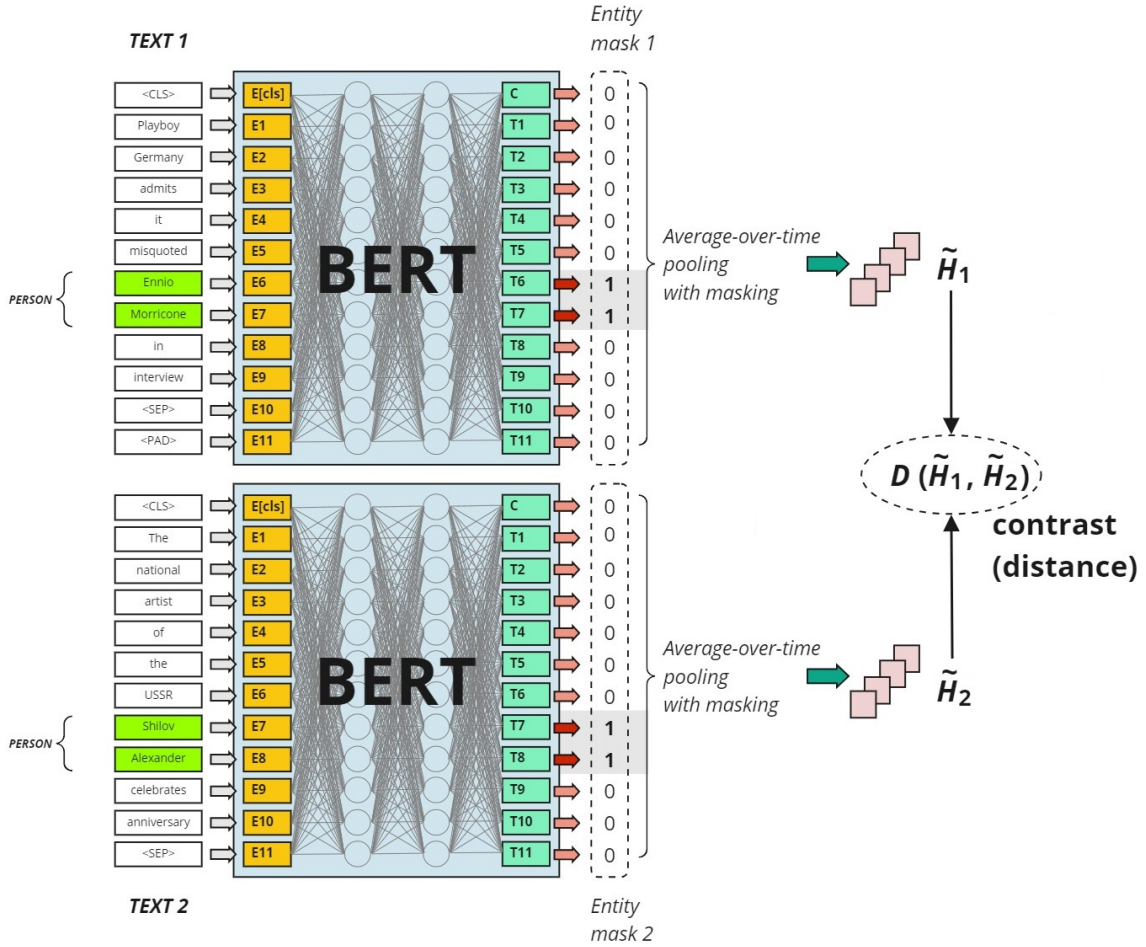


Figure 1: Entity masking example for the sequence outputs of two BERT sub-networks in the contrastive learning with Siamese BERT.

Unlike the well-known Sentence BERT (Reimers and Gurevych, 2019), we work with pairs of entities such as text segments instead of pairs of whole texts. Each text segment is specified as a pair $Ent = \langle S, M \rangle$, where S corresponds to a tokenized sentence containing the entity, and M specifies a token-in-text mask that defines the bounds of the entity. Thus, the named entity embedding generated from the last hidden layer of BERT is described using the following expression:

$$H(S, M) = F_{\text{BERT}}(S) \circ M, \quad (1)$$

where (\circ) is a masking operator which implements an element-wise multiplication between matrices of the same sizes (a masked matrix of embeddings and a masking matrix of zeros and ones).

After that we apply L_2 normalization to the named entity embedding to regard it as point on a unit hypersphere:

$$\tilde{H}(S, M) = \frac{H(S, M)}{\|H(S, M)\|} \quad (2)$$

Finally, we formulate the probability of the classes of the entities in the pair matching based on the euclidean distance between their embeddings as points on the unit hypersphere:

$$p(\tilde{H}_1, \tilde{H}_2) = \frac{1 + \exp(-m)}{1 + \exp(D(\tilde{H}_1, \tilde{H}_2) - m)}, \quad (3)$$

where m is the margin parameter to inflict of a penalty on the matched pairs with a too large distance (usually it equals to 1). Then we can use the log-loss as in the classification case. This distance based logistic (DBL) loss for Siamese neural networks was firstly proposed for a special computer vision task, concerned with localizing street views on satellite images. (Vo and Hays, 2016)

In comparison with a "classical" contrastive loss function, which is popular for Siamese neural networks, the DBL loss function is more effective owing to quicker convergence. In contrast to classification loss function after Siamese network structure such as Sentence-BERT, the DBL loss function does not need an additional trainable layer what allows us to stay focused on fitting the BERT sub-network only. Unlike the popular N -pairs loss and the SupCon loss function (Khosla et al., 2020), the DBL loss function provides a more stable training process on small mini-batches, and it is less inclined to the exploding gradients problem in such situations.

3.2 Final fine-tuning for NER

We use the BERT model fine-tuned according to the principles from subsection 3.1 as the base for a multi-head sequence classifier where each head corresponds to one of 29 named entity classes. Since a named entity can consist of several tokens, then we use the **BILOU** tagging scheme and describe the named entity class as a system of five token classes: the **Outside**, the **Beginning**, the **Inside** and the **Last** tokens of multi-token chunks (see example 5) as well as **Unit-length** chunks (see example 4).

For example, the text "*В Китае отметили 170-летие публикации «Коммунистического манифеста»*" (in English, "China celebrated the 170th anniversary of the publication of the Communist Manifesto") contains several named entities of different classes. Examples 4 and 5 illustrate applying BILOU tagging to the named entities in this text, which consist of single token and multiple tokens, accordingly:

- (4) *В Китае отметили 170-летие публикации «Коммунистического манифеста»*
 O U O O O O O O
 ‘BILOU labeling for the unit-length entity of the COUNTRY class’
- (5) *В Китае отметили 170-летие публикации «Коммунистического манифеста»*
 O O O O O B I I L
 ‘BILOU labeling for the multi-token entity of the WORK_OF_ART class’

We have two reasons for using BILOU instead of well-known BIO (the Beginning - the Inside - the Outside) tagging scheme:

1. some preceding experiments demonstrate that the BILOU formalism outperforms the BIO tagging scheme. (Ratinov and Roth, 2009)
2. but also more importantly, the BILOU scheme brings more a priori knowledge about the structure of a natural language in trainable model.

As a result of the above, the second fine-tuning stage is defined as training to solve 29 tasks of 5-class token classifications. The trainable algorithm consists of:

- the shared hidden layer (Transformer base) fine-tuned on the previous stage
- 29 different time-distributed dense layers (neural heads) initialized randomly.

The total loss function for the trainable algorithm is formulated as the sum of particular loss functions (named entity losses) associated with each of 29 neural heads. In this case the key problem is determining the named entity loss function. We propose a special loss as weighted combination of the usual categorical crossentropy for multiclass classification and the dice loss for binary classification **O vs. non-O** (i.e. all entity tokens including the Beginning, the Inside, the Last, and the Unit are opposed to the Outside):

$$NEL = -\alpha \cdot \sum_{i \in T} (\mathbf{y}_i \cdot \log(\mathbf{p}_i)) - \frac{2 \cdot \sum_{i \neq O} \mathbf{p}_i \cdot \sum_{i \neq O} \mathbf{y}_i + \gamma}{\left(\sum_{i \neq O} \mathbf{p}_i\right)^2 + \left(\sum_{i \neq O} \mathbf{y}_i\right)^2 + \gamma} + 1.0 \quad (6)$$

where $T = \{O, B, I, L, U\}$ is the set of all BILOU tags for the named entity, α is the weight of the cross-entropy item, and γ is the smoothing factor in the nominator and denominator of the dice loss.

The dice loss item is included in the total formula of the named entity loss to reduce the influence of the background-object label imbalance in data (evidently, frequency of the Outside tag is far greater than frequency of all entity tags, and it is a severe issue). As is well known, the dice loss has the class re-balancing property (Li et al., 2020b). Thus, the proposed named entity loss combines two advantages:

- the dice loss item attaches the robustness to imbalance,
- the categorical cross-entropy loss item accounts for the BILOU scheme which leads to better modeling the multi-token entities.

It should be clarified that the first advantage is important for the named entity recognition task, since labels of any named entity class are very unbalanced (the number of words labeled as entities is extremely less than number of non-entity words). However, the dice loss conforms to the binary classification problem, while the BILOU scheme can be implemented as the multiclass classification problem only. Consequently, we need to preserve the multiclass component in the formulated loss function. This implies the significance of the second of these advantages.

3.3 Rescoring with Viterbi algorithm

The posterior probability distribution $P(L|W)$ of the BILOU tag sequence L given the input word sequence W is calculated using Bayes' theorem:

$$P(L|W) = \frac{P(W|L) \cdot P(L)}{P(W)}, \quad (7)$$

where $P(W|L)$ is an observation likelihood estimated by the corresponding neural head of the multi-head NER from subsection 3.2, $P(L)$ is a prior distribution of BILOU tags determined by the language structure, and $P(W)$ is a marginal distribution which does not depend on any BILOU tag sequence (hence we can neglect it). Thus the best way to find an optimal sequence of the BILOU tags is based on well-known Viterbi search:

$$L^* = \arg \max_L (p(W|L)P(L)). \quad (8)$$

The description of Viterbi algorithm is trivial and can be found in various papers and books on natural language processing and speech recognition (for example, see (Jurafsky and Martin, 2008)). Nevertheless, it would be interesting to describe a technique for forming the prior distribution of BILOU tags. The fact is that the input sequence does not consist of true words. Text units of the sequence are sub-words built with a byte pair encoding according to the tokenization algorithm for the BERT model. Each true word can consist of one or multiple such sub-words. Correspondingly, we define four cases:

- the sub-word is equal to the true word,
- the sub-word is the first part in the true word;
- the sub-word is the middle part in the true word;
- the sub-word is the last part in the true word.

These cases are illustrated by the example 9. The true word "*импичмент*" (in English, "impeachment") is represented by one sub-word, but another true word "*Руссефф*" (in English, "Rousseff", which is the surname of the 36th president of Brazil) consists of three sub-words, i.e of the first, middle and last parts of the true word, accordingly.

(9) *Президенту Бразилии Дилме Руссефф грозит импичмент*
 президенту брази ##лии ди ##л ##ме рус ##сеф ##ф грозит импичмент
 'Possible tokenization cases (the SberDevices RuBERT-large tokenizer is used)'

Possibility transitions from one BILOU tag to another are different for each of these cases, and the prior distribution $P(L)$ is defined differently too. The discrete-time Markov chains for these four cases are shown on Figure 2.

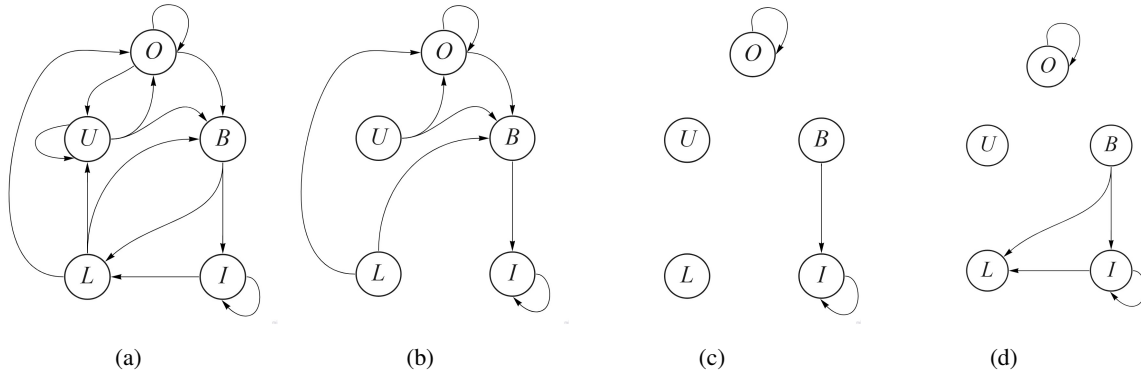


Figure 2: Four discrete-time Markov chains, which describe the BILOU tag transitions for: (a) the sub-word equals the whole word, (b) the first, (c) the middle, and (d) the last sub-word in the word.

Using the Viterbi search instead of the simplest greedy search allows us to rescore the output probabilities of the neural head accounting for some a priori knowledge, and the abovementioned way to determine the prior distribution of BILOU tags rationally specifies this knowledge.

4 Analysis and discussion

We evaluated the quality of several versions of CoNER on the NEREL dataset (Loukachevitch et al., 2021) in the context of the abovementioned RuNNE competition (Artemova et al., 2022). The F1 score is the relevance criterion of the quality, and it allows us to compare different NER algorithms to confirm or deny our hypothesis about effectiveness of the contrastive fine-tuning. Nevertheless, to explain the reasons of effectiveness of the the contrastive fine-tuning in CoNER, we need more than just comparing its F1 score to other NERs. In addition, it is important and necessary to analyze CoNER’s mistakes in an attempt to find some patterns and regions of the algorithm errors. These issues will be the subject of further discussion in this section.

4.1 CoNER vs. standard NER: what’s better?

We proposed the described two-stage fine-tuning to improve a NER quality. We explained the theoretical basis of the improvment. However, practice is the criterion of truth. We organized a series of experiments to compare two fine-tuning schemes:

- **standard NER:** fine-tuning BERT as a sequence classifier only;
- **CoNER:** two-stage fine-tuning BERT including the contrastive-based learning as the first stage.

Both fine-tuning schemes were started from three pretrained BERT models: the DeepPavlov RuBERT and two variants (base and large) of the SberDevices RuBERT. Here is a brief note about the difference between these pretrained models. The DeepPavlov model was trained on the Russian part of Wikipedia and news data. The SberDevices team added the Taiga corpus and a fiction corpus into the training data for both of its models. The SberDevices RuBERT-large is larger than the other two models. Also, in contrast to the DeepPavlov RuBERT, all the models of SberDevices are lowercased.

According to the description in Section 1, there were two formulations of the problem, or two kinds of tasks: the main task and the few-shot task. All entity classes in the few-shot formulation were very rare in the training data, which led to a significantly greater imbalance of data set in relation to these entities.

Both tasks were solved using the same NER algorithm which was trained on the common training set. After that, the quality of this NER for solving any task was evaluated on the same test inputs, but test labels for each task were different depending on the corresponding entity classes. The results of the experiments in the main formulation (main results) are presented in Table 1, and the results of analogous experiments for the few-shot task (few-shot results) are presented in Table 2.

Type of pre-trained model	Standard NER	CoNER
DeepPavlov RuBERT	0.7202	0.7425
SberDevices RuBERT base	0.6931	0.7233
SberDevices RuBERT large	0.7089	0.7113

Table 1: Main results (F1-macro scores) after different fine-tuning schemes from different pre-trained BERT models on the test data (i.e. at the RuNNE test phase).

Type of pre-trained model	Standard NER	CoNER
DeepPavlov RuBERT	0.3231	0.4037
SberDevices RuBERT base	0.3320	0.5099
SberDevices RuBERT large	0.4372	0.5256

Table 2: Few-shot results (F1-macro scores) after different fine-tuning schemes from different pre-trained BERT models on the test data (i.e. at the RuNNE test phase).

Results of CoNER appear to be better for either type of the task (main and few-shot). We used the Wilcoxon signed-rank test for statistically significant acceptance of this statement. To test the null hypothesis that there is no difference between quality measurements of a standard NER and CoNER, we applied the two-sided test. Evidently, size of both measurement sets is 6 as sum of 3 samples for the main task and 3 samples for the few-shot task. As a result, the test statistic was 0.0 with p -value of 0.03125. Hence, our null hypothesis was rejected at a confidence level of 0.05, and the differences in quality were confirmed.

The following conclusions and summary can be made from these experiments:

1. CoNER is better than a standard NER, and inclusion of contrastive learning in the fine-tuning scheme improves the generalization ability of any NER in any formulation of the problem.
2. This advantage of CoNER is more apparent for the few-shot task.
3. Using case-insensitive pretrained model similar to all SberDevices models reduces the NER quality for the main task. However, it likely is more effective in the few-shot formulation of the problem owing to suppression of the over-fitting on very imbalanced text datasets.

For the final submission to the RuNNE leaderboard, we selected CoNER based on the BERT model pretrained by the DeepPavlov team. This submission turned out to be the third out of 10 in the leaderboard for main task, and the eighth out of 10 in the few-shot formulation of the problem.

4.2 Why does contrastive-based fine-tuning matter?

It can be seen that CoNER is better than a standard NER, and the contrastive fine-tuning works. Nevertheless, we would like to explain these results. In section 2, we proposed a two-stage fine-tuning with the contrastive first stage based on our supposition that Siamese neural network as a typical kind of the contrastive learning algorithm has better discriminative ability in comparison to a standard classifier.

As they say, "a picture is worth a thousand words", and this picture is shown on Figure 3. It illustrates the compactness of the entity word representations in different feature spaces obtained from the sequence output of different BERT models. MONEY is used as a typical example of named entity class in this figure. As Figure 3a demonstrates, entity words and other words cannot be well separated in the usual pretrained BERT embedding space. Figure 3b shows that they have become better separated after fine-tuning BERT as standard sequence classifier. Better, but not by much. And only inclusion of the contrastive learning in the fine-tuning scheme, i.e. fine-tuning BERT as Siamese neural network, significantly increases the compactness and separability of entities in the word embedding space (this effect is clearly visible in Figures 3c and 3d).

Besides the abovementioned visual explanation, we analyzed the quality of representations of entity

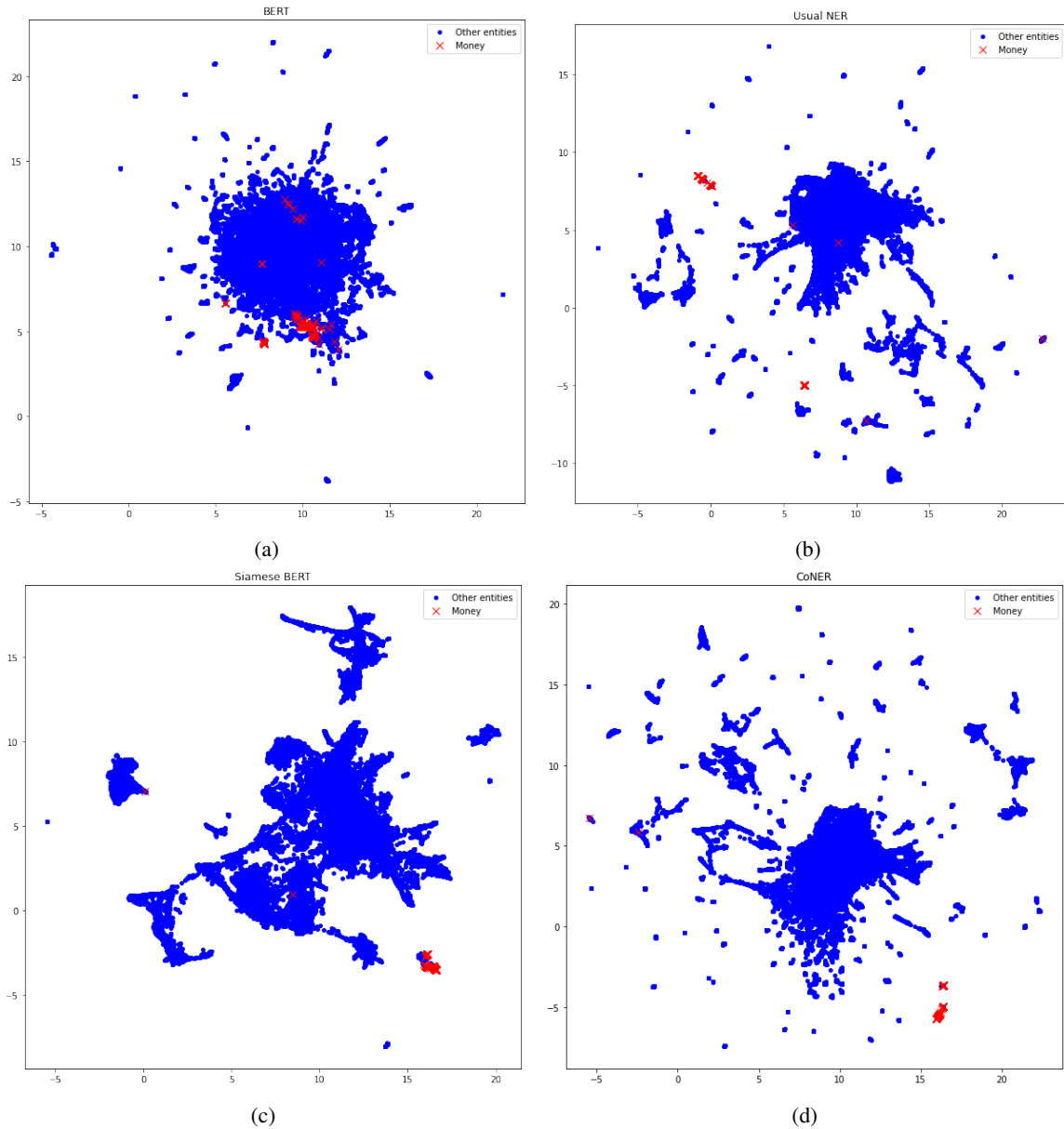


Figure 3: UMAP projections of 768-dimensional contextual embeddings of test words associated with MONEY entities and everything else, where (a) the embeddings before any fine-tuning RuBERT, (b) after fine-tuning as the standard NER, (c) after the first and (d) second stage of the proposed two-stage fine-tuning. The red crosses present words of money entities, and the blue dots are everything else, i.e. words of other entity classes and various background words.

words with and without the contrastive-based fine-tuning stage using the Silhouette Coefficient, well-known as a representative metric for the clustering performance evaluation. We formulated a hypothesis that the contextual word embeddings generated by BERT from CoNER allow us to build a more separable entity space compared to BERT from a standard NER. To test this hypothesis, we did the following steps:

1. We calculated contextual word embeddings from the last hidden output of the standard NER for all words in the test set. Thereby, we built a new feature space of words from named entities.
2. We separated all these words in the formed feature space into two clusters according to manual entity labeling: entity and non-entity. We did it for each of 29 entity classes, and we obtained 29 kinds of clustering.

3. We evaluated the Silhouette Coefficient for each kind of clustering from 29. Thus, we formed 29 measure samples.
4. Then we repeated these three steps with CoNER instead of the standard NER. As a result, we got yet another measure sampling with 29 elements.

The two measurement samplings are shown in Table 3. The superiority of representations from CoNER is observable to the unaided eye. Nevertheless, we applied the dependent *t*-test for paired samples for statistically significant acceptance of difference between these two samplings (they comes from a normal distribution, and their sizes are too large to effectively apply the Wilcoxon signed-rank test). The test statistic was -2.2866 with *p*-value of 0.029985. Hence, our null hypothesis that the contrastive-based fine-tuning does not improve separability of entities in the embedding space was rejected at a confidence level of 0.05, and the differences in quality were confirmed. The inclusion of fine-tuning BERT as Siamese neural network matters for BERT-based NER.

Entity class	Word number in training texts	Standard NER	CoNER
AGE	1506	0.4363	0.4507
AWARD	757	0.4959	0.5014
CITY	1432	0.4678	0.4639
COUNTRY	2671	0.4847	0.4788
CRIME	482	0.4064	0.4443
DATE	7087	0.4042	0.3886
DISEASE	53	0.3022	0.4220
DISTRICT	207	0.5490	0.5297
EVENT	5002	0.3102	0.3123
FACILITY	893	0.5049	0.5196
FAMILY	34	0.3242	0.4240
IDEOLOGY	368	0.4638	0.4510
LANGUAGE	45	0.3438	0.4014
LAW	1609	0.4601	0.4750
LOCATION	462	0.5308	0.5331
MONEY	635	0.4982	0.5240
NATIONALITY	434	0.4716	0.4271
NUMBER	1410	0.4745	0.4597
ORDINAL	710	0.5275	0.5456
ORGANIZATION	7127	0.4638	0.4588
PENALTY	73	0.4538	0.5251
PERCENT	216	0.5028	0.5166
PERSON	8063	0.3553	0.3439
PRODUCT	398	0.3594	0.3636
PROFESSION	7970	0.4314	0.4259
RELIGION	101	0.5377	0.5609
STATE OR PROVINCE	485	0.5562	0.5741
TIME	693	0.4993	0.4914
WORK OF ART	74	0.4396	0.4796

Table 3: The Silhouette Coefficient as the quality of different entity classes representation in the feature space generated with CoNER and the standard NER. Entity classes with better value of the Silhouette Coefficient for the standard NER are grey colored. Entity classes for the few-shot formulation of the problem are bolded.

4.3 When does CoNER make errors?

CoNER is good, but not the best according to the RuNNE competition leaderboard. CoNER makes some errors. Confucius said that *"people make errors according to the type of person they are. By observing their errors, you can understand humaneness"*. Similarly, by observing algorithm errors, we can understand its generalization ability and some patterns of its work.

Most errors can be divided into two types:

1. **The algorithm does not recognize nested entities of same entity class.** For example, the phrase *"Центральный комитет Коммунистического союза молодёжи Китая"* (in English, "the Central Committee of the Communist Youth League of China") describes the organization, and also it contains three nested organizations and one nested location. Nested organizations are *"Центральный комитет"* (in English, "the Central Committee"), *"Коммунистического союза молодёжи Китая"* (in English, "the Communist Youth League of China"), and *"Коммунистического союза молодёжи"* (in English, "the Communist Youth League"). The nested location is *"Китай"* (in English, "China"). And our CoNER correctly recognizes the "parent" organization and its nested location, but it cannot find any "organization-in-organization".
2. **The algorithm come across an ambiguous manual labeling.** For example, the full text *"пресс-служба филиппинского президента"* (in English, "press service of the Philippine President") was labeled as organization, but the same words *"пресс-служба"* (in English, "press service") were not labeled as a part of the organization entity in the text *"пресс-служба Светлогорского городского суда"* (in English, "press service of the Svetlogorsk City Court").

The first disadvantage is not related to any fine-tuning scheme, and it was determined by the common architecture of proposed solution: entity outputs of neural network were developed without consideration of entity nesting of the same class (we considered that only entities of different classes could be nested). We are going to improve our algorithm on the basis of a special syntactical post-processing of the recognized entity that allows to find nested entities of the same class using noun groups in the "parent" entity.

The second disadvantage is not really significant, because it was explained by the incorrect manual labeling. Furthermore, this effect can demonstrate greater robustness of CoNER, because it does not allow us to discover non-existent patterns by over-adapting to manual labeling.

Also, the algorithm makes a lot of false negative errors for all entities in the few-shot task. This may be explained by the under-fitting effect for very rare entity classes.

5 Conclusion. What's next?

We have confirmed our hypothesis about the contrastive fine-tuning for the NER task. We have also successfully performed error analysis and can apply this to improve our approach.

Further work may include two directions. First, recognition of nested entities of the same class will be implemented (for example, using a special syntactical-based postprocessing). Second, modeling of linguistic concept hierarchy (morphology - syntax - semantics) using a hierarchical multitask learning on both fine-tuning stages will be analyzed, because this technique can increase the generalization ability and reduce the vanishing gradient problem.

Acknowledgements

The work is supported by the Mathematical Center in Akademgorodok under the agreement No. 075-15-2022—282 with the Ministry of Science and Higher Education of the Russian Federation.

Especially the author would like to thank his student and fellow-thinker Alexey Pauls for his help and support.

References

- Ekaterina Artemova, Maksim Zmeev, Natalia Loukachevitch, Igor Rozhkov, Tatiana Batura, Pavel Braslavski, Vladimir Ivanov, and Elena Tutubalina. 2022. RuNNE-2022 Shared Task: Recognizing Nested Named Entities. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog”*.
- Sarkar Snigdha Sarathi Das, Arzoo Katiyar, R. Passonneau, and Rui Zhang. 2021. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. *Computing Research Repository*, arXiv:2109.07589.
- Dan Jurafsky and James H. Martin. 2008. *Speech and language processing - an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall; 2nd edition, Hoboken, New Jersey, USA.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. // *Advances in Neural Information Processing Systems*, volume 33, P 18661–18673, online. Curran Associates, Inc.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020a. A Unified MRC Framework for Named Entity Recognition. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 5849–5859, Online. Association for Computational Linguistics.
- Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020b. Dice Loss for Data-imbalanced NLP Tasks. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 465–476.
- Natalia V. 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. // *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2021)*, P 876–886, Held Online. Association for Computational Linguistics.
- Lev Ratnikov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. // *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, P 147–155, Boulder, Colorado, USA. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, P 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Nam N. Vo and James Hays. 2016. Localizing and Orienting Street Views Using Overhead Imagery. // *Proceedings of the 14th European Conference on Computer Vision*, P 494–509, Amsterdam, The Netherlands.
- Renjie Zhou, Qiang Hu, Jian Wan, Jilin Zhang, Qiang Liu, Tianxiang Hu, and Jianjun Li. 2022. WCL-BBCD: A Contrastive Learning and Knowledge Graph Approach to Named Entity Recognition. *Computing Research Repository*, arXiv:2203.06925.

The dataset for presuicidal signals detection in text and its analysis

Igor Buyanov
FRC CSC RAS
Moscow, Russia
buyanov.igor.o@yandex.ru

Ilya Sochenkov
FRC CSC RAS
Moscow, Russia
sochenkov@isa.ru

Abstract

The paper says about dataset for presuicidal signal detection in Russian posts from social media. To the best of our knowledge, it is a first dataset of a such type for this language. We develop a collection methodology and conduct linguistic analysis of completed dataset. We also build a classification baseline with machine learning models to solve the detection task.

Keywords: suicide; dataset; methodology; natural language processing

DOI: 10.28995/2075-7182-2022-21-81-98

Датасет для задачи распознавания пресуицидальных сигналов в тексте и его анализ

Буянов Игорь
ФИЦ ИУ РАН
Москва, Россия
buyanov.igor.o@yandex.ru

Соченков Илья
ФИЦ ИУ РАН
Москва, Россия
sochenkov@isa.ru

Аннотация

В статье представлен датасет для распознавания пресуицидальных сигналов в постах на русском языке в социальных сетях. Насколько нам известно, это первый датасет такого типа на русском языке. Мы разработали методологию сбора и провели лингвистический анализ полученного датасета. Мы также провели эксперименты по распознаванию пресуицидальных сигналов методами машинного обучения.

Ключевые слова: суицид, датасет, методология, обработка естественного языка

1 Introduction

Despite the rapid development of technologies that have significantly improved the living conditions of many people, the WHO reports 800,000 thousand annual suicides worldwide [13]. Suicide, in addition to tangible economic losses for the state, is hard to experience by the surrounding people and the action that can no longer be corrected. Judging by the fact that a recent study showed an increase in the level of depression [14], the problem of suicide will become more serious, as depression is considered to be one the suicidal factors [16] [17].

There are various non-profit organizations both in Russia and abroad that prevent such a terrible outcome by searching for potential suicides and carrying out preventive actions with them. The main source of search for such people is social networks such as Twitter, VK, Instagram, Telegram, etc., where people, mostly young ones, along with memes, sometimes post their experiences, even very frank.

Often, for people on the verge of suicide, expressing their feelings on a social network is a kind of valve that allows them to relieve the tension a little. In addition to direct expressions, a person who has decided to commit suicide sometimes leaves notes about his decision with information about the place and the chosen method. If such information could be detected immediately, it would be possible to save these people. In a less extreme case, it would be possible to track the individual problems like early

depression or emerged physical self-harm before these problems severely damage individual mental health.

In recent years, a large number of papers have been published where the authors study the problems of detecting depressive behaviour based on data from social networks. Unfortunately, most of these works concentrate on English and concern the prediction of certain outcome like whether the person will commit a suicide in a predetermined time period. In this work to the best of our knowledge, we present a first Russian language dataset built from Twitter that is dedicated to a study of signals that people shows on their road to a possible decision of suicide.

As a result of the study, the following results were obtained:

- We collect a dataset, containing texts of messages in Russian from personal pages showing suicidal intentions or close to this condition. The dataset contains markup on the presence of features by which volunteers assess the condition of people.
- We discovered some language characteristics that are specific for people with a risk to commit a suicide at least on Twitter.
- We proposed the baseline implementations solving the aforementioned tasks of presuicide signal detection. The code and dataset are available.¹

2 Related work

Applying the NLP techniques in the mental health domain is vastly possible with access to social media data. As a common source of data including post texts, the researchers utilize Reddit and Twitter. The former has a subcategory that is dedicated to a mental health problem so sometimes users directly report their diagnosis there which can be used to build a quality dataset. Almost the same happens on Twitter where users may post their diagnoses to find emotional support [8]. However, these posts had to be verified in order to be sure that the post contains no jokes, sarcasm and other unrelated phenomenon [18] [19].

This approach allows researchers to build a dataset for the identification of users having depression or PTSD [4] [18], a dataset with signs of depression [5] based on which the task of Early depression detection (eRisk) was organized, and a unique suicidal dataset [6] created from died and survived from committing suicide person's Twitter account. A list of currently available datasets for the mental health domain can be found in chapter 3.1 of a survey [7].

In this work [1] authors show that there is a statistical value between mental health and using Offensive Language. Again, the source of the data was Reddit.

Another dataset building method is to create a questionnaire application based on popular social network like Facebook. The users, who want to take a participation, give agreement under Terms of Services to collect their publically available data such status text, gender, age, etc. This approach was applied to study linguistic difference in user's personality [20].

Speaking of Russian language based works, it's worth to mention the paper, in which authors collect the depression posts from Vkontakte by utilizing a list of depression-related keywords and provide analysis of collected data [21]. In other work [22], authors managed to collect essay that was written on neutral topic by persons with a diagnosed depression. They provide analysis of dataset by showing the difference in a set of depression markers between depressive essays and control ones.

3 Task definition

The task set in the study is as follows. Having submitted the text to the input, the machine learning model should assign the text to one of five categories. During the paper we will refer to categories as next indices.

1. **Texts describing negative events that occurred with the subject in the past or in the present** — messages that are factual, describing negative moments that can happen to a person, such as attempts and facts of rape, problems with parents, the fact of being in a psychiatric hospital, facts of self-harm, etc.

¹ https://github.com/Astromis/research/tree/master/presuicidal_detection_dataset

2. **Current negative emotional state** — messages containing a display of subjective negative attitude towards oneself and others, including a desire to die, a feeling of pressure from the past, self-hatred, aggressiveness, rage directed at oneself or others.
3. **Messages about the intention of suicide** — messages containing an explicit declaration of suicidal actions. Messages that contain questions about suicide methods also fall into the same category.
4. **Messages with a suicidal theme** — the text of messages that are not directly related to the user but have a suicidal topic.
5. **Neutral** is the category in which messages that are not included in the above list fall.

Here we explain how we form these categories. In the course of the work of the non-profit organization, volunteers process accounts in social networks, in the post of which a third-party search engine found matches with keywords that carry a suicidal meaning. Processing consists of searching account posts containing signals about the possible presence of suicidal behaviour. Such signals can be indirect, such as, for example, stories about constant problems in the family or a university, and direct — the clear expression of a suicide intension. After evaluating the founded signals, the volunteer assigns to a particular user his suicidal status having three levels: low, medium, and critical (the highest level). The formulation of these categories based on volunteer’s needs when they try to classify the user status.

The first category was formed from the considerations that negative events can leave an emotional trigger that can destabilize a person's psyche, increasing the likelihood of suicide if such thoughts arise. The more such triggers, the more vulnerable a person is. The second category is an indirect indicator of a person's mental state, which is also cumulative – if the density of messages with similar content increases, then the person becomes mentally unstable. The third category is self-explainable in a view of finding people with suicidal behaviour. Sometimes people don’t expose direct emotions but uses death-related poetry or expressions. We can’t include it in previous categories so we allocate a fourth one.

Notice that the second category is similar for the more general task of sentiment prediction where the task is to identify whether the text is either negative, neutral or positive, but in this work, we narrow the definition of text negativity. In our dataset, some texts also can be assessed as negative. It may be, for example, statements that a character from a game or TV series is annoying, but such negative texts do not carry meaningful information for our task.

4 The methodology of dataset creation

Using the collected database of annotated users, we download the texts of Twitter users' posts that had a medium and critical status. Further, all texts were annotated manually by several trained and guided non-psychologists annotators. At the time of writing, there were a critically small number of volunteers engaged directly in detecting users, and there was also no unified data collection software where volunteers could immediately mark messages with the necessary features. For this reason, outside people were hired and trained to annotate downloaded texts.

One of the benefits of the hired annotators is their personal responsibility that increase the quality compared to the crowdsourcing, and direct communication, that allows to give them a feedback on their work. Moreover, based on a feedback from the annotators, we constantly improve an instruction. The major drawback is the high cost, that didn’t allow us to annotate the dataset with an overlap so we don’t report inter-annotator agreement. We will remove this drawback in a future version of dataset as it is constantly improved.

We compile an instruction, which describes the categories, phenomena falling under the certain categories, as well as some general recommendations. The annotation was divided in several rounds. In each round the annotator receives data block consisting of 3-5 thousand texts, annotates it and sends it back. We manually verify 5% of each block and if the number of errors was no more than three cases per thousand examples, we accept the block and send the annotator a feedback.

Among the problems faced by the annotators, we can highlight attempts to interpret texts based on their own beliefs and personal experience, ambiguous meaning of some texts, texts containing complex phrasal expressions and sarcasm, and texts representing two classes.

To compensate an absence of inter-annotator agreement and ensure the quality of the dataset, after the annotation was finished, we apply a cleaning procedure using the TracIn [12] algorithm. Originally developed and tested in the field of computer vision, the algorithm can be adapted to any type of data, including text.

5 Analysis of the collected dataset

In this section, we provide some remarkable findings that we discovered during a dataset analysing process. First, Table 1 summarizes class distribution in a resulting dataset. We see that the neutral text is a majority class in a dataset despite the source of texts being persons with medium and high suicidal risk. From the perspective of our task, we, unfortunately, couldn't gather a comparable amount of texts that represent classes three and four so these categories will not be considered. However, the remaining categories also have a rather small number of examples compared to neutral texts, creating a strong imbalance of classes. This can be explained by the fact that the social network as a whole is not a "book of complaints" — people write there on various topics, including to distract themselves.

Class name	Amount of examples
Neutral text	27619
Current negative emotional state	2809
Texts describing negative events	2131
Messages with a suicidal theme	205
Messages about the intention of suicide	21

Table 1 – Class distribution of a collected dataset

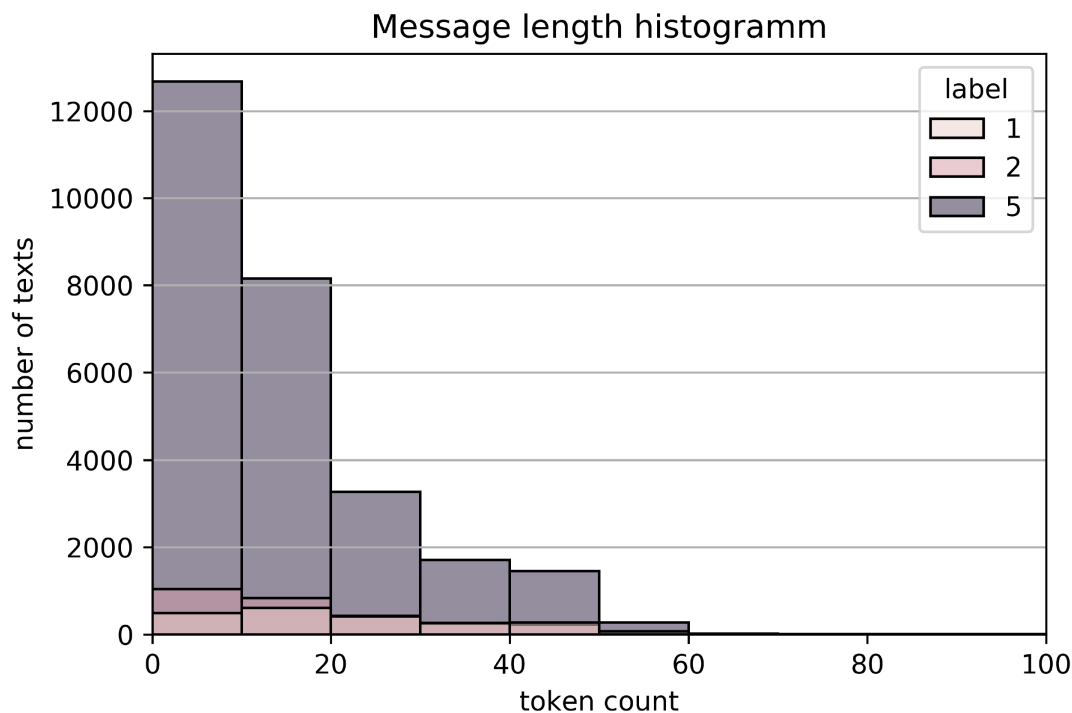


Figure 1 – Count distribution of texts by classes

Later, when we will build the baseline, we unite first and second classes in order to increase representativeness. To visualize how we transform classes for different purposes, we provide the Table 2.

Source set of categories	Set for dataset analysis	Set for baseline
{1,2,3,4,5}	{1,2,5}	1∪2,5

Table 2 – Dataset label transformation

In Figure 1 we can see the distribution of token length of text. As we can see, the distribution is convinced with our expectations that Twitter is a microblogging platform.

As part of texting, emojis have become an essential component in text communications. The main goal of emojis is to help better express person’s feels, intentions sometimes even art. During the collection of the corpus, we preserve emojis in text and do a basic analysis.

In our dataset we got, 12551 emojis with over 483 unique set. We chose the top 10 emojis by frequency and build a count to class ratio distribution that is depicted in Figure 2. We excluded the Triangular red flag because it appears that about 1200 times of usage is distributed through 23 posts. We see that the Loudly crying face has the highest value for the second class which is consistent with our expectation of class semantics. On the other hand, we see that the Pleading face which we might expect to be also important for the second class has the highest value for the five class. We also see that heart-related emojis also lies in five class that also looks coherent.

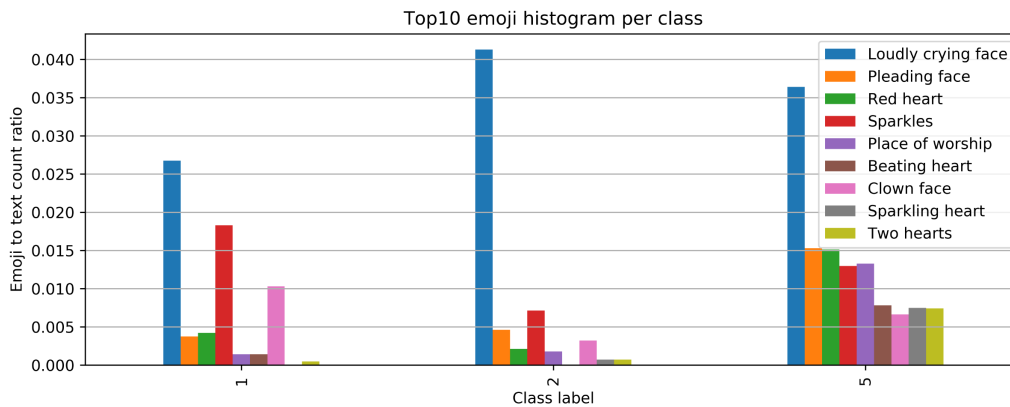


Figure 2 – Emoji to text count ratio distribution

We examine the lexicon of the dataset by a comparing method. To do this we took a more general twitter dataset that is used for sentiment analysis [9]. The first thing we investigated is unique words that characterize the language of the dataset. We acquire these words by substituting the set of general dataset words from a set of ours. We highlight the top 20 such words by frequency in Table 3.

Word	Count	Word	Count
мем	102	бсд	52
дазай	97	секси	50
геншин	82	ментальный	50
мью	81	пж	49
тикток	76	атсума	48
краш	74	рпш	48
рп	65	фд	46
хорни	63	осама	45
дилюк	57	эстетика	43
вайб	55	косплей	40

Table 3 – Specific words for our dataset

From this table, we can see special Twitter language like «мью» - transliterated short version of word “mutual” that means the person with which user has a mutual subscription with another one. We also can see some meme-words like «хорни» (transliteration from «horny» that means sexual arousal), shortcuts like “рп, пж, фд”. We also see a name «геншин» which is a name of online videogame Genshin Impact and a Bungou stray dogs shortcut “бсд” which is a name for manga and anime TV show. There are also names of persons from these two universals.

Another method of lexicon analysis we applying is recently proposed allotaxonomy[10]. The goal behind this method is a comparison of any two systems, entity of which has a rank and this rank is distributed according to Zipf law. As part of that comparison, rank-divergence metric was proposed to

understand the most important entities from two systems. Given two rank list R_1, R_2 of two systems with entity τ and hyperparameter α the rank-divergence metric can be computed as follow

$$D_{\alpha}^R(R_1||R_2) = \frac{1}{N_{1,2;\alpha}} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{[r_{\tau,1}]^{\alpha}} - \frac{1}{[r_{\tau,2}]^{\alpha}} \right|^{\frac{1}{\alpha+1}}$$

where $N_{1,2;\alpha}$ is a normalization factor (see the formula 7 from [10]). We again use the general sentiment analysis Twitter dataset as the opposite system. We construct an intersectional vocabulary from two corpora removing stop words and normalizing tokens. Then we compute the rank-divergence metric with $\alpha = 1/3$ as it was reported to deliver a good balance between entities with high and low ranks.

Table 4 shows the top 50 words sorted by Rank-Turbulence divergence and also shows the word rank in both corpuses that help two understand a direction of a rank change. From this table, we can clearly see that obsessive lexicon is a vital component of texts from our corpus. We might assume that this phenomenon relates to findings in work [1] in which authors show that there is a statistically significant relationship between mental health and offensive language usage.

index	word	rtd	rank in our corpus	rank in common one	index	word	rtd	rank in our corpus	rank in common one
0	бл**ь	3.070	7	175	25	обидно	1.241	868	131
1	завтра	2.429	70	8	26	день	1.240	8	4
2	блин	2.403	82	9	27	зачет	1.239	5573	334
3	на**й	1.968	44	745	28	заболеть	1.239	763	122
4	болеть	1.941	136	19	29	киев	1.227	12436	473
5	скучать	1.834	517	44	30	аниме	1.226	249	2786
6	сегодня	1.688	15	5	31	приходиться	1.218	387	83
7	человек	1.632	4	11	32	комп	1.217	1995	214
8	жаль	1.553	429	57	33	просто	1.215	5	10
9	жалко	1.553	1075	92	34	дома	1.209	232	60
10	пробка	1.516	3516	165	35	винд	1.208	25752	640
11	нг	1.463	2243	149	36	друг	1.208	27	76
12	тип	1.456	75	558	37	по**й	1.207	183	1435
13	школа	1.382	123	32	38	печально	1.194	5553	365
14	свой	1.375	6	14	39	заканчиваться	1.194	328	77
15	жизнь	1.360	16	46	40	снег	1.186	971	152
16	серия	1.346	556	87	41	ппц	1.181	9842	478
17	буквально	1.338	178	2093	42	выздоровливать	1.171	8477	460
18	печаль	1.338	2509	194	43	блин	1.169	22574	672
19	нету	1.332	2371	191	44	выходной	1.163	505	106
20	чел	1.309	219	2915	45	обновлять	1.153	4337	354
21	пятница	1.305	1683	169	46	скоро	1.138	125	42
22	е**ть	1.291	111	740	47	ау	1.138	820	32308
23	пи***ц	1.254	41	144	48	личность	1.138	457	7089
24	суицид	1.251	615	31353	49	порез	1.121	800	26578

Table 4 – Rank-Turbulence divergence values for the top 50 words

In addition, we compute a thematic value characteristic (TVC) [11]. TVC represents the value of a word w for some particular topic σ compared to all other topics in a given corpus c . A TVC value ΔI^+ can be computed as next

$$\Delta I(w, c, \sigma) = IDF(w, c \setminus \sigma) - IDF(w, \sigma)$$

$$\Delta I^+(w, c, \sigma) = \Delta I(w, c, \sigma) * X(\Delta I(w, c, \sigma))$$

where IDF is inverse document frequency, X is a Heaviside function. Table 5 contains words with the highest TVC value for three classes.

Finally, we examine the common (pseudo-)syntactic patterns of the sentences by mining POS trigrams associated with a certain label. We use the Russian POS tagger from the NLTK package. Having got the POS tags we create trigrams and then compute the PMI score between each trigram and text label where a certain trigram occurs. In Table 6 a list of the top 10 trigrams² for three labels is presented.

² <https://yandex.ru/dev/mystem/doc/grammemes-values.html>

To explore the statistical significance of these findings we compute the Mann–Whitney U test statistics for these tag sets, with values being presented in Table 7. As we can see the difference between the tagset for 1 and 2 classes against 5 class and vice versa has statistical significance, although the difference between class 1 and 2 has not.

	Class 1	Class 2	Class 5		Class 1	Class 2	Class 5
0	прл	сато	мью	10	диагноз	прорыдать	дажай
1	антидепрессант	чудовище	чують	11	селфхармить	упорно	петь
2	рвота	усталый	сезон	12	препарат	рас***рить	солнышко
3	больничка	медосмотр	вайб	13	желчь	здохнуть	фанфик
4	галлюцинация	поплакать	хорни	14	выстраивать	ничтожный	спи**ить
5	побочка	подпускать	добавлять	15	тревожка	пусто	геншин
6	бессонница	забываться	картинка	16	трезвый	шататься	рт
7	частичка	унижение	читатель	17	биполярка	уе**сь	косплей
8	кп	подавлять	вкус	18	до**ывать	кулак	мило
9	порез	комок	ау	19	перечить	будовать	тикток

Table 5 – Top 20 words by TVC

Class 1	Class 2	Class 5
V S-PRO S	V S-PRO ADV	PR V S
CONJ S-PRO S-PRO	CONJ V S-PRO	CONJ ADV PR
CONJ V S-PRO	PR A-PRO=m S	A=p S NONLEX
S-PRO ADV-PRO V	ADV V S	S-PRO PR A-PRO=p
S CONJ PART	S NONLEX NONLEX	PRAEDIC V <none>
V V S-PRO	ADV V PR	CONJ S-PRO A=n
ADV-PRO S-PRO V	V CONJ S	A=n CONJ S-PRO
PART PART V	PRAEDIC <none> <none>	CONJ CONJ S
V CONJ PR	ADV V V	S S-PRO A=m
S CONJ PR	ADV V CONJ	S-PRO S-PRO CONJ

Table 6 – Postag trigrams for classes

	Class 1	Class 2	Class 5
1 class tagset	1	8.5e-1	1.8e-4
2 class tagset	7.0e-1	1	3.1e-3
5 class tagset	2.4e-3	2.4e-3	1

Table 7 – The Mann–Whitney U test results

6 The baseline classifier

In this work, we also provide a baseline for solving the established problem. In this section, we describe of a whole pipeline.

At first, we preprocess the dataset by removing all punctuation, set the case to lower, filter emojis and non-alphabetic characters. We also remove all text that contains only one token. As a vectorization procedure, we employ count vectorization which is essentially a vector with dimensionality equal to

power of vocabulary and entities representing the number of times a certain word occurs in the given text. We also used BERT distilled model for the Russian language named rubert-tiny2³. As embeddings, a CLS token from the last hidden state were used. As we mentioned before, due to the lack of third and fourth categories, we exclude them from considerations. Moreover, as we mentioned before, we combine first two classes into one that makes it more representative. We assume that this new class will carry certain negatively gained emotions anyway so it might be distinguished from the neutral class.

			precision		recall		f1		f1 macro	
			mean	std	mean	std	mean	std	mean	std
method	vec	class								
Isolation Forest	BERT	0	0.558	0.351	0.000	0.000	0.001	0.000	0.334	0.000
		1	0.500	0.000	1.000	0.000	0.667	0.000		
	Count	0	0.000	0.000	0.000	0.000	0.000	0.000	0.333	0.000
		1	0.500	0.000	1.000	0.000	0.667	0.000		
Local Outlier Factor	BERT	0	0.301	0.022	0.009	0.001	0.017	0.001	0.338	0.001
		1	0.497	0.001	0.979	0.003	0.659	0.001		
	Count	0	0.502	0.000	0.997	0.000	0.668	0.000	0.344	0.002
		1	0.768	0.033	0.011	0.002	0.021	0.004		
Logistic Regression	BERT	0	0.948	0.002	0.765	0.005	0.847	0.003	0.680	0.003
		1	0.382	0.004	0.777	0.011	0.512	0.004		
	Count	0	0.909	0.007	0.754	0.018	0.824	0.013	0.617	0.018
		1	0.313	0.021	0.598	0.030	0.410	0.024		
LogReg Stack	BERT	0	0.947	0.002	0.774	0.004	0.852	0.002	0.680	0.005
		1	0.391	0.010	0.770	0.006	0.518	0.009		
	Count	0	0.923	0.035	0.632	0.299	0.693	0.302	0.617	0.018
		1	0.292	0.078	0.656	0.190	0.384	0.065		
OneClassSVM	BERT	0	0.486	0.015	0.577	0.220	0.512	0.108	0.685	0.005
		1	0.491	0.013	0.401	0.202	0.414	0.130		
	Count	0	0.621	0.006	0.374	0.001	0.467	0.001	0.538	0.180
		1	0.552	0.002	0.772	0.007	0.644	0.003		
Random Forest	BERT	0	0.856	0.004	0.994	0.001	0.920	0.002	0.558	0.006
		1	0.770	0.032	0.112	0.007	0.196	0.010		
	Count	0	0.856	0.005	0.991	0.001	0.918	0.003	0.545	0.010
		1	0.671	0.028	0.098	0.012	0.171	0.019		
XGBoost	BERT	0	0.899	0.004	0.931	0.003	0.915	0.001	0.703	0.007
		1	0.548	0.018	0.445	0.016	0.491	0.013		
	Count	0	0.899	0.004	0.923	0.003	0.911	0.002	0.693	0.005
		1	0.516	0.016	0.442	0.009	0.476	0.009		

Table 8 – Experiment results

We experiment with several models including traditional classification methods like Random Forest, Logistic Regression, and XGBoost and models for outlier detection like Isolation Forest, Local outlier factor and, One class SVM. The motivation for the latter is a significant class imbalance, so we can view

³ <https://huggingface.co/cointegrated/rubert-tiny2>

a neutral class as a class representing text that person types being mentally stable. We assume that a number of mentally stable people largely outnumber the number of unstable ones. Moreover, we assume that even high suicidal risk person's messages are not always exposed informative signals [15]. On the other hand, texts from combined classes can be treated as a non-common case. Finally, we use a composition of Logistic regressions called stacking. We split the dataset into equal blocks, where each block consists of a full number of outlier class and an equally sized normal class. On all blocks, we separately train the logistic regression model. After that, we train a final logistic regression model on a whole dataset using predicted probabilities from early train models as features.

We also estimate class weights and set them as hyperparameter in classification models. Other hyperparameters we left as default.

As a metric we use precision, recall and F1-measure. We evaluate all models ten times each time mixing random state of models and train/test split. Table 8 summarize the results.

From this table, we can see that Isolation Forest and Local Outlier Factor doesn't work in this setting as perfect recall with a half precision says that classifier assigns one class for all examples. Another observation is that in almost all settings the BERT embeddings as expected outperform the simple count vectorization method except OneClassSVM for detecting both classes. The best result by precision shows RandomForest based on BERT, although Random forest shows the worst result by recall. The best recall showed the Logistic Regression. The OneClassSVM with count vectorizer shows the best F1 score for the first class. It's interesting, that tree based ensemble methods show high recall for zero class. Unlike Random Forest, BERT based XGBoost classifier shows a much better result by recall for the first class that leads to the best macro F1 across all settings and, thus, setting our baseline for the task.

7 Conclusion and future work

In this work, we introduce a new task of detecting messages that express some clues about possible person mental instability. We develop a methodology for collecting a Russian dataset from open data from social media. We also analyse the dataset and found various language features that characterize such texts. Finally, we investigate various settings to build a baseline classifier. Overall we see that this task is quite challenging as the highest precision we can archive is only 0.75 on the classification task. On the other hand, we see that outlier detection method One Class SVM shows the best performance by the F1 score for a class of interest so it might be a promising way to continue to work with this task in outlier detection setting. Nevertheless, the best macro F1 shows the XGBosst classifier. Probably, with accurate hyper parameter search it is possible to archive better results.

In the future, we plan to collect more data which will include not only the text but also images, audio and social interactions. We believe that multimodality brings new findings and ideas to better understand the behaviour of people with high suicidal risk and thus give us more accurate methods to find and help them.

Acknowledgements

We would like to thank to our data annotators Ermakova Darina, Ulyanova Irina, Kalinovskaya Tatiana, our volunteer Brodskaya Alexandra for user seeking and psychologist Elizaveta Kluchikova for meaningful consultation. The reported study was funded by RFBR according to the research project № 21-011-44242.

Reference

- [1] Ana-Maria Bucur, Marcos Zampieri, Liviu P. Dinu. An Exploratory Analysis of the Relation Between Offensive Language and Mental Health // Computing Research Repository. — 2021. — Vol. arXiv: 2105.14888. — version 2. Access mode: <https://arxiv.org/abs/2105.14888>
- [2] Siyang Liu et al. Towards Emotional Support Dialog Systems // Computing Research Repository. — 2021. — Vol. arXiv: 2106.01144. — version 1. Access mode: <https://arxiv.org/abs/2106.01144>
- [3] Ning Wang et al. Learning Models for Suicide Prediction from Social Media Posts // Computing Research Repository. — 2021. — Vol. arXiv: 2105.03315. — version 1. Access mode: <https://arxiv.org/abs/2105.03315>
- [4] Glen Coppersmith et al. CLPsych 2015 Shared Task: Depression and PTSD on Twitter // Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Denver, Colorado, 2015. — P. 31–39.
- [5] Losada D.E., Crestani F., A Test Collection for Research on Depression and Language Use. — Springer, Cham, 2016. — Vol. 9822
- [6] Sean MacAvaney et al. Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task // Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology. Online, 2021. — P. 70–80.
- [7] Muskan Garg. Quantifying the Suicidal Tendency on Social Media: A Survey // Computing Research Repository. — 2021. — Vol. arXiv: 2110.03663. — version 1. Access mode: <https://arxiv.org/abs/2110.03663>
- [8] Moreno, Megan A. et al. Feeling bad on Facebook: depression disclosures by college students on a social networking site. *Depression and anxiety*, 2011. — Vol. 28,6
- [9] Rubtsova U. (2012), An automatic construction and analysis of short text corpus (microblog posts) for the task of developing and training of sentiment classifier. [Avtomaticheskoye postroenie i analiz korpusa korotkikh tekstov (postov mikroblogov) dlya zadachi razrabotki i trenirovki tonovogo klassifikatora], Knowledge engineering and semantic web technologies [Injeneria znanii i tekhnologii semanticheskogo weba], Saint Petersburg, pp. 109-116
- [10] P. S. Dodds et al. Allotaxonomy and rank-turbulence divergence: A universal instrument for comparing complex systems // Computing Research Repository. — 2021. — Vol. arXiv: 2002.09770. — version 1. Access mode: <https://arxiv.org/abs/2002.09770>
- [11] D. A. Devyatkin et al. (2013) Method of thematic clustering of large-scale collections of scientific and technical documents. [Metod tematicheskoy klasterizatsii mashtabnikh kollektiy nauchno-tekhnicheskikh dokumentov], ITCS [ITiVS], Moscow, pp. 68-78
- [12] Garima Pruthi et al. Estimating Training Data Influence by Tracing Gradient Descent // Computing Research Repository. — 2020. — Vol. arXiv: 2002.08484. — version 3. Access mode: <https://arxiv.org/abs/2002.08484>
- [13] Dévora Kestel and Mark van Ommeren et al. Suicide in the world — World Health Organization, 2019. — Vol. 1
- [14] Bollen J. et al. Historical language records reveal a surge of cognitive distortions in recent decades. — Proc Natl Acad Sci USA, 2021 — Vol. 1
- [15] Cavazos-Rehg PA et al. A content analysis of depression-related Tweets. — *Comput Human Behav*, 2016 — Vol. 1
- [16] Craig J. Bryan and M. David Rudd, Brief Cognitive-Behavioral Therapy for Suicide Prevention. — Guilford Press, 2018 — Vol. 1
- [17] Popov U. V., A.A. Pichikov, Suicidal behavior in adolescents. [Suicidalnoe povedenie u podrostkov] — *SpecLit*, 2017 — Vol. 1
- [18] Glen Coppersmith et al. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses // Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, Colorado, 2015 — P. 1–10
- [19] De Choudhury M. et al. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media // Proceedings of the SIGCHI conference on human factors in computing systems, 2016 — P. 2098-2110
- [20] H. Andrew Schwartz et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach — *PloS one*, 2013 — vol. 8
- [21] Narynov S. et al. Dataset of depressive posts in Russian language collected from social media // *Data in Brief*, 2020 — vol. 29
- [22] Stankevich M., Smirnov I. et al. Predicting Depression from Essays in Russian // Proceedings of “Computational Linguistics and Intellectual Technologies” DIALOGUE, 2019 — P. 637-647

Appendix A Инструкция для разметчиков

Общая информация

Вам будет представлен ряд сообщений, собранный с просторов русскоязычного Твиттера в формате Excel таблицы. Необходимо их распределить в соответствии с ниже заданной классификацией. Разметка нужна для создания классификатора, который упростит волонтерам поиск людей, находящихся на грани самоубийства, для их последующего консультирования и оказания помощи. Без преувеличения можно сказать, что, выполняя эту работу, вы вносите вклад в спасение чьей-то жизни.

В случае если у вас возникнет вопрос как именно следует классифицировать сообщение, то необходимо такие сообщения сгруппировать отдельно, поставить ту метку, к которой склоняетесь больше всего, и описать ваши сомнения. Полученный файл прислать заказчику на супервизию.

Классификация

Жирным шрифтом выделены сами классы, а маркированным список указаны **не исчерпывающие** подклассы класса, которые призваны помочь составить представление о контенте. Критичность указывает приоритет класса при совместном появлении классов в одном сообщении.

1) [*Критичность: средняя*] **Исторические или текущие негативные события** – сообщения, носящие фактический характер, описывающие негативные моменты, которые могут произойти с человеком, такие как:

- попытки и факты изнасилования,
- проблемы с родителями (ненависть к ним, непонимание, алкоголики, насилие с их стороны),
- проблемы с друзьями/отношениями (отсутствие друзей, разрыв любовных и дружественных отношений, конфликты)
- издевательства в школе и травля,
- факты применения физической силы,
- факт нахождения в психиатрической больнице,
- психиатрический диагноз (депрессия("депра"), шизофрения, биполярное расстройство(биполярка), тревожность, СДВГ, ПТСР),
- факт употребления медикаментов (антидепрессантов, успокаивающих и т.д.)
- попытки в прошлом или фантазии о суициде,
- факт употребления наркотиков, алкоголя,
- проблемы со сном,
- проблемы со здоровьем
- факты самоповреждений - текст, в котором говорится о том, что человек причиняет себе физическую боль. Чаще всего это выражается в порезах.
- проблемы с питанием - анорексия, заявления о том, что стошнит после приема пищи, невозможность нормально есть
- бедность (личная или семейная)
- проживание с больным родственником
- выраженная низкая самооценка
- пережитое недавно эмоциональное потрясение
- факты криминального характера

2) [*Критичность: низкая*] **Текущее негативное эмоциональное состояние** – сообщения, содержащие отображение субъективного негативного отношения к себе и окружающим:

- заявления о том, что нет сил, терпения,
- желание умереть,
- ощущение одиночества,

- ощущение, что "все плохо",
- рассуждения о тщетности жизни,
- ощущение давления прошлого
- фантазии и высказывания желаний о порезах
- ненависть к себе
- агрессивность, ярость, направленная на себя или на других

3) [*Критичность: высокая*] **Сообщения о намерении суицида**, отличается от "желания умереть" именно декларацией действий, например, "завтра в 7 иду на железную дорогу, всем спасибо за внимание", «завтра набираю ванну и беру нож» или поиск способов типа "какую веревку выбрать" или "смертельная доза таблеток", «насколько глубоко надо резать вены»

4) **Суицидальная тематика** - все то, что как-то связано с суицидом, но трудно в определении или не попадает в другие категории. Например,

- "Зачем нужны парные луки, татухи и всякая лабудень, если можно просто совершить парный суицид"
- "нежелаете совершить со мной суицид?"
- "соверши чистое, весёлое и энергичное самоубийство"

5) **Сообщения, не имеющие отношения к суицидальной тематике**

Антипаттерны

- Выражение эмоций, связанных с сохранением жизни, следует отнести к положительным примерам

Замечания

1. Сообщения на иностранном языке необходимо пометать, как не имеющие отношения к суициду
2. Тег `<emoji></emoji>` обозначает эмоджи, использованные автором. Хотя они довольно сильно могут мешать, они необходимы для изучения взаимосвязи использования эмоджи с состоянием людей. Просьба проявить терпение
3. В случае, если встречается текст, который попадает под две категории, то следует выбирать ту, которая более критична. Критичность категории указана в ее описании. Например, фраза: «Я БОЛЬШЕ ТАК НЕ МОГУ!!! Очередь эти еб**ые флешбеки про изнасилование» несет в себе категории 1 и 2, но гораздо важнее то, что у человека есть факт переживания насильственных действий. Или же пример «Все, на**й все, вы все меня за**али! пошла на мост!», где пересекаются 2 и 3, но нам важно, что человек собрался совершить суицид.
4. При оценке сообщения недопустимо пытаться их интерпретировать, опираясь на личный опыт или знания – нельзя только лишь по твиту сказать действительно ли сообщение является попыткой привлечения к себе внимания, обманом или реальным положением дел. Посему, если текст попадает под одну из категорий, то необходимо его пометить этой категорией, вне зависимости от контекста.

Discourse-aware text classification for argument mining

Elena Chistova

FRC CSC RAS / Moscow, Russia
chistova@isa.ru

Ivan Smirnov

FRC CSC RAS / Moscow, Russia
ivs@isa.ru

Abstract

We show that using the rhetorical structure automatically generated by the discourse parser is beneficial for paragraph-level argument mining in Russian. First, we improve the structure awareness of the current RST discourse parser for Russian by employing the recent top-down approach for unlabeled tree construction on a paragraph level. Then we demonstrate the utility of this parser in two classification argument mining subtasks of the RuARG-2022 shared task. Our approach leverages a structured LSTM module to compute a text representation that reflects the composition of discourse units in the rhetorical structure. We show that: (i) the inclusion of discourse analysis improves paragraph-level text classification; (ii) a novel TreeLSTM-based approach performs well for the computation of the complex text hidden representation using both a language model and an end-to-end RST parser; (iii) structures predicted by the proposed RST parser reflect the argumentative structures in texts in Russian.

Keywords: Discourse parsing, RST, text classification, argumentation mining

DOI: 10.28995/2075-7182-2022-21-93-105

Классификация текстов с учетом дискурсивной структуры для анализа аргументации

Елена Чистова

ФИЦ ИУ РАН
Москва, Россия
chistova@isa.ru

Иван Смирнов

ФИЦ ИУ РАН
Москва, Россия
ivs@isa.ru

Аннотация

В работе демонстрируется эффективность автоматического дискурсивного анализа для анализа аргументации в текстах на русском языке. Улучшенный за счет применения современного метода построения неразмеченных риторических структур метод дискурсивного анализа применяется в классификации документов на примере двух подзадач анализа аргументации в соревновании RuARG-2022. Предлагаемый подход к классификации на основе структурной LSTM предусматривает обучение векторного представления текста, отражающего композицию его фрагментов в дискурсивном дереве. В ходе исследования показано, что: (1) учет предсказанной дискурсивной структуры позволяет улучшить качество классификации текста на уровне абзаца; (2) предложенный подход на основе TreeLSTM эффективен при обучении векторного представления абзаца с использованием языковой модели и автоматического дискурсивного анализатора; (3) предсказанные анализатором риторические структуры в целом отражают аргументативную структуру текстов.

Ключевые слова: Дискурсивный анализ, теория риторических структур, классификация текстов, анализ аргументации

1 Introduction

As an attention module of an advanced classification model traverses a complex sentence or a document sequentially, it may become confused as to which phrases pertain to the document's class and which represent a different class from the document's meaning, or whether certain phrases argue for or against the author's position. It is possible to uncover relations between text parts with discourse parsing. The

Rhetorical Structure Theory (RST [22]) is the discourse framework suggesting that texts have a hierarchical, connected structure, with both intra- and inter-sentential relations. A rhetorical tree shows how elementary discourse units (EDUs) and non-elementary units combine to form the overall meaning of a document (see Figure 1).

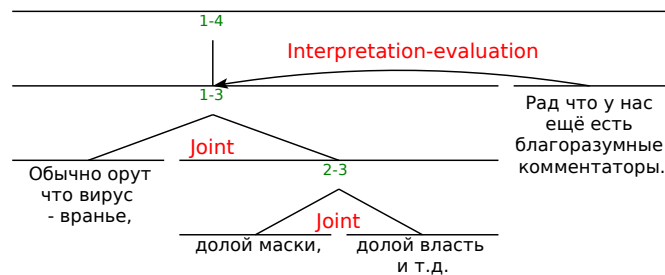


Figure 1: RST parsing result for a short example from RuArg-2022: *[Usually, they yell that the virus is a scam,]₁ [down with the masks,]₂ [down with the government, etc.]₃ [Glad there are still sane commenters out there.]₄*. In this example, there is only one mononuclear rhetorical relation, where the left constituent (EDU₁₋₃) is the nucleus, and the right constituent (EDU₄) is the satellite.

Text classification methods adopt one of two main approaches based on the discourse parsing: either (1) weighting tokens based on their position in an unlabeled discourse tree for lexicon-based analysis [4, 13, 32], or (2) combining phrases based on the discourse structure to determine an overall class score [16, 20]. These methods focus mainly on sentiment analysis; several studies have also identified a connection between the rhetorical and argumentation structures [8, 5, 10].

This paper investigates the impact of discourse parsing on document classification for argument mining in social media texts in Russian. We first improve the performance of the current RST parser for Russian by introducing it to the top-down paragraph parsing. Then, we investigate the text classification applying a Tree LSTM [31] module on the predicted discourse structures. We use this module to correct the predictions of the fine-tuned language model. The classification methods are tested on the RuARG-2022 [17] shared task.

Our contributions can be highlighted as follows:

- To the best of our knowledge, we are the first to explicitly analyze the effect of discourse among opinion mining and argument classification in Russian.
- We achieve a significant improvement in RST discourse parsing for Russian using a top-down algorithm for unlabeled tree construction at the paragraph level.
- We propose a new method to utilize RST discourse structure in Tree LSTM for paragraph-level text representation learning.

Our code is publicly available¹.

2 Related work

Rhetorical structure in text classification: A number of early studies investigated the possibility of opinion mining using shallow text structures derived from the discourse connectors vocabulary [24, 29] or based on the manual discourse annotations [2]. The development of automatic discourse parsers for English has strengthened research in this area. Finding the most common nucleus of rhetorical relations in each sentence, authors of [32] investigate whether the sentiment lexicon can be weighted based on RST structure. Assuming that the nuclei of each relation encapsulate the general idea of the text, they explain the lack of performance improvement through the poor accuracy of the early RST parser SPADE [30]. However, the results obtained with the SPADE and HILDA [28] parsers in [13] demonstrate an improvement in sentiment classification when weighting words based on the depth of corresponding subtree and nuclearities in it. Markov logic and the sentence-level discourse trees predicted by the

¹<https://github.com/tchewik/discourse-aware-classification>

HILDA parser are used in [34] to calculate the sentiment score using information about contrastive (Contrast, Concession) and non-contrastive (the rest of RST-DT relations) rhetorical relations occurring between elementary discourse units.

More recent approaches explore the integration of the explicit discourse structure in deep learning models. In [4], sentiment scores are propagated recursively up the RST tree to the root via a neural network with architecture specific for each parse, and scalar parameters related to particular relations are tuned. They do not construct latent representations of discourse units and also train a simplified version of the DPLP RST parser [15], focusing only on distinguishing contrastive and non-contrastive relations. Method [16] exploits the trainable representations of discourse units at all levels. The authors propose to build a shared text vector representation for a discourse tree node based on the composition of representations of individual EDUs and subtrees. A weighting of the importance of individual discourse units is automated through the attention mechanism. They test the method on multiple text classification tasks. In [9], the authors apply Tree LSTM to the unlabeled sentence-level RST trees with nuclearities. Binary Tree LSTM in their method does not process left and right children of the current node, but rather its predefined nucleus and satellite (or two nuclei). The use of the DPLP parser to construct structural neural networks is also demonstrated in [20]. They construct RecNN [12] and Tree LSTM [31]. To reduce the complexity of the neural network to be constructed, they consider individual sentences rather than EDUs as leaves of the discourse tree. In the representation of each discourse tree node, the text embedding and the rhetorical relation embedding are concatenated; the sentence embeddings are trained independently. In [19] the authors propose a Tree LSTM model similar to the one proposed in [9] with additional tree nodes augmentation. In their study, they predict the polarity of each EDU using dictionaries and word embeddings and found that incorporating embeddings leads to strong overfitting in the Tree LSTM models.

Argument mining using RST annotation: Argument mining is known to benefit from discourse analysis. It has been shown [8] that certain semantic groups of discourse connectors are indicative of either claims or premises and can be used to differentiate between the two. There are certain argumentative relations in RST that represent supportive, incentive, justification, and persuasion arguments, as outlined in [3]. Communicative discourse structure inspired by RST is used in [10] to categorize texts as being either argumentative or non-argumentative. The authors of [5] propose combining a BERT-based classifier with a gradient boosting model based on a rhetorical relation label in the root of the discourse tree. Examples of how the classifier on discourse relations corrects the predictions of BERT are given in order to illustrate how some RST relations, such as Evaluation or Antithesis, correlate with argumentative ones. TreeLSTM over RST structure is probed for argumentation mining in [6]. This module is used to obtain a vector representation of the text (the root of the rhetorical tree) based on EDU embeddings, which are formed by concatenating word, sentence, and part-of-speech tag embeddings.

In this work, we propose a TreeLSTM-based text classification method for argument mining. Current text classification methods using TreeLSTM over RST structures, usually designed for sentiment analysis, are subject to strong overfitting due to the high dimensionality of discourse unit embeddings trained jointly with the recursive neural module. The key difference between our work and previous work is that we do not train TreeLSTM from scratch in conjunction with the text encoder, but instead use the module to refine predictions of a high-performance sequential text classifier on documents with rhetorical structure.

3 Improving discourse parsing for Russian

This section describes the end-to-end RST parsing method we later use in text classification. We propose constructing unlabeled trees at the paragraph level by using a top-down approach, which improves the structure awareness of the recent discourse parser for Russian.

Method: RST parser for Russian recently proposed in [7] is proven to be highly accurate for relation classification and EDU segmentation, although its greedy bottom-up tree-building algorithm limits its overall performance for document parsing. However, the method takes on the challenge of segmenting

long texts into separate discourse trees despite weakly paragraph related tree boundaries, a feature of the Russian RST corpus RuRSTreebank [27] that disallows direct application of state-of-the-art unlabeled tree construction methods (1 document = 1 tree) developed for other languages. Therefore, for our experiments, we reproduce [7], but replace the sentence- and paragraph-level unlabeled tree construction methods in the parser with the recent top-down parsing approach proposed in [26] under the assumption that each paragraph corresponds to a separate subtree. As opposed to prior top-down discourse parsing methods [21, 33] which considered each span separately at each time step, the novel method allows for comparison of subtree candidates globally at the full-tree level by computing all span boundary representations in text at each time step and using beam search to find the best subtree candidate.

Data, Results, and Discussion: We use the standard RuRSTreebank corpus [27] for training and evaluation, focusing on two genres: news and blogs, and selecting 15% of data for the test. Since there is no available language model for long documents in Russian, we rely on character and pretrained word2vec embeddings for the initial representation of the document. For training on gold segmentation, we use the following parameters: beam size = 20, batch size = 4000 tokens. In Table 1, we compare the end-to-end discourse analysis performance at the different granularity levels between the system using greedy bottom-up paragraph parsing [7] and the one proposed in this study using micro-averaged standard Parseval metric [23]. In both cases, we use the same BiLSTM-CRF discourse segmentation model on pretrained ELMo embeddings, achieving 88.4% F1 on the test set. We use word2vec and ELMo pretrained models provided by RusVectors². For our structure-aware classification method, the parser’s most important feature is its ability to retrieve discourse structure regardless of labeled relations. The top-down approach improves the unlabeled tree construction (span identification) performance by 10.5% F1 at the sentence level, 10.4% F1 at the paragraph level, and 8.9% F1 at the document level, taking into account that the relations between paragraphs are in both cases detected by applying the same greedy bottom-up algorithm. The full end-to-end parsing performance increases by 10.6%, 7.0%, and 6.2% F1, respectively. We publish the source code for the end-to-end parser used in our experiments³.

Method	Sentence level				Paragraph level				Document level			
	span	nuc	rel	full	span	nuc	rel	full	span	nuc	rel	full
Greedy [7]	58.0	38.9	27.8	27.1	49.4	31.0	20.4	20.3	43.6	27.3	18.0	17.7
Beam search	68.5	50.6	38.1	37.7	59.8	38.8	27.5	27.3	52.5	34.2	24.2	23.9

Table 1: Performance of end-to-end RST parsing using different paragraph-level unlabeled tree construction methods

In this study, improving parsing performance at both the sentence and paragraph levels is crucial. Analysis of the RuArg-2022 dataset reveals that each example corresponds to a single automatically identifiable sentence. However, the sentence segmenter often fails to segment social media comments properly, because some sentences end with emojis, parentheses, ellipses, or no punctuation at all. In addition, social media users often write extremely long sentences that could be broken down into several grammatically correct shorter ones. Therefore, in some situations, it may be necessary to analyze inter-sentential discourse relations.

4 Discourse-aware classification method

In this section, we detail our proposed method for stance and argument classification, addressing the limitations of unstructured full-text classification methods. We discuss the pipeline-based framework for the classification of texts with or without recognizable rhetorical structure. The first stage involves fine-tuning the sequential model on the dataset including texts of different lengths and complexity. In the second stage, we freeze the base model and then train a discourse-aware neural module on top of it for the classification of texts with discourse structure.

²<https://rusvectors.org/>

³https://github.com/tchewik/isanlp_rst/releases/tag/v2.0

4.1 BERT

For text classification based on token sequences, we adapt the multitask baseline model architecture proposed by the competition organizers, where two outputs are being trained simultaneously in a classifier based on a language model. We use the DeepPavlov RuBERT Conversational⁴ along with BERT pooling to encode the document. This particular language model was chosen because it is pretrained on dialogue and social media texts, so it is well suited for encoding social media comments. The hidden representation is then passed through two fully-connected layers for stance and argument prediction; all parameters are trainable.

The model as it stands is used in the final pipeline for predicting labels for structure-lacking sentences (EDUs). It is also used in the structure-aware model for the initial encoding of discourse tree nodes.

4.2 RST-LSTM

RST parsers represent discourse as a binary constituency tree. If the binary discourse tree is traversed from the bottom up, information from the left and right constituents can be combined to represent the tree node at the upper level and all the way up to the root. Our structure representation module is based on the Binary Tree LSTM network [31]. In Binary Tree LSTM, a non-elementary discourse unit’s hidden and cell states are determined by the hidden and cell states of its left and right constituents rather than the sequence of words inside it. It allows computation over self-contained phrases within a complex discourse. We draw inspiration from previous work on Tree-LSTM over RST structure for document classification, but instead of classifying each node in an unlabeled RST tree based on text features [9, 14], or dictionary-based class scores [4, 19], we use outputs of a pretrained classifier and a type of rhetorical relation as input features of each node to predict the only label for the rhetorical root of the document. A single overall class label is defined for the entire text in the tasks presented in this paper. Hence, we propose a deep model for aggregation of the class labels predicted for all the discourse units in a document by a sequential text classifier. First of all, this allows for a strong sequential text classification method, one that itself takes into account some aspects of discourse [18]. Additionally, the methods based on training high-dimensional EDU representations simultaneously with Tree-LSTMs are found to be prone to strong overfitting [19]. Therefore, it is important to produce DU representations that are as compact and informative as possible, which the proposed method achieves by encoding them with a pre-trained classifier.

The six types of fine-grained relations in the RuRSTreebank corpus outlined in [27] are used in the initial feature representation of each node. These include Coherence (Background, Elaboration, Restatement, Interpretation-evaluation, Preparation), Causal-argumentative:Contrastive (Concession, Contrast, Comparison), Causal-argumentative:Causal (Purpose, Evidence, Cause-effect), Causal-argumentative:Condition (Condition), Structural (Sequence, Joint, Same-unit), and Attribution.

Considering RST tree t and the current nonterminal node (nonelementary discourse unit) $u_i \in t$, its left and right constituents u_{i_1} and u_{i_2} sharing relation $r_i = (r_{i_1}, r_{i_2})$ (e.g., Attribution_NS = (Attribution_Nucleus, Attribution_Satellite)) are initially encoded into representations U_{i_1} and U_{i_2} as follows:

$$U_{i_j} = [\text{FC}_{stance}(\text{Enc}(u_{i_j})); \text{FC}_{premise}(\text{Enc}(u_{i_j})); r_{i_j}] \text{ for } j = 1, 2. \quad (1)$$

An additional Root relation is introduced to encode a root node that is not a constituent. We derive both the BERT-based text encoder Enc and the fully-connected layers for preliminary labels predictions FC from the sequence-level base model with frozen weights. Since all Structural relations are multinuclear and do not have satellites, the one-hot vector r_{i_j} of discourse unit labels (Coherence_Nucleus, Coherence_Satellite, Root, etc.) in our model has a length of 12. Binary Tree LSTM is then applied to these representations $U_k \in t$. The model uses a Tree LSTM hidden representation of the root discourse unit for both stance and argument prediction and has two output feedforward layers as with the BERT model.

⁴<https://huggingface.co/DeepPavlov/rubert-base-cased-conversational>

5 Experiments

5.1 Data

The dataset for joint stance and premise classification is provided by the RuArg-2022 competition organizers. The stance label represents the point of view of the author in relation to the given claim. The presence of arguments for, against, or mixed in the text is indicated by the premise (argument) label. There are three claims in the dataset regarding the COVID-19: “Wearing masks is beneficial for society”, “Vaccination is beneficial for society”, and “The introduction and observance of quarantine is beneficial for society”.

Figure 2 illustrates the length distribution of data by elementary discourse units derived with RST parsing. Since the texts in the dataset do not have paragraph breaks, each text is considered to belong to a single tree. Thus, if the text is l elementary discourse units long, its rhetorical structure contains $l - 1$ relations. Each subset of the data contains about 25% simple sentences with no automatically recognizable discourse structure. From this, we hypothesize that for 75% of the data, the classification performance can be improved by analyzing the coherence structure within the text. Most examples are found to have only one discourse relation between two elementary units; in the official test set, this is the case in 35.9% of examples.

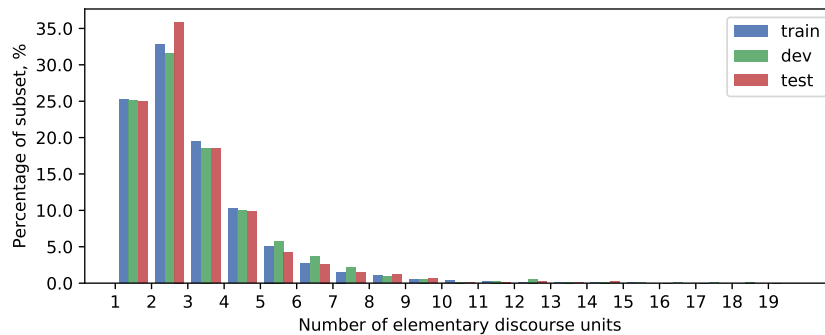


Figure 2: Distribution of text lengths in RuArg-2022

Figure 3 shows the distribution of text lengths in different topic-irrelevant classes across the labeled train set. Complex texts with a rhetorical structure are the most common way in which polar opinions are expressed in the corpus. The simplest sentences are most common among the examples of the mixed class *Other*, with this difference being particularly evident in the premise (argument) classification subtask (Fig. 3b). It demonstrates that most examples in this class lack any argumentation typically [5, 8, 25] expressed by causal, conditional, or any other meaningful discourse relations, which is consistent with the definition of the *Other* in the premise classification subtask description. One interesting observation is that the examples in which the author expresses a positive stance or argument tend to have the most complex structures in the train set.

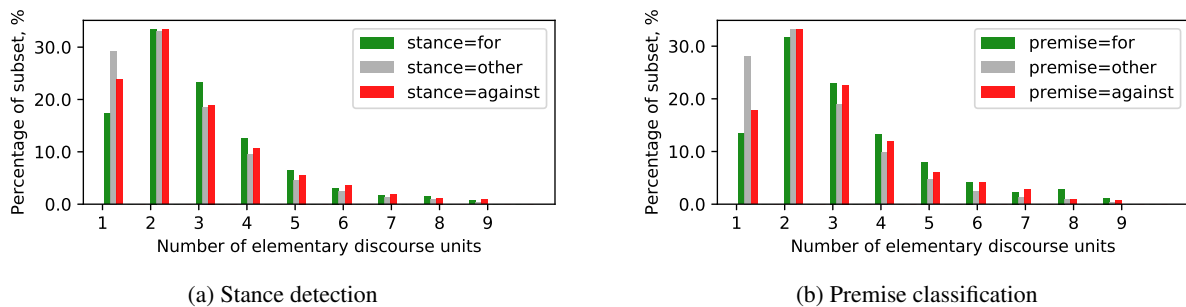


Figure 3: Distribution of text lengths in the train set

5.2 Settings

In the data, the number of examples of various classes is unbalanced, with a major predominance of the *Irrelevant*. We choose to add class weights to the loss in both BERT and RST-LSTM models to prevent unbalanced learning. These weights are adjusted to correspond to the overall class weights in the train data. We use the Optuna optimization framework [1] for automated hyperparameter tuning in both BERT fine-tuning and RST-LSTM training. The optimal hidden size of Tree-LSTM for the three topic-related models is found to be between 50 and 125 units. We use PyTorch and AllenNLP libraries [11] for implementation and a single Nvidia GeForce RTX 2080 Ti GPU. In our experimental setup, RST-LSTM takes on average 2.4 times less time to run one training epoch than BERT, with 2 to 5 epochs total.

5.3 Evaluation Procedure

In our evaluation, we use the metric proposed by RuArg-2022: macro F1 excluding the score for label *Irrelevant*. We use a 5-fold cross-validation over the labeled train set to accurately compare approaches that employ and do not employ rhetorical structure. For the official test and development sets, the final predictions are obtained by averaging predictions from five models trained on cross-validation. This is similar to an ensemble, where each model is trained using 80% of the train data.

6 Results and Discussion

In Table 2 we compare the results of the model with Tree LSTM over RST structure with the baseline BERT model.

Non-EDU classification	Performance on non-EDU				Overall performance			
	Masks	Vaccines	Quar.	Mean	Masks	Vaccines	Quar.	Mean
Stance detection								
BERT	59.8 ± 2.7	62.4 ± 3.4	54.5 ± 3.4	58.9 ± 2.3	60.6 ± 2.6	64.4 ± 2.2	56.4 ± 2.8	60.5 ± 1.9
+ RST-LSTM	61.3 ± 2.7	63.4 ± 4.2	55.6 ± 2.7	60.1 ± 2.3	61.7 ± 2.6	65.1 ± 3.0	57.5 ± 2.4	61.4 ± 1.8
Premise classification								
BERT	66.4 ± 2.9	61.7 ± 4.3	56.4 ± 2.8	61.5 ± 2.2	66.0 ± 2.4	62.6 ± 2.7	57.0 ± 2.3	61.9 ± 1.6
+ RST-LSTM	68.1 ± 2.1	60.4 ± 3.3	57.6 ± 2.0	62.0 ± 1.3	67.5 ± 1.9	61.5 ± 2.3	58.3 ± 2.1	62.4 ± 0.9

Table 2: Performance (F1, mean ± std) during cross-validation

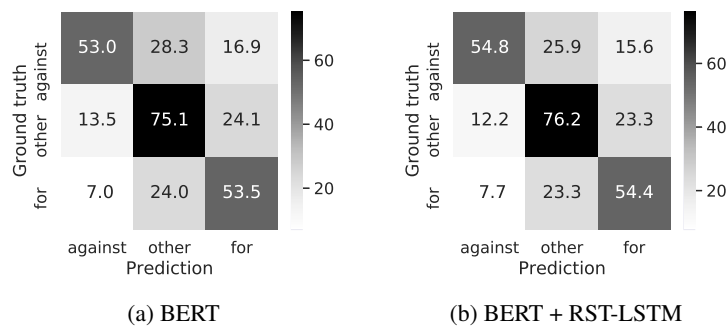


Figure 4: Averaged cross validation confusion matrices for stance detection

Stance detection: Introducing discourse structure to this model leads to an improvement across all topics. The method resulted in an averaged 1.2% mean F1 improvement in the classification of texts with discourse structure and a 0.9% mean F1 improvement for all texts. Figure 4 shows the averaged cross-validation confusion matrices for stance detection in the non-elementary samples of the train set, topics not separated, excluding label *Irrelevant*. The model with RST-LSTM shows improved ability to distinguish between *For* and *Against* polar stance labels and the mixed label *Other*. For the sequential

classifier, *Other* is the most challenging label. First, it can mean that the text has examples of both polar classes; second, it is the most frequent. We include the examples where discourse structure helps to differentiate the stance labels in Figures 5 and 6 in Appendix.

Premise classification: On average, the classification performance improved by 0.5% both for texts with identifiable rhetorical structure and for all texts. Classification performance improved significantly for the topics Masks (+1.6% F1) and Quarantine (+2.0% F1) and worsened for the Vaccines (-1.3% F1), which may indicate a drawback of the approach in which a single structure representation for two targets is trained simultaneously; it is also worth noting that the scores related to the Vaccines topic have the highest deviation in both the sequential and structural classification methods on complex (non-EDU) examples. The examples where discourse structure helps to differentiate the premise labels are illustrated in Figures 7 and 8 in Appendix.

Evaluation on the official dev and test sets: In Table 3 we compare the methods on the official dev and test sets of RuArg-2022. Both sets are treated as unseen, so the official development set was not used for the parameters adjustment. The results confirm that the RST-LSTM is capable of capturing the overall polarity of stance and arguments in a document based on the rhetorical structure. Vaccines-related text classification continues to produce the least stable results. On both dev and test sets, the BERT model enhanced with the RST-LSTM module achieves the best performance for premise classification.

Non-EDU classification	Dev				Test			
	Masks	Vaccines	Quar.	Mean	Masks	Vaccines	Quar.	Mean
BERT	66.0 / 66.1	66.8 / 58.5	58.4 / 58.8	63.7 / 61.1	70.0 / 76.4	68.3 / 63.4	61.0 / 71.6	66.5 / 70.6
+ RST-LSTM	67.3 / 68.2	67.8 / 56.3	56.3 / 59.8	63.8 / 61.4	70.0 / 76.5	66.0 / 63.8	61.7 / 72.5	65.9 / 71.0

Table 3: Performance (F1, stance / argument) on dev and test sets of RuArg-2022

The public RuARG-2022 test leaderboard (Table 4) shows only the results of the last model evaluated. In our case, it is a model different from the one evaluated in Tables 2 and 3. Specifically, it is an additional variant of RST-LSTM where nuclei of asymmetric relations are marked as Span (instead of Attribution_Nucleus, Coherence_Nucleus, etc.). As our method aggregates the rhetorical relations with similar semantics and nuclearity discrepancies, such as the causal relations Purpose⁵ and Cause-Effect⁶, or contrastive Concession⁷ and Contrast⁸, this idea was later dismissed. However, according to Table 3, our final method described in this paper also ranks 4th in stance prediction (65.9%) and 3rd in argument classification (71.0%) in the competition leaderboard on the official test set.

Stance detection			Premise classification		
1	camalibi	69.7	1	camalibi	74.0
2	sevastyannm	68.2	2	sevastyannm	72.4
3	iamdenay	66.8	3	<u>ursdth (ours)</u>	70.6
4	<u>ursdth (ours)</u>	65.7	4	iamdenay	65.6
5	sopilnyak	56.0	5	dr	60.4
6	kazzand	55.5	6	kazzand	56.0
7	morty	53.5	7	morty	54.5
8	invincible	58.9	8	invincible	54.3
9	dr	47.5	9	Baseline	43.6
10	Baseline	41.8			

Table 4: Public leaderboard of RuArg-2022 (F1)

Ablation study: We inspect the importance of the rhetorical relation labels and nuclearities in our method in Table 5. We found that excluding particular relation types individually, i.e. replacing the

⁵Satellite represents the intended *result* behind the situation described in the nucleus.

⁶Nucleus represents the actual *result* after the situation described in the satellite.

⁷Mononuclear relation in which additional information in satellite creates expectations that the situation in the nucleus would be opposite.

⁸Multinuclear relation in which nuclei describe alternative situations.

only type in the trees with a structural type, marginally affects the classification performance on the discourse trees. However, the simultaneous substitution of all semantic relations for a multinuclear relation Structural leads to a 0.2% F1 decrease in the stance identification performance and a 0.6% F1 decrease in argument classification, demonstrating the importance of the features related to the labeled rhetorical structure in argument mining. We note that although the RST parser is far from perfect in recognizing the labeled trees, it is capable of identifying argumentative structures.

Discourse relations	Stance detection	Argument classification
All	60.1	62.0
- Coherence	- 0.1	- 0.1
- Contrastive	- 0.1	- 0.1
- Causal	- 0.0	- 0.1
- Condition	- 0.1	- 0.1
- Attribution	- 0.1	- 0.0
Only structural	- 0.2	- 0.6

Table 5: Ablation study on the rhetorical relations types during cross-validation (F1, mean)

7 Conclusion

Sequential text classifiers typically perform well when applied to short texts, but their performance degrades for longer texts due to the complexity of discourse. In order to form an accurate hidden representation of a complex text, we propose a method leveraging both a pretrained language model and an end-to-end RST parser. Additionally, we improve the rhetorical parsing for Russian using a recent top-down algorithm for paragraph parsing and report fine-grained RST scores for different text granularities. The improved RST parser is used to show the utility of rhetorical parsing in stance detection and premise classification on social media comments.

The architecture we propose shows the effectiveness of a two-step rhetorical-driven approach, where the base text classification method can be any advanced neural network or feature-based machine learning model. Future work should investigate the suitability of discourse parsing in Russian for other tasks requiring argument extraction and processing.

Acknowledgements

This paper is supported by the Research Program of the National Center for Physics and Mathematics (project no. 9).

References

- [1] Akiba Takuya et al. Optuna: A next-generation hyperparameter optimization framework // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. — 2019. — P. 2623–2631.
- [2] Asher Nicholas et al. Appraisal of opinion expressions in discourse // *Linguisticae Investigationes*. — 2009. — Vol. 32, no. 2. — P. 279–292.
- [3] Azar Moshe. Argumentative text as rhetorical structure: An application of rhetorical structure theory // *Argumentation*. — 1999. — Vol. 13, no. 1. — P. 97–114.
- [4] Bhatia Parminder et al. Better Document-level Sentiment Analysis from RST Discourse Parsing // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 2212–2218.
- [5] Chakrabarty Tuhin et al. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — 2019. — P. 2933–2943.

- [6] Chernyavskiy Alexander and Ilvovsky Dmitry. Recursive Neural Text Classification Using Discourse Tree Structure for Argumentation Mining and Sentiment Analysis Tasks // International Symposium on Methodologies for Intelligent Systems / Springer. — 2020. — P. 90–101.
- [7] Chistova Elena et al. RST Discourse Parser for Russian: An Experimental Study of Deep Learning Models // In Proceedings of Analysis of Images, Social Networks and Texts (AIST). — 2020. — P. 105–119.
- [8] Eckle-Kohler Judith et al. On the role of discourse markers for discriminating claims and premises in argumentative discourse // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 2236–2242.
- [9] Fu Xianghua et al. Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis // Asian conference on machine learning / PMLR. — 2016. — P. 17–32.
- [10] Galitsky Boris et al. Argumentation in text: discourse structure matters // CICLing 2018. — 2018.
- [11] Gardner Matt et al. AllenNLP: A Deep Semantic Natural Language Processing Platform. — 2017. — arXiv:1803.07640.
- [12] Goller Christoph and Kuchler Andreas. Learning task-dependent distributed representations by backpropagation through structure // Proceedings of International Conference on Neural Networks (ICNN'96) / IEEE. — 1996. — Vol. 1. — P. 347–352.
- [13] Hogenboom Alexander et al. Using rhetorical structure in sentiment analysis // Communications of the ACM. — 2015. — Vol. 58, no. 7. — P. 69–77.
- [14] Huber Patrick and Carenini Giuseppe. From Sentiment Annotations to Sentiment Prediction through Discourse Augmentation // Proceedings of the 28th International Conference on Computational Linguistics. — 2020. — P. 185–197.
- [15] Ji Yangfeng and Eisenstein Jacob. Representation learning for text-level discourse parsing // Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). — 2014. — P. 13–24.
- [16] Ji Yangfeng and Smith Noah A. Neural Discourse Structure for Text Categorization // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2017. — P. 996–1005.
- [17] Kotelnikov Evgeny et al. RuArg-2022: Argument Mining Evaluation // Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”. — 2022.
- [18] Koto Fajri et al. Discourse Probing of Pretrained Language Models // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2021. — P. 3849–3864.
- [19] Kraus Mathias and Feuerriegel Stefan. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees // Expert Systems with Applications. — 2019. — Vol. 118. — P. 65–79.
- [20] Lee Kangwook et al. A discourse-aware neural network-based text model for document-level text classification // Journal of Information Science. — 2018. — Vol. 44, no. 6. — P. 715–735.
- [21] Lin Xiang et al. A Unified Linear-Time Framework for Sentence-Level Discourse Parsing // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — P. 4190–4200.
- [22] Mann William and Thompson Sandra. Rhetorical structure theory: A theory of text organization. — University of Southern California, Information Sciences Institute Los Angeles, 1987.
- [23] Morey Mathieu et al. How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT // Conference on Empirical Methods on Natural Language Processing (EMNLP 2017). — 2017. — P. pp–1330.
- [24] Mukherjee Subhabrata and Bhattacharyya Pushpak. Sentiment analysis in twitter with lightweight discourse analysis // Proceedings of COLING 2012. — 2012. — P. 1847–1864.

- [25] Musi Elena et al. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018.
- [26] Nguyen Thanh-Tung et al. RST Parsing from Scratch // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2021. — P. 1613–1625.
- [27] Pisarevskaya Dina et al. Towards building a discourse-annotated corpus of Russian // Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue. — 2017. — P. 23.
- [28] Prendinger Helmut et al. A novel discourse parser based on support vector machine classification // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. — 2009. — P. 665–673.
- [29] Somasundaran Swapna and Wiebe Janyce. Recognizing stances in online debates // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. — 2009. — P. 226–234.
- [30] Soricut Radu and Marcu Daniel. Sentence level discourse parsing using syntactic and lexical information // Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. — 2003. — P. 228–235.
- [31] Tai Kai Sheng et al. 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 (Volume 1: Long Papers). — 2015. — P. 1556–1566.
- [32] Voll Kimberly and Taboada Maite. Not all words are created equal: Extracting semantic orientation as a function of adjective relevance // Australasian Joint Conference on Artificial Intelligence / Springer. — 2007. — P. 337–346.
- [33] Zhang Longyin et al. A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — P. 6386–6395.
- [34] Zirn Căcilia et al. Fine-grained sentiment analysis with structural features // Proceedings of 5th International Joint Conference on Natural Language Processing. — 2011. — P. 336–344.

Appendix

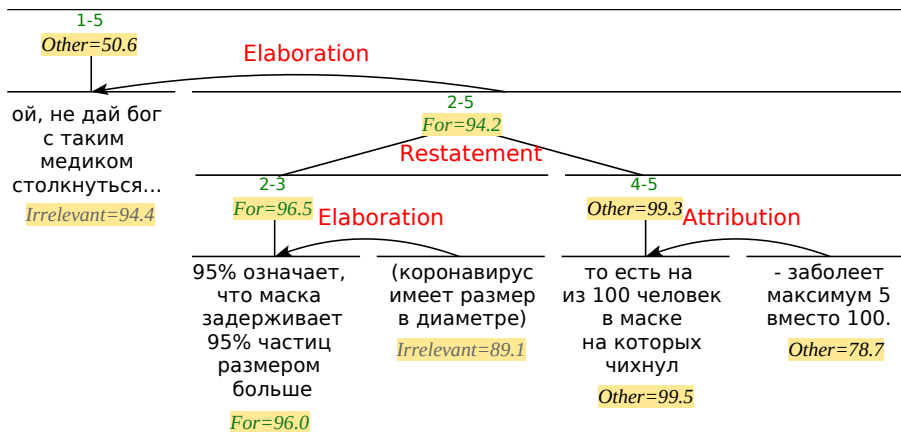


Figure 5: Example of RST parsing for a claim regarding masks from RuArg-2022: [oh, god forbid anyone encountering such a physician...]₁ [95% means that the mask blocks 95% of particles larger]₂ [(coronavirus has a diameter dimension)]₃ [that is, out of 100 people wearing masks who were sneezed on]₄ [- at most 5 will get sick instead of 100.]₅. For this example, the BERT-based classifier predicts a false stance label *Other*, but RST-LSTM predicts a true stance label *For*.

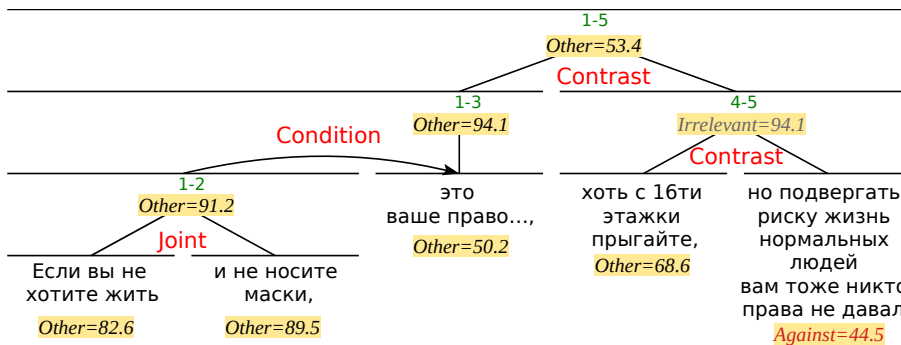


Figure 6: Example of RST parsing for a claim regarding masks from RuArg-2022: [If you don't want to stay alive]₁ [and are not wearing masks,]₂ [that's your right...],₃ [you are free to jump off a 16-story building as well,]₄ [but no one authorized you to endanger normal people's lives!]₅. Yellow boxes indicate BERT stance label predictions and their probabilities. For this example, the BERT-based classifier predicts a false stance label *Other*, but RST-LSTM predicts a true stance label *For*.

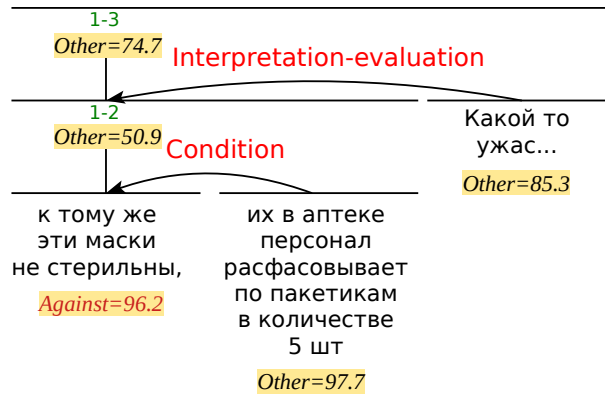


Figure 7: Example of RST parsing for a claim regarding masks from RuArg-2022: [*in addition, these masks are not sterile,*]₁ [*the pharmacy staff packs them in bags of five pieces*]₂ [*What a nightmare...*]₃. Yellow boxes indicate BERT premise label predictions and their probabilities. For this example, the BERT-based classifier predicts a false premise label *Other*, but RST-LSTM predicts a true premise label *Against*.

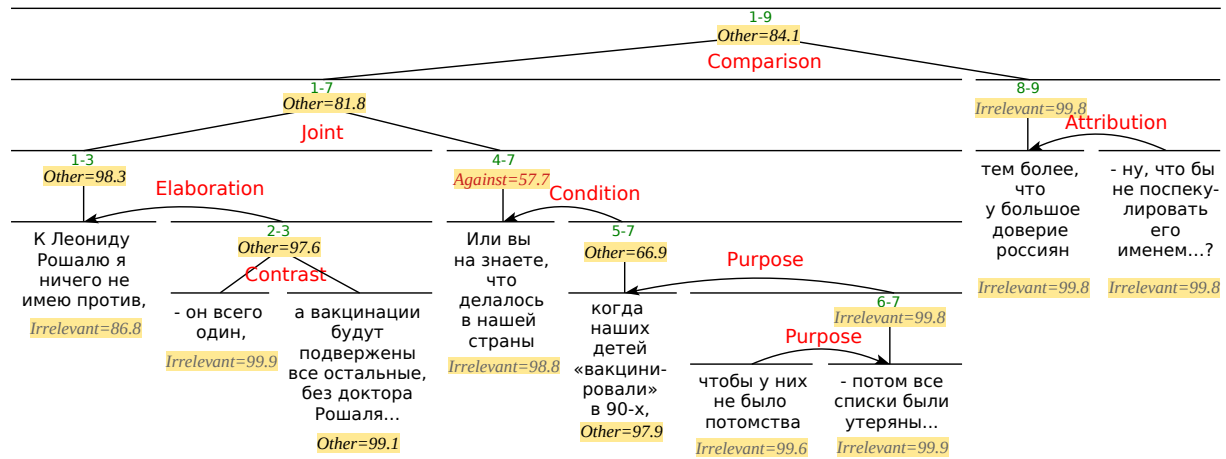


Figure 8: Example of RST parsing for a claim regarding vaccines from RuArg-2022: [*I have nothing against Leonid Roshal*]₁ [*- he is just one person,*]₂ [*and everyone else will be vaccinated, without Dr. Roshal...*]₃ [*Or don't you remember what happened in our country*]₄ [*when they “vaccinated” our children back in the 90s*]₅ [*so that they wouldn't have children*]₆ [*- then all the records went missing...*]₇ [*especially given how much trust Russians have for him*]₈ [*well, why not speculate on his name...?*]₉. Yellow boxes indicate BERT premise label predictions and their probabilities. For this example, the BERT-based classifier predicts a false premise label *Other*, but RST-LSTM predicts a true premise label *Against*.

Russian Verbal Affixation in Mental Lexicon: Priming Study and Its Online Replication with True and Stem-Modified Relative Prime Verbs

Chuprina A. O.
NRU HSE, Moscow
a.o.chuprina@gmail.com

Abstract

While suffixed and prefixed words share common lexical features with their base word in the mental lexicon, the two derivational processes have their own properties. Whether their differences are reflected in the mental storage of the group of relative words or not is one of the topical psycholinguistic questions. My experimental results indicate that memory representations of the derivatives differ: while between the stem and the suffixed relative, the relationship is closer and based on transparency of the derived meaning, the relationship between the stem and the prefixed derivative is rather formal. The results also signal that the decompositional route is not a preferred strategy in complex verb processing for a Russian speaker. I base this conclusion on the results of two in-person experiments and their online replicas. Additionally, the data suggest that lexical organization is modified through the aspectual information of family members. These findings need to be taken into account in future studies, both within psycholinguistic and computational fields, using verbal material of the Russian language.

Keywords: Russian verbal affixation; lexical access; mental lexicon; speed-accuracy trade-off; online experimentation

DOI: 10.28995/2075-7182-2022-21-106-113

1 Background

My study investigates how morphologically related verbs are accessed in the mental lexicon while reading a word without a context. Beyond the theoretical objective, the research deals with the problem of result replication. For this, I use the power analysis on the available data in order to plan an online replication study to support the obtained major effects and critically evaluate the emergent trends. This paper concludes with a diagnostic comment on the data collected through online platforms.

Studies of mental lexicon deal with the question of how various types of information carried by a word affect the storage and access to a word in memory both in comprehension and production. The majority of morphological processing studies provide evidence for some effects of universal importance among languages like a facilitatory effect of word frequency, inhibitory effect of word length or age of acquisition ([9] for a review). However, language specific effects are far from being substantially accounted for. Moreover, once they are, the experimental findings fail to be uniformly explained, especially within the theory of decomposition and its mechanisms of affix stripping [22] and edge-aligned embedded word activation [13], [18]. This not only revives alternative theories of whole-word storage [12] and a dual-route model [2], [7] but invites a revision of results through the lens of the morphological phenomenon under investigation. Theoretical conclusions were primarily drawn upon a mixture of grammatical classes, while later experimental data confirmed that the information incorporated into the memory varies depending on the part of speech of a word. Particularly, in Hebrew a priming effect was observed for verbs but not nouns in the condition when both prime and target shared a stem [11]. Similar discrepancies in lexical storage were found for Russian verbs but not deverbal nouns [25] and Bantu nouns but not verbs [4].

My research explores the storage properties of base verbs in relation to their prefixed and suffixed derivatives in Russian. As the base word undergoes various changes, accompanying either of the processes, these differences may play a role for lexical access to the base word. I study this, using an in-person or offline traditional mode and an online platform for the replication study.

2 Material of Russian Verbs

Suffixation and prefixation differ in formal and semantic changes they bring onto the base [14]. A suffixed verb does not share the inflectional class of the base verb, which results in its distinct inflectional paradigm; and its stress may shift from its positional morpheme in the base. The base verb undergoes phonological transformations (vowel/consonant truncation, insertion, and alternation) for the morpheme boundary of the neighboring stem and suffix to comply with the phonological rules of Russian. However, the meaning of the related word in suffixal transformations of two most common types that I use in this study is uniformly predictable: the suffix *-nu-* signals a single act of the base verb-denoted action, and the suffix *-va-* indicates the process of imperfectivization.

The prefix, on the other hand, does not imply as many changes of the base verb: the two relative verbs share the verbal class and most often the stress pattern. On the contrary, the related verb is not always semantically transparent outside the context (*затолкать* ‘push/push into’ in the meaning of ‘pushing’, ‘forcefully hitting somebody in a crowd’ or ‘starting to push’). In addition, the aspect of the base verb changes only when an imperfective base verb undergoes the prefixation, but not the perfective one.

Additionally, an exploratory corpus investigation showed a difference in use of these two derived forms. The dominant form of the prefixed verb surfaced as that of the participle (1:2, for each finite form there are two non-finite forms of a prefixed verb). On the contrary, a suffixed verb primarily appears as a fully inflected form, which is comparable to the syntactic behaviour of the base verb (1:3). Overall, this suggests a functional difference between the two relative verbs.

The study of verbs and their derivatives in Russian unavoidably includes aspectual information into the research frame. Prior studies of verbs [23] and theoretical discussion of the aspectual pairs ([27] among many others) suggests a closer mental organization between a verb and its imperfective *-va-* derivative, while the connection between the base and its prefixed relative verb might depend on a particular prefix or overall might be disconnected from the base, acquiring its own semantics with a distant relation to the base [16]. In the first case, the imperfective verbs will provide a processing advantage to their bases, resulting in shorter reaction times upon reading the base. In the second case, prefixed verbs will give very little facilitation to their base during lexical access and show a longer decision time to the base.

To sum up, using a group of verbs of Russian I compare the properties of lexical access to them after a brief exposure to their related affixed verbs. The main prediction is that the faster the access is to the base, the closer the storage between the base and its derivative. If the suffixed derivative is closer, the mental connection between the relative verbs is based on the clearly related semantics between the two, easily overriding formal changes of the derivational process. If the prefixed derivative is closer, the mental organization depends more on the formal resemblance between the two verbs. The continuation of the study, using the copies of the derivatives as primes but with an orthographically transformed stem, will test the presence of the decompositional strategies that can be used during access to the base verb. In particular, the affix stripping and edge-aligned embedded word activation operations will put a connection between the base and its prefixed derivative into an advantageous position during access. While none of these particular operations could account for the type of the suffixal connection in Russian, if the access is still faster from the quasisuffixed verb, it will point at a somewhat two-stage decomposition that strips a verbal suffix *-t* and spots *-nu* suffix simultaneously. If decomposition is not a preferred strategy for accessing a Russian complex verb, I cannot expect any facilitation from neither of the relative verbs.

3 Data and Procedure

I use the experimental procedure of the lexical decision task with masked morphological priming [10], see the Data and Procedure section for the details. This experimental procedure does not demand a substantial expenditure of the resources on the part of a researcher and has validated itself on many decades of research on different languages. It is not surprising that online data collection platforms now allow researchers to use the available code or create it from scratch and run their priming experiments [1]. I ran two offline experiments and their two online replicas. In Experiment 1 ‘Real Affixed Verbs’ primes were suffixed and prefixed derivatives of the base verbs as targets. Experiment 2 ‘Quasistem Verbs’ used the same target verbs, but the primes were nonwords obtained by replacing a letter within a stem and an intact affix of the real verb used in Experiment 1 (see Table 1).

Stimulus	Suffixed Prime	Prefixed Prime	Unrelated prime
KÁPAT' (Imperf.) 'to drop'	kápnut' (Perf.) *kabnut' (Perf.)	zakápat' (Perf.) *zakabat' (Perf.)	krjákat' (Imperf.) *krjagat' (Imperf.)
ZAVÍT' (Perf.) 'to curl sth'	zavivát' (Imperf.) *zarivat' (Imperf.)	podzavít' (Perf.) *podzarit' (Perf.)	perepít' (Perf.) *porepít' (Perf.)

Table 1: An example of stimuli-prime pairs in Experiment 1 'Real Affixed Primes' (the top prime) and in Experiment 2 'Quasistem Primes' (the bottom prime)

For the target material I selected 39 target base verbs. The length of the stimulus words was five to nine letters, their own frequency was in the range from zero to 70 ipm (occurrences of a word per million in the main corpus of the Russian National Corpus). The frequency was taken from the Frequency Dictionary of the Modern Russian Language, edited by O. N. Lyashevskaya and S. A. Sharov [20]. A group of unrelated verbs was chosen to control the results for related prime conditions. Afterwards, an additional group of 13 fillers was added to equal the lists in related and unrelated conditions overall. To these 52 targets, I added the same number of nonword targets, preceded by real word primes in Experiment 1 and nonword primes in Experiment 2. It should be noted that nonword targets in Experiment 2 appeared in the related and unrelated conditions, i.e. preceded by relative and unrelated nonword primes [6]. All nonword primes were constructed, following a natural phonetic structure of the Russian language.

The experiment consisted of performing a lexical decision task with visual masked priming. Each experimental attempt by a subject consisted of seeing a screen with an asterisk, serving as a fixation point (varied between 1000ms and 1500ms); a 500ms mask, the length of which corresponded to the average length of the target words (seven bars, #); a 60 ms prime and a following 20ms mask. The function of grids or masks is to avoid the visual trail of words appearing on the screen. Finally, a 500ms target word screen appeared between two blank screens of 30ms and 2000ms correspondingly. To respond, a subject pressed the key – either 'left Ctrl' or ARROW RIGHT, which were located in the bottom row of the keyboard. The participant's task was to decide as quickly as possible whether they see a real word, or whether the word does not exist by pressing the corresponding key. The speed and correctness of the answers were measured.

The experiment was carried out on a DELL personal computer using the OpenSesame program [21]. The overall duration of the procedure amounted to seven minutes.

The sample consists of 39 participants in Experiment 1 and in Experiment 2 of 25 participants. In online replication studies the samples consist of 56 and 41 in Experiment 1 and 2 correspondingly.

All speakers speak Russian as their native language. The age range of the entire sample is between 18 and 59 years of age with female participants constituting two thirds of the whole sample. All participants completed the task in full and were not aware of the objectives of the study except for the general psycholinguistic research into the process of reading.

4 Results

I analyzed the reaction of the subjects according to two parameters: the accuracy of the choice "word or nonword" and the speed of the response. Based on the general distribution of the respondents' answers, I determined the threshold of 80% accuracy, below which it could be uncertain that the participant in the experiment gave thoughtful answers. As a result, the data of nine participants in Experiment 1 and four participants in Experiment 2 were not included in the statistical analysis. To describe the data, I used the generalized additive model as in package *mgcv* [26] in the R programming environment for statistical data processing [24].

I now present the data of both experiments with real (Experiment 1) and quasistem primes (Experiment 2) for immediate comparison of the effects.

In Experiment 1 the response accuracy showed a subject random effect ($\text{Chi.sq}=132.1$, $p<0.01$) as well as main effects of target frequency ($\text{Est.}=0.7$, $\text{SE}=0.09$, $p<0.01$) and target aspect ($\text{Est.}=-1.02$, $\text{SE}=0.23$, $p<0.01$), imperfective group having an advantage over the perfective one. An effect of prime type was not observed. In Experiment 2 the analysis showed the significant random effects of the subject ($\text{Chi.sq}=28.1$, $p<0.01$) and target ($\text{Chi.sq}=31.6$, $p<0.01$) as well as main effects of target's frequency ($\text{Est.}=0.4$, $\text{SE}=0.15$, $p=0.002$), a positive effect of length of the target ($\text{Est.}=0.86$, $\text{SE}=0.33$, $p=0.01$) and the negative one of the prime ($\text{Est.}=-0.7$, $\text{SE}=0.3$, $p=0.01$). From the group of related verbs, the prefixed prime showed a positive effect on accuracy in comparison to unrelated controls ($\text{Est.}=0.98$, $\text{SE}=0.4$, $p=0.02$).

Before analysing the reaction time data, I excluded the data points for incorrect answers and extreme values greater than 2500ms, following an established data curation procedure [4], [5], [19]. To this, I lost 10% of each experiment's data sample. Also, a quadratic transformation of the dependent variable was performed, guided by the boxcox plot suggestion [3].

The following analysis revealed the random effect of the subject ($F=24.16$, $p<0.01$) in Experiment 1 and subject ($F=20.69$, $p<0.01$) and the target verb factors ($F=0.49$, $p=0.05$) in Experiment 2. Beyond main effects of frequency ($\text{Est.}=-4.8-05$, $\text{SE}=9.1-06$, $p<0.01$) and aspect, perfective targets recognized more slowly ($\text{Est.}=1.4-03$, $\text{SE}=3.4-04$, $p<0.01$), in Experiment 1, the morphological type of the prime verb also showed an explanatory power not only in comparison to the unrelated verb condition (prefix facilitation: $\text{Est.}=-1.02-03$, $\text{SE}=3.2-04$, $p<0.01$; suffix facilitation: $\text{Est.}=-1.8-03$, $\text{SE}=3.2-04$, $p<0.01$) but also among themselves, a suffixed verb having an advantage over a prefixed derivative ($\text{Est.}=-7.2-04$, $\text{SE}=3.3-04$, $p=0.03$). As for Experiment 2, the effects of frequency ($\text{Est.}=-8.5-04$, $\text{SE}=1.5-04$, $p<0.01$) and aspect ($\text{Est.}=1.5-03$, $\text{SE}=3.9-04$, $p<0.01$) were the only ones to account for the variance in the data sample.

There are various methods to calculate the plausibility of the null hypothesis, depending on the statistical model and the type of data. For example, power analysis, using simulations, can be applied for generalised linear effect models, while Bayesian inference – for more nuanced complex modelling. To fit into the bigger picture of morphological processing research, where the data are mainly analysed with linear models, I chose the power analysis with simulations. The effect of interest after running the offline study was the suffix' advantage over prefix'. The power calculation for the model to detect the suffix effect of this model was 60% which is good but not sufficient for the certainty in theoretical conclusions. With the help of *extend* method from the *simr* package [15] I identified the necessary subject sample size of 40 to detect the effect with about 80% confidence.

To program the replication of Experiment 1 and 2, I adjusted the masked priming experiment template available at PCIBex, an online experiment builder [28], and to recruit the participants I used Toloka, a Yandex crowdsourcing platform [29]. Due to technical differences between programming procedures for the original study and its replication, some issues arised. One of them is that the duration of the fixation point was always 1500ms in comparison to its varied 1100–1500ms timing in the offline experiments. To see whether this factor should be corrected for, I included it into the initial offline model and saw that it does not affect the data variation. The other issue was that the online response could only be given upon the disappearance of the target off the screen, i.e. there was no chance for a participant to react with a key press within the target duration of 500ms. This could have been a problem if the RTs' distribution and its central tendency measures suggested an anomaly in the online data sample in comparison to the offline data (offline RTs: mean (695.2ms), sd (237.1), median (643); online RTs: mean (920.2ms), sd (364), median (835)). This was not the case. The online RTs data were overall slower than in the offline sample but both samples were distributed identically.

As before the data for subjects that gave too many false responses or did not finish an experiment were extracted from the further analysis, making a 45-subject sample in Experiment 1 and a 34-subject sample in Experiment 2. In online Experiment 1 the response accuracy revealed the subject random effect ($\text{Chi.sq}=207.8$, $p<0.01$) as well as main effect of frequency ($\text{Est.}=0.29$, $\text{SE}=0.08$, $p<0.01$) without the aspect effect as it was observed in the offline version. In Experiment 2 the significant random effect was the one of the subject ($\text{Chi.sq}=55.02$, $p<0.01$). The main effects in online Experiment 2 were of target verb frequency ($\text{Est.}=0.57$, $\text{SE}=0.1$, $p<0.01$), length of the target ($\text{Est.}=0.59$, $\text{SE}=0.24$, $p=0.01$) and of the prime ($\text{Est.}=-0.44$, $\text{SE}=0.22$, $p=0.05$) in the same direction as in the offline mode. As before, the prefixed prime showed a significant positive effect in comparison to unrelated controls ($\text{Est.}=0.57$,

SE=0.31, $p=0.07$) but this time, including the contrast to the suffixed prime, suffixed primes causing more mistakes (Est.=-0.8, SE=0.3, $p=0.02$).

After cutting incorrect responses and extremely high values (14% of the data sample in online Experiment 1 and 13% in online Experiment 2) I analysed the RT data. The analysis revealed the random effect of the subject ($F=20.64$, $p<0.01$) and target ($F=0.8$, $p<0.01$) in Experiment 1. The same factors were observed in Experiment 2. Among the main effects in Experiment 1, there emerged an effect of frequency (Est.=-0.0005, SE=0.0001, $p<0.01$) and both relative primes in comparison to unrelated controls but not among themselves as it was in the offline sample (prefix facilitation: Est.=-0.001, SE=0.0003, $p<0.01$; suffix facilitation: Est.=-0.0008, SE=0.0003, $p=0.003$). As for Experiment 2, the effect of frequency (Est.=-0.0005, SE=0.0002, $p<0.01$) was the only robustly significant effect while the others including the related primes emerged at the level of tendency.

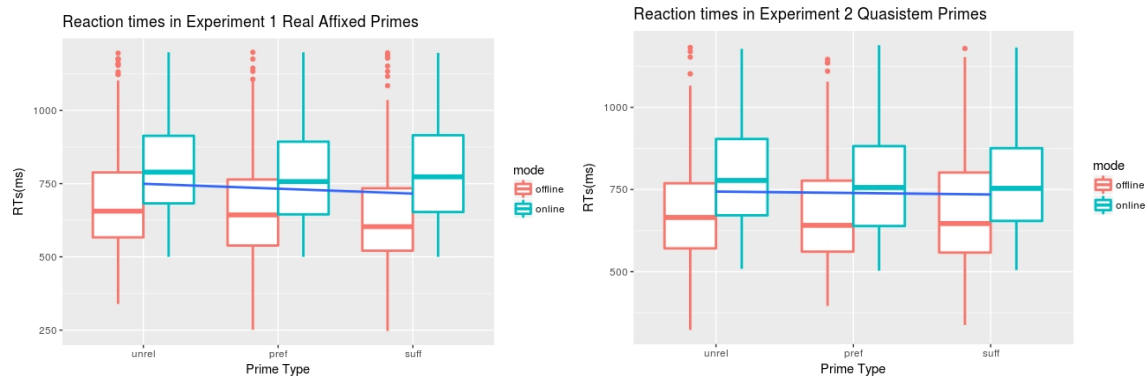


Figure 1: Boxplots for reaction time data based on linear models for Experiments 1 and 2

To evaluate the main finding of suffix' greater facilitation over prefixes in the offline mode for the real verbs' pair and the effects at tendency levels in the online mode after the quasistem primes, I combined offline and online datasets into one and included the mode of the experiment as an independent variable. In Experiment 1 with real primes the effect of the mode of the experiment was not observed in the accuracy data. However, the speed of online responses was significantly slower than the offline ones (Est.=4.6-03, SE=6.8-04, $p<0.01$). The effects of frequency and target aspect were preserved as well as the ones of prime type, suffix condition being faster when compared to the unrelated primes (Est.=-1.8-03, SE=3.1-04, $p<0.01$) and their prefixed counterparts (Est.=-7.1-04, SE=3.2-04, $p=0.03$). In addition, the interaction between the mode and prime type showed a significant slow down for suffixed primes online (in the comparison to the unrelated controls: Est.=9.8-04, SE=4.1-04, $p=0.02$; and to the prefixed primes: Est.=9.5-04, SE=4.1-04, $p=0.02$). However, calculating the difference between the two slopes and accounting for the standard errors, the suffix effect was still facilitatory in contrast to both the unrelated and the prefixed primes.

The combined data analysis for accuracy in Experiment 2 did not show data variation for the effect of the experimental mode and the previously observed effects of frequency and lengths of target and prime were robust as well as the beneficial effect of prefixes over the unrelated condition (Est.=0.9, SE=0.4, $p=0.02$). The reaction times data also showed a significant slow down in the online sample (Est.=3.6-03, SE=7.6-04, $p<0.01$) as well as the main effects of frequency (Est.=-5.4-05, SE=1.4-05, $p<0.01$) and target aspect (Est.=1.7-03, SE=5.1-04, $p=0.01$). An interaction surfaced between the mode and the prime time which indicated a greater facilitation from prefixed primes over unrelated ones online in comparison to the identical case in the offline experiment (Est.=-9.1-04, SE=4.5-04, $p=0.3$).

Finally, the error analysis of data on nonword prime-target pairs in the offline Experiment 2 showed the significance of the same random factors as in the analysis of the material of real words. The morphological type of the verb was not found to be significant in explaining either the variability in the correctness data or the reaction times.

5 Discussion

Morphologically related words are stored especially closely in the memory of speakers of different languages. My study assessed whether for Russian speakers a particular type of the morphological relation is reflected in the mental lexicon. My study on derived suffixed and prefixed verbs shows that there are psychological differences in the organization of storage of and access to the base verb, channeled through either of the derivatives, a result similar to the findings in Korean [19].

The robust observed effects showed an importance of the stem morpheme for the transference of lexical information from an affixed word to its base: once the lexical relatedness was cancelled by a single-letter modification within the stem of the derivative, the advantage from the related primes over unrelated primes was cancelled as well with no regard to the intact affixal morpheme. Instead, we could observe purely orthographic effects, ones of target's and prime's lengths, to emerge for accuracy of response. The fact that these effects were accompanied by the prefix effect in the quasi-affixed prime data could also point to the orthographic nature of the base-prefixed derivative connection. An approximation of the decompositional route that a Russian verb may go through is an affix stripping or more specifically, a prefix-stripping mechanism. However, the cognitive nature of this mechanism deserves a closer look as the prefixal advantage was observed only for accuracy measure.

The speed of response results show that suffixed derivatives are better facilitators of the base verb's recognition. The presence of a real suffix on the root that deviated from the shared stem by a single letter did not give a processing advantage over an incompletely unrelated letter string. Together it shows that the psychological representation of the connection between the base and its suffixed derivative is localized inside the shared root of the pair. This, in contrast to the prefix's orthographic relation, indicates the lexical nature of the suffix relation to its base.

The emergence of the facilitatory suffix effect for the reaction time in the experiment with real primes and of the beneficial prefix effect in the accuracy results in the experiment with nonword primes suggest that multiple memory processes mediate morphological connections of a base word with its derivatives. This is in line with the studies on speed-accuracy trade-off [17]. This also brings about the next observation about the verb's aspect factor, which emerged in the offline data samples and was more robust for the real verb pairs. I cannot extrapolate this to a single verb recognition due to the methodological two-word nature of the study, but my results signal more common information stored in memory between the imperfective verb and its perfective derivative than inside the pair of the reverse relationship. This result along with a previously found aspect effect in sign language [8], deserves further investigation for the benefit of both theoretical and cognitive linguistic fields.

As for the available modes for experimentation, the researcher should be aware of the discrepancies that might arise. Although the online mode is an undeniably valid tool for the data collection and a valuable platform for replication studies, one should take care of the smaller volatile effects. Due to the longer speed of response in the offline mode, one can possibly detect another processing procedure than the one hypothesized. Specific to my case, this is the effect of the suffix in relation to prefix, showing discontinuity in the online mode of experimentation. Since the extra time, taken online before an actual button press, allows for more prime-target information becoming available, I can tentatively conclude that the lexical properties being incorporated at later time points might also differ for either of the morphological processes.

Altogether, my study not only confirms that mental lexicon reflects various means of organization of words but also is one of the few that demonstrates that even within a homogeneous group of a single category of verbs one can expect a substantial variation in the mental organization.

Acknowledgements

I am deeply grateful to my PhD supervisor professor Nataliia Slioussar at HSE (Moscow, Russia) for introducing me to the research of complex word storage in the mental lexicon, for suggesting the research method, and for editing the initial verb lists.

References

- [1] Angele B., Baciero A., Gómez P., Perea M. Does online masked priming pass the test? The effects of prime exposure duration on masked identity priming. — *Behavior Research Methods*, 2022.
- [2] Baayen R. H., Schreuder R. War and peace: Morphemes and full forms in a non-interactive activation parallel dual route model. — *Brain and Language*, 1999. — 68.
- [3] Box G., Cox D. An Analysis of Transformations. — *Journal of the Royal Statistical Society. Series B (Methodological)*, 1964. — 26(2).
- [4] Ciaccio L. A., Kgoro N., Clahsen H. Morphological decomposition in Bantu: a masked priming study on Setswana prefixation. — *Language, Cognition and Neuroscience*, 2020. — 35:10.
- [5] Chuprina A. (2019), The problem of related verbs in mental lexicon [K probleme rodstvennyh glagolov v mental'nom leksikone], *St. Tikhon's University Review. Series III: Philology [Vestnik Pravoslavnogo Sviato-Tikhonovskogo gumanitarnogo universiteta Serii III: Filologiya]*, Vol. 59, pp. 36-53.
- [6] Chuprina A. (2022), Affixation in the mental lexicon: morphological priming in Russian verbs with orthographic changes [Processy affiksacii v mental'nom leksikone: morfologicheskij prajming pri orfograficheskikh izmenenijah v rodstvennyh glagolah russkogo jazyka], *St. Tikhon's University Review. Series III: Philology [Vestnik Pravoslavnogo Sviato-Tikhonovskogo gumanitarnogo universiteta Serii III: Filologiya]*, Vol. 70, pp. 63-91.
- [7] Diependaele K., Morris J., Serota R. M., Bertrand D., Grainger J. Breaking boundaries: Letter transpositions and morphological processing. — *Language and Cognitive Processes*, 2013. — Vol. 28(7).
- [8] Emmorey K. *Language, cognition and the brain: insights from sign language research*. Psychology Press, 2001.
- [9] Feldman L.B. *Morphological aspects of language processing*. Psychology Press, 2013.
- [10] Forster K.I., Davis, C. Repetition priming and frequency attenuation in lexical access. — *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1984. — 10.
- [11] Frost R., Kugler T., Deutsch A., Forster K. Orthographic Structure Versus Morphological Structure: Principles of Lexical Organization in a Given Language. — *Journal of experimental psychology. Learning, memory, and cognition*, 2005. — Vol. 31.
- [12] Giraud H., Grainger J. Priming complex words: Evidence for supralexical representation of morphology. — *Psychonomic Bulletin & Review*, 2001 — 8.
- [13] Grainger J., Beyersmann E. Edge-aligned embedded word activation initiates morpho-orthographic segmentation. In B. H. Ross (Ed.), *Psychology of learning and motivation*, 2017. — Vol. 67.
- [14] *Grammar of Russian [Russkaya grammatika]. Phonetics. Phonology. Stress. Intonation. Derivation. Morphology. [T. 1: Fonetika. Fonologiya. Udarenie. Intonaciya. Slovoobrazovanie. Morfologiya.]* Editors: N. Yu. Shvedova (chief editor), N.D.Arutyunova, A.V.Bondarko, Val.Vas.Ivanov, V.V. Lopatin, I.S.Uluxanov, F.P.Filin. M.: Nauka, 1980.
- [15] Green P, MacLeod CJ. “simr: an R package for power analysis of generalised linear mixed models by simulation.” — *Methods in Ecology and Evolution*, 2016. — 7(4).
- [16] Janda L. A., Endresen A., Kuznetsova J., Lyashevskaya O., Makarova A., Nessel T., Sokolova S. *Why Russian aspectual prefixes aren't empty. Prefixes as verb classifiers*. Bloomington: Slavica Publishers, 2013.
- [17] Kahana M., Loftus G. Response Time versus Accuracy in Human Memory. Sternberg, R. *The Nature of Cognition*. In *Cognitive Linguistics Bibliography (CogBib)*. Berlin, Boston: De Gruyter Mouton, 2010.
- [18] Kazanina N., Dukova-Zheleva G., Geber D., Kharlamov V., Tonciulescu K. Decomposition into multiple morphemes during lexical access: a masked priming study of Russian nouns. — *Language and Cognitive Processes*, 2008. — Vol.23.
- [19] Kim S. Y., Wang M., Taft M. *Morphological Decomposition in the Recognition of Prefixed and Suffixed Words: Evidence From Korean*. — *Scientific Studies of Reading*, 2015.
- [20] Lyashevskaya O. N., Sharov S. A. (2009), *Frequency dictionary of modern Russian language (based on National corpus of Russian) [Chastotnyj slovar' sovremennogo russkogo yazyka (na materialah Nacional'nogo korpusa russkogo yazyka)]*, Moscow; Azbukhovnikh.

- [21] Mathôt S., Schreij D., Theeuwes J. OpenSesame: An open-source, graphical experiment builder for the social sciences”. — *Behavior Research Methods*, 2012. — 44(2).
- [22] Rastle K., Davis M. H. Morphological decomposition based on the analysis of orthography. — *Language and Cognitive Processes*, 2008. — 23(7–8).
- [23] Riekhakainen E. I. Verbal aspectual pairs and associative connections in the mental lexicon of the Russian native speaker // XLII Mezhdunarodnaya filologicheskaya konferentsiya. Izbrannye trudy. Bogdanov S. I., Men'shikova Yu. V. (eds.) St. Petersburg State Univ.: Philology Department, 2014. — P. 265—274.
- [24] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>, 2020.
- [25] Slioussar N., Chuprina A. How derivational links affect lexical access evidence from Russian verbs and nouns. — *Italian journal of linguistics*, 2016. — Vol. 28.
- [26] Wood S. *Generalized Additive Models: An Introduction with R*, 2 edition. — Chapman and Hall/CRC, 2017.
- [27] Zaliznjak Anna A., Mikajeljan I. L., Shmelev A. D. Russian aspectology: In defense of aspectual pair [Russkaja aspektologija: V zashhitu vidovoj pary.] — M.: Slavic Culture Languages [M.: Jazyki slavjanskoj kul'tury], 2015.
- [28] Zehr, J., & Schwarz, F. PennController for Internet Based Experiments (IBEX), 2018 <https://doi.org/10.17605/OSF.IO/MD832>
- [29] <https://toloka.yandex.com/>

RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora

Daryna Dementieva^{1,5*}, Varvara Logacheva^{1*}, Irina Nikishina¹,
Alena Fenogenova³, David Dale¹, Irina Krotova², Nikita Semenov²,
Tatiana Shavrina^{3,4}, and Alexander Panchenko¹

¹Skolkovo Institute of Science and Technology (Skoltech), ²Mobile TeleSystems (MTS),
³SberDevices (Sber), ⁴AI Research Institute (AIRI), ⁵Technical University of Munich (TUM)
{daryna.dementieva, v.logacheva, irina.nikishina, d.dale, a.panchenko}@skoltech.ru,
fenogenova.a.s@sberbank.ru, i.krotova@mts.ai, nikita.semenov@mts.ru, rybolos@gmail.com

Abstract

Text detoxification is the task of rewriting a toxic text into a neutral text while preserving its original content. It has a wide range of applications, e.g. moderation of output of neural chatbots or suggesting less emotional version of posts on social networks. This paper provides a description of RUSSE-2022 competition of detoxification methods for the Russian language. This is the first competition which features (i) parallel training data and (ii) manual evaluation. We describe the setup of the competition, the solutions of the participating teams and analyse their performance. In addition to that, the large-scale evaluation allows us to analyse the performance of automatic evaluation metrics.

Keywords: Style transfer, detoxification, corpus, text generation, evaluation, competition, metrics analysis
DOI: 10.28995/2075-7182-2022-21-114-131

RUSSE-2022: первое соревнование по детоксификации русских текстов на основе параллельного корпуса

Дарина Дементьева^{1,5*}, Варвара Логачева^{1*}, Ирина Никишина¹,
Алена Феногенова³, Давид Далей¹, Ирина Кротова², Никита Семенов²,
Татьяна Шаврина^{3,4}, и Александр Панченко¹

¹Сколковский институт науки и технологий (Сколтех),
²Мобильные ТелеСистемы (МТС), ³SberDevices (Сбер), ⁴AI Research Institute (AIRI),
⁵Технический университет Мюнхена (ТУМ)
{daryna.dementieva, v.logacheva, irina.nikishina, d.dale, a.panchenko}@skoltech.ru,
fenogenova.a.s@sberbank.ru, i.krotova@mts.ai, nikita.semenov@mts.ru, rybolos@gmail.com

Аннотация

Детоксикация текста — это задача преобразования токсичного текста в нейтральный текст с сохранением его исходного содержания. Технологии детоксификации имеют широкий спектр применений, таких как модерация вывода чатботов или перефразирование эмоционального комментария в социальной сети. Данная статья посвящена описанию соревнования моделей для детоксификации текстов RUSSE-2022. Это первое подобное соревнование, в рамках которого были доступны (1) параллельный обучающий корпус и (2) ручная оценка моделей. Мы описываем в данной работе соревнование и модели, участвовавшие в нем, а также анализируем их работу. Кроме того, проведенная ручная оценка качества позволила нам проанализировать автоматические метрики, используемые для оценки качества моделей детоксификации и переноса стиля.

Ключевые слова: Перенос стиля, детоксификация, корпус, генерация текста, оценка качества, соревнование, анализ метрик

* Equal contribution

1 Introduction

Identification of toxicity in user texts is an active area of research. Today, social networks such as Facebook¹, Instagram², and VK³ are trying to address the problem of toxicity. However, they usually simply block such kinds of texts. We suggest a proactive reaction to toxicity from the user. Namely, we aim at presenting a neutral version of a user message which preserves meaningful content. We denote this task as *detoxification*.

Detoxification can be solved with Text Style Transfer (TST) (Jin et al., 2020; Hu et al., 2020) methods. This task aims at transforming the text so that its content stays the same, and its *style* (which can refer to text sentiment, author profile, degree of politeness or formality) changes. For the majority of style transfer tasks there exists no parallel data, which makes researchers train TST models on non-parallel texts (Shen et al., 2017; Wang et al., 2019; Xu et al., 2021).

Detoxification task is usually considered a variety of TST task from *toxic* to *neutral* style. There already exist unsupervised approaches to detoxification (Dementieva et al., 2021a; Dale et al., 2021) for the Russian and English languages. However, the output of these models is often of bad quality.

Russian IT company Yandex already tried to address the detoxification problem and launched the first detoxification competition. However, we extended their setup in several directions:

- We collected a **new parallel corpus** of toxic sentences and their manually written non-toxic paraphrases. This allows to solve the detoxification task using the methods developed for parallel training data (in particular, for machine translation).
- We use an **established and tested automatic evaluation setup** (Krishna et al., 2020) which agrees with the formulation of style transfer and takes into account all aspects of transfer quality. In addition to that, we use the reference-based evaluation.
- We follow the common assumption of low reliability of automatic evaluation of style transfer and make the final decision on the models quality based on the **manual evaluation**. Our work is the first attempt to use crowdsourcing for large-scale manual evaluation of a text generation model. We describe and analyse our evaluation setup.

All general information about the presented competition as well as all used code, data, and the final results can be obtained via official website.⁴

2 Related Work

RUSSE’2022 is the first competition on detoxification based on parallel corpora for Russian and has no analogies in any languages. As for the Russian, the first detoxification was launched by Yandex in november 2021 (Yandex, 2021). However, the dataset did not include parallel data that prevented participants from using seq2seq models. Moreover, their evaluation setup was weak as it only included toxicity measuring as well as similarity to the initial text and was hacked by participants.

At the same time, a lot of attempts have been made in studying toxicity for the English language. The earliest ones were several Kaggle competitions from the Jigsaw/Conversation AI team on toxicity: the “Toxic Comment Classification Challenge” (Jigsaw, 2018) in 2018, the “Unintended Bias in Toxicity Classification Challenge” (Jigsaw, 2019) in 2019 and the “Multilingual Toxic Comment Classification Challenge” (Jigsaw, 2020) in 2020. The organizers present the largest English toxicity datasets with multiple types of toxicity (toxic, obscene, threat, insult, identity hate, etc) and a multilingual test set for other languages such as Spanish, French, Italian, Russian, etc.

Since 2019 toxicity and offensive language becomes one of the central topics at SemEvals. SemEval-2019 Task 6 and SemEval-2020 Task 12 on Identifying and Categorizing Offensive Language in Social

¹<https://edition.cnn.com/2021/06/16/tech/facebook-ai-conflict-moderation-groups>

²<https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse>

³<https://vk.com/press/stickers-hate-speech>

⁴<https://russe.nlpub.org/2022/tox>

Media (OffensEval) (Zampieri et al., 2019; Zampieri et al., 2020) attracted about 115 teams for the first year and 145 for the second year. The next SemEval competition devoted to the toxicity held in 2021: Toxic Spans Detection⁵ (Pavlopoulos et al., 2021) is a task which aims at identifying the spans that make a text toxic instead of classifying whole texts in comparison to OffensEval and Jigsaw competitions. Highlighting such toxic spans can assist human moderators who often deal with lengthy comments. The 2022 year is also quite eventful for competitions on toxicity. For instance, there is a competition called “Multimedia Automatic Misogyny Identification (MAMI)”⁶ which aims at identification of misogynous memes, raising a topical issue of systematic inequality and discrimination of women online. The competition examines memes as a form of hate against women, taking advantage of both text and images available as source of information.

Another SemEval 2022 task: “Patronizing and Condescending Language Detection”⁷ focuses on categorizing sentences in context (paragraphs), extracted from news articles, in which one or several pre-defined vulnerable communities are mentioned. The task is to identify whether the unfair treatment in the media is expressed in the text and the correct category of the Patronizing and Condescending Language.

Additionally, we pay attention to the SemEval competition of this year called “iSarcasmEval: Intended Sarcasm Detection In English and Arabic”⁸. Sarcasm is omnipresent on the social web and often present in toxic texts. Determination of sarcastic texts could be also beneficial for the text detoxification process.

It can be seen that none of the previous competitions provide parallel datasets for performing detoxification and only aim at text classification and not paraphrasing. Our competition has been inspired by the Machine Translation shared tasks as it also applies parallel data and adopts some of the evaluation techniques from Machine Translation (MT) (Akhbardeh et al., 2021). It is the first parallel dataset in Russian on the topic of detoxification. In this work we present such dataset for the first time as well as the results of shared task on that data.

3 Parallel Detoxification Dataset

To perform training and automatic evaluation we provide a parallel detoxification dataset. The dataset is the core innovation of our shared task as previous detoxification shared task relied on non-aligned text corpora.

3.1 Definition of Toxicity

Our shared task deals with one particular style - toxicity. Namely, the goal is to rewrite text from toxic to neutral. What is and what is not toxic is a crucial question which shapes the training dataset and influences the performance of detoxification models. In our work we decide to consider only cases of open toxicity: open offences, use of swear and rude words. We do not focus on subtle forms of toxicity such as sarcasm or passive aggression, since they are difficult to identify not only for machines, but also for untrained human assessors. We leave work on these types of toxicity for future work.

We should warn against conflating toxicity with sentiment. Non-toxic sentences are not necessarily pleasant, they can still contain criticism such as *bad person*, *liar*, etc. Since our task is to detoxify a text while saving its content, we allow keeping negative content.

It is important to explain our understanding of toxicity to crowd workers. We use the example-based approach. Namely, instead of definitions of what is toxic we give users examples of sentences which we consider offensive and neutral. We do so in the instruction which workers need to read before doing tasks and which they can refer to later (the full text of the instruction is given in Appendix B.1). Also, since we noticed that users often skip the instruction, we ask them to take the training. It consists of examples of toxic and neutral sentences with the explanation of their label (toxic/neutral). See the examples of training questions in Appendix B.2. After that, the user passes an exam which shows if she understands the notion of toxicity correctly. We only admit users who have the result of above 80%.

⁵<https://sites.google.com/view/toxicspans>

⁶<https://competitions.codalab.org/competitions/34175>

⁷<https://sites.google.com/view/pcl-detection-semeval2022>

⁸<https://sites.google.com/view/semeval2022-isarcasmeval>

Further during the labelling, we control users by occasionally giving them control questions and reinforce their understanding of toxicity by giving training questions.

3.2 Dataset Summary

We take source (toxic) sentences for our dataset from the Russian datasets of toxic messages from various social media: Odnoklassniki (Belchikov, 2019), Pikabu (Semiletov, 2020), and Twitter (Rubtsova, 2012). The target part of the dataset are the same messages which were manually rewritten by crowd workers to eliminate toxicity.

The dataset is divided into train, development, and test sets of the following sizes:

- training: 6 948 source (toxic) sentences,
- development: 800 source (toxic) sentences,
- test: 875 source (toxic) sentences.

For each toxic sentence we have 1–3 variants of detoxification. The examples of samples collected for the task are presented in Appendix C.

3.3 Data Collection Pipeline

To collect the dataset for this competition we hired workers via Yandex.Toloka platform. We use the pipeline for the parallel detoxification data collection which was described in the work (Dementieva et al., 2021b) and tested for English. In this work we improved this pipeline and adapted it for the Russian language.

The pipeline consists of three tasks:

- **Paraphrase generation** — the workers are asked to write a neutral paraphrase of the input text. They can also select not to rewrite the input if the text is already neutral or it is difficult to extract non-toxic content. The paraphrases generated by crowd workers can be of poor quality. Therefore, we validate them using the next two tasks.
- **Content preservation check** — given two texts (the original toxic sentence and its crowdsourced paraphrase) an annotator should indicate if the content of the texts matches.
- **Toxicity classification** — given the generated paraphrase, an annotator should label it as toxic or neutral.

During the dataset collection we tried to exclude examples which are impossible to detoxify. These are (i) sentences whose meaning is offensive, (ii) sentences which aren't toxic so can't be detoxified, and (iii) sentences with unclear meaning. See the following examples:

- **Toxic content:**
 - пристрелить этих уродов без суда и следствия (*shoot these freaks without trial*)
 - а что ты с*ка умеешь, только ноги раздвигать... (*and what can you b*tch, you can only spread your legs*)
 - п*доры они в квадрате с*ки. (*f*gs are squared b*tches.*)
- **Unclear meaning:**
 - ч оз тема ч о класс ответить д лёка продаю п*зду дочери комментарий (*h oz topic h about class answer d loka sell pussy daughter comment*)

Paraphrasing sentences with toxic content cannot remove toxicity, and if we manage to remove it, the sense of such sentence will be very different from the original one.

To increase the reliability of crowdsourcing, we have each example labelled by three crowd workers. In case of paraphrase generation this gives us multiple paraphrases (some of them are filtered out later). When doing content and toxicity checks, we get multiple judgments on each example. They are further aggregated with Dawid-Skene aggregation method (Dawid and Skene, 1979) which defines the true label iteratively giving more weight to the answers of workers who agree with other workers more often. Besides the true label, this method returns the label confidence. We consider a paraphrase correct with respect to content and toxicity if it is labelled as such with the confidence of over 90%.

4 Shared Task Description

4.1 Task Formulation

Text detoxification can be considered as a kind of textual style transfer task. The style transfer task is formulated as follows. We would like to rewrite a text so that it keeps most of its content, but one particular attribute of this text (denoted as *style*) changes. The “style” can refer to various features of the text such as the level of formality, politeness, simplicity, the presence of bias or the features of the author (e.g. gender or membership in a political party). The task is usually to transfer between two “opposite” styles (toxic–neutral, formal–informal, ancient–modern), but there can exist models which support multiple exclusive or non-exclusive styles. More formally, the notion of a “style” is defined below. We deliberately, rely on a practical notion assuming that style is an automatically measurable text attribute. A more comprehensive formal definition of all various styles is a challenging task beyond the scope of our work.

Style transfer task can be formally defined as follows. We have a set of styles $S = \{s_{src}, s_{tg}\}$ ⁹ and two collections of documents: the source corpus $D^{src} = \{d_1^{src}, \dots, d_n^{src}\}$ and the target corpus $D^{tg} = \{d_1^{tg}, \dots, d_m^{tg}\}$ in the styles s_{src} and s_{tg} , respectively. Let us also define the following functions. The style of a sentence is measured with $\sigma : D \rightarrow S$. A binary function $\delta : D \times D \rightarrow \{0, 1\}$ indicates the equivalence of meanings of the two styles. Finally, the function $\theta : D \rightarrow \{0, 1\}$ defines if a text belongs to well-formed sentences.

Text style transfer task is thus defined as a function $\alpha : S \times S \times D \rightarrow D$. Given a text d^{src} and its source and target styles s_{src} and s_{tg} it transforms the text to a new text d^{tg} such that:

- the style of the text is changed from the source s_{src} to the target s_{tg} : $\sigma(d^{src}) \neq \sigma(d^{tg}), \sigma(d^{tg}) = s_{tg}$,
- the contents of the original and the transformed sentences match: $\delta(d^{src}, d^{tg}) = 1$,
- the resulting sentence is well-formed (fluent): $\theta(d^{tg}) = 1$.

Therefore, a style transfer model has to optimize all three functions. Analogously, to evaluate the performance of a style transfer model, we need to check that all three conditions hold: the style is appropriately changed, the content stayed intact, and the text is fluent.

4.2 Competition Rules

The competition was opened on December 15, 2021 and lasted until February 28, 2022. It consisted of the following stages:

- **Development stage** — this stage lasted from December 15, 2021 to January 31, 2022. At this stage we made available the training and development data. The participants were invited to train their models and submit their outputs for the development set to the public leaderboard at Codalab.¹⁰ At this stage, the models were evaluated with the automatic metrics.

⁹Style transfer task can be generalized for S with more than two styles or for continuous styles. We use the binary case for simplicity.

¹⁰<https://codalab.lisn.upsaclay.fr/competitions/642>

- **Test stage** — this stage lasted from February 1 to 14, 2022. At the beginning of this stage the participants were given access to the source part of the test set. They had two weeks to run their best-performing models on the test set and submit their answers to Codalab. The test stage leaderboard was hidden until the end of the competition.
- **Manual evaluation stage** — this stage lasted from February 14 to 28, 2022. At this stage we conducted the manual evaluation of the test answers submitted by participants and the baseline answers. The evaluation was performed via crowdsourcing. At the end of this stage we released the final leaderboard based on the results of manual evaluation.

We allowed participants to use detoxification models of any architecture. Participants were allowed to use any additional data and existing pre-trained models under open source licences.

Once results were submitted, we required participants to provide their source code and model via GitHub and also write its short description.

4.3 Baselines

We provide four baselines for detoxification task: a trivial Duplicate baseline, a rule-based Delete approach, fine-tuning on the ruT5 model and the continuous prompt tuning approach for ruGPT3 model.

Duplicate This is a trivial baseline which consists in leaving the input text intact. It provides a lower threshold for models.

Delete Delete is an unsupervised method that eliminates toxic words based on a predefined toxic words vocabulary. The idea is often used on television and other media: rude words are bleeped out or hidden with special characters (usually an asterisk). We provide both the vocabulary and the script that performs the replacement.

RuT5 Baseline Another approach is the supervised baseline based on the T5 model. We fine-tune the ruT5-base model¹¹ on the training part of the provided dataset.

RuPrompts This baseline is based on the ruPrompts library¹² for fast language model tuning via automatic prompt search. The Continuous Prompt Tuning method (Konodyuk and Tikhonova, 2021) consists in training embeddings corresponding to the prompts. Such approach is cheaper than classic fine-tuning of big language models. We tune the prompts for the ruGPT3-large model.¹³ Pre-trained prompts for detoxification task are available online.¹⁴

5 Evaluation

We use two evaluation setups: automatic evaluation with reference-free and reference-based metrics and manual multi-aspect evaluation.

5.1 Automatic Evaluation

In our automatic evaluation we follow the state-of-the-art evaluation strategies. Namely, we replicate the setup of (Krishna et al., 2020). We evaluate the three parameters of style transfer quality: style of text, content preservation, and fluency of text.

Note that these three parameters exactly correspond to the TST definition components as formulated in Section 4.1: namely functions $\sigma(\cdot)$, $\delta(\cdot, \cdot)$, and $\theta(\cdot)$. The three metrics are then aggregated to a joint score. We use the following techniques.

Style (STY_a) is evaluated with a BERT-based classifier for toxicity detection. We fine-tune the ruBERT model (Kuratov and Arkhipov, 2019) on the Odnoklassniki (Belchikov, 2019) and Pikabu (Semiletov, 2020) datasets. Style accuracy is denoted as $\sigma(\cdot)$ in Section 4.1.

¹¹<https://huggingface.co/sberbank-ai/ruT5-base>

¹²<https://sberbank-ai.github.io/ru-prompts>

¹³<https://github.com/sberbank-ai/ru-gpts>

¹⁴https://huggingface.co/konodyuk/prompt_rugpt3large_detox_russe

Content (SIM_a) is evaluated as the cosine similarity of embeddings of the source and the transformed sentences. We use embeddings generated by LaBSE model (Feng et al., 2020) because in our preliminary experiments they showed the best performance for Russian. We prefer the embedding distance over BLEU-like metrics, because (Yamshchikov et al., 2021) showed that embedding-based metrics are better correlated with human judgments than ngram-based metrics such as BLEU. We do not use references for the evaluation of content to mimic the setup where references are unavailable, which is very common for style transfer tasks. Content similarity is denoted as $\delta(\cdot, \cdot)$ in Section 4.1.

Fluency (FL_a) Although fluency is usually evaluated as perplexity, we follow (Krishna et al., 2020) and use an acceptability classifier. In this work this classifier was trained on CoLA dataset (Warstadt et al., 2019). Since there is no such dataset for Russian, we create synthetic examples of corrupted sentences by randomly replacing, deleting or shuffling words in sentences as suggested by (Kann et al., 2018). We choose this method over perplexity, because it ranges from 0 to 1 and its greater values mean higher quality, just like metrics we use for evaluating toxicity and content. This makes combining the three metrics easier. Fluency is denoted as $\theta(\cdot)$ in Section 4.1.

Joint (J_a) Following (Krishna et al., 2020), we combine the three metrics at the sentence level by multiplying them. Since all scores are binary, the joint score is 1 only if all three metrics are 1. Therefore, it indicates fully acceptable sentences.

$$\mathbf{J} = \frac{1}{n} \sum_{i=1}^n \mathbf{STA}(x_i) \cdot \mathbf{SIM}(x_i) \cdot \mathbf{FL}(x_i) \quad (1)$$

ChrF We provide an additional reference-based metric which follows the Machine Translation evaluation setup. We choose ChrF (Popović, 2015) over BLEU, because it compares character ngrams and is more suitable for languages with rich morphology, such as Russian.

5.2 Manual Evaluation

The manual evaluation follows setups used in state-of-the-art works. We separately evaluate the three parameters of the transferred sentences, namely, their style, content, and fluency. We conduct the evaluation via crowdsourcing. For the evaluation we also use Yandex.Toloka platform.

5.2.1 Evaluation Metrics

All three parameters are evaluated at the sentence level in terms of a binary scale, where 0 refers to the bad quality in terms of the parameter and 1 is the good quality. Assessors are given the following guidelines.

Toxicity (STY_m) The toxicity level is defined as:

- **non-toxic** (1) — the sentence does not contain any aggression or offence. However, we allow covert aggression and sarcasm. Note also that toxicity should not be mixed with the lack of formality. Even if a sentence is extremely informal, it is non-toxic unless it attacks someone.
- **toxic** (0) — the sentence contains open aggression and/or swear words (this also applies to senseless sentences).

Content (SIM_m) In terms of content, sentences should be classified as:

- **matching** (1) — the output sentence fully preserves the content of the input sentence. Here, we allow some change of sense which is inevitable during detoxification (e.g. replacement with overly general synonyms: *idiot* becomes *person* or *individual*). It should also be noted that content and toxicity dimensions are independent, so if the output sentence is toxic, it can still be good in terms of content.

- **different** (0) — the sense of the transferred sentence is different from the input. Here, the sense should not be confused with the word overlap. The sentence is different from its original version if its main intent has changed, (cf. *I want to go out* and *I want to sleep*). The partial loss or change of sense is also considered a mismatch (cf. *I want to eat and sleep* and *I want to eat*). Finally, when the transferred sentence is senseless, it should also be considered *different*.

Fluency (FL_m) The fluency evaluation is different from the other metrics. We evaluate it along a ternary scale with the following values:

- **fluent** (1) — sentences with no mistakes, except punctuation and capitalisation errors.
- **partially fluent** (0.5) — sentences which have orthographic and grammatical mistakes, non-standard spellings. However, the sentence should be fully intelligible.
- **non-fluent** (0) — sentences which are difficult or impossible to understand.

However, since all the input sentences are user-generated, they are not guaranteed to be fluent in terms of this scale. People often make mistakes, typos and use non-standard spelling variants. We cannot require that a detoxification model fixes them. Therefore, we consider an output of a model fluent if the model did not make less fluent than the original sentence. Thus, we evaluate both the input and the output sentences and define the final fluency score as **fluent** (1) if the fluency score of the output is greater or equal to that of the input, and **non-fluent** (0) otherwise.

Joint Score (J_m) Finally, We aggregate the three metrics in the same Joint score as it was done for automatic evaluation.

Note that, in manual evaluation setup, we again resort to the original TST formulation based on three functions as defined in Section 4.1: $\sigma(\cdot)$, $\delta(\cdot, \cdot)$, and $\theta(\cdot)$. However, in this case, their outputs are defined not in an automatic way but rather using human judgements.

5.2.2 Crowdsourcing Setup

Each of the three parameters is evaluated in a separate crowdsourcing project. For all the projects we hire only native speakers of Russian.

Crowdsourcing tasks In the toxicity detection task (see Figure 1) we show workers the transferred sentence and ask them if it is offensive. Then, in the content similarity task we show both sentences and ask if they mean the same. Finally, we apply the fluency evaluation task to both the source and the target and compute the final fluency score from the source and target scores. While here we provide English interfaces examples, the original interfaces are presented in Appendix A.

Each sentence in each of the projects is labelled by 10 to 12 workers. We aggregate their result using Dawid-Skene aggregation method (Dawid and Skene, 1979). It takes into account the dynamically defined reliability of workers. For each example with multiple labels Dawid-Skene method returns the label and its confidence. We use only labels whose confidence is above 90%. The other labels (around 3% of all examples) are later filled by experts.

Quality Control Before admitting users to accomplishing tasks we need make sure they understand them correctly. For that purpose we devise a pipeline of training and exam tasks. First, a user needs to pass training (a set of tasks with a known label and an explanation of the task shown if the user makes a mistake) and exam (same as training, but no explanations are shown). We only admit users whose exam score is above 80%. Similarly, we control their performance with control questions during labelling. We ban users whose performance on these control question is below 70%.

Finally, we use other heuristics to control the user performance:

- **captcha** — prevents workers from using scripts and bots for labelling,
- **fast answers** — we ban users who accomplish a page of tasks in less than 15 seconds (this usually means that the user is not reading the task and is giving random answers),

- **skipped tasks** — we ban users who skip 5 or more task pages (this indicates a user who does not understand the task).

6 Participating Systems

Ten teams participated in the final phase of the competition. Here we briefly describe them. For the easier navigation in the leaderboard, we provide the models aliases which summarise the methods they use.

orzhan (ruT5-finetune) approach is based on the ruT5-base model¹⁵. It was fine-tuned on the part of competition train data with a learning rate 1e-5 on 15 epochs. Only the samples with fluency, similarity, and accuracy higher than 0.5 were selected from the train set. The best output is selected from 32 generated samples using beam search. It was decided not to use sampling.

NSU team (ruGPT3-filter) This team’s solution uses a model based on ruGPT3. The authors filtered the dataset released by the organizers with the following heuristics: (i) cosine similarity between the original and transformed sentences ranges from 0.6 to 0.99; (ii) ROUGE-L between the sentences ranges from 0.1 to 0.8; (iii) the transformed sentence length is less or equal to the original sentence length. This dataset was used to fine-tune ruGPT3.

Mindful Squirrel (lewis) solution is based on the LEWIS framework (Reid and Zhong, 2021), a coarse-to-fine editor for style transfer that transforms text using Levenshtein edit operation. First, the sequence of coarse-grain Levenshtein edit types (keep, replace, delete or insert) was predicted for each sentence pair. Next, the resulting tags were used to train the conversational RuBERT¹⁶ for the sequence tagging task. The ruT5-base model was trained to fill in the tokens for coarse-grain edit type *replace*.

king_menin (ruGPT3-XL) trained RuGPT3 XL¹⁷ to generate a non-toxic text on the competition train data. The input is the concatenation of the toxic and non-toxic sentences.

¹⁵<https://huggingface.co/sberbank-ai/ruT5-base>

¹⁶<https://huggingface.co/DeepPavlov/rubert-base-cased-conversational>

¹⁷<https://huggingface.co/sberbank-ai/rugpt3xl>

The figure shows three separate user interface forms for crowdsourcing tasks. Each form has a light gray background and rounded corners.

- Top Left Form:** The question is "Does this text contain offenses or swear words?". The text input field contains "I don't care about that." Below the input are two radio buttons labeled "Yes" and "No".
- Top Right Form:** The question is "Do these sentences mean the same?". The text input field contains "I don't f*ckin care about that shit" (highlighted in red) and "I don't care about that" (highlighted in green) on separate lines. Below the input are two radio buttons labeled "Yes" and "No".
- Bottom Form:** The question is "Is this text grammatical?". The text input field contains "I don't care about that." Below the input are three radio buttons with labels: "YES, there are no or only minor mistakes", "PARTIALLY, there are mistakes, but the text is intelligible", and "NO, the text is difficult to understand".

Figure 1: Design of crowdsourcing user interfaces of the toxicity detection (top left), content check (top right), and fluency check (bottom) tasks. Forms were translated from Russian to English for readability.

anzak (RoBERTa-replace) solution is based on the RoBERTa-large¹⁸. The logistic regression model on the FastText vectors trained on the competition data was used as a toxic words classifier. Toxic tokens were substituted by RoBERTa-large model, where the best candidates were chosen by the cosine similarity between the candidate and the toxic token. In case it was not possible to find an acceptable candidate, the toxic word was removed from the sentence.

SomethingAwful (ruT5-clean) used the ruT5-large model¹⁹ improved by data cleaning. The preprocessing stage consists of emoticons and smiley filtering and removing duplicate characters. The Levenshtein Transformer (Gu et al., 2019) was used as an extra step in preprocessing to clean the ruT5-large model output.

FRC CSC RAS (ruT5-large) modified the t5 baseline. RuT5-base was replaced by ruT5-large with beam search used as inference algorithm. 20 candidates were generated for each toxic sentence, the best candidate was selected by the largest J-score metric.

barracudas (ruT5-preproc) This solution is based on ruT5-base model with additional pre- and post-processing of the texts.

gleb_shnshn (adversarial) This team devised an adversarial training setup where the training data was enriched with the artificially generated sentences which attained the highest scores of the automatic metrics.

ruPrompts-plus (ruPrompts-plus) This team advanced over the ruPrompts baseline. The solution is based on RuGPT3-XL²⁰ adapted to the task via prompt tuning. In particular, the participant prepended 100 and appended 20 trainable embeddings to the toxic text and passed it to the model, which was expected to output the detoxified version. These embeddings were directly optimized by gradient descent.

7 Results

The primary goal of our competition is to evaluate the models and understand which approach is more promising. Here we compare the performance of models in terms of manual and automatic metrics. Besides that, since we have both manual and automatic scores, we evaluate the performance of metrics themselves.

7.1 Models Performance

Table 1 shows the performance of the participating models and our baselines in terms of the automatic metrics. The adversarial example generation (**gleb_shnshn**) turns out to be very effective — it attains the highest scores of all metrics, thus yielding the highest J_a score. The next three places in the leaderboard are taken by the models based on our baseline ruT5 system (**orzhan**, **FRC CSC RAS**, and **SomethingAwful**). This suggests that this model is very efficient. Notice that the human references are below the majority of models in terms of all metrics except ChrF whose score for the human references is the highest by a large margin.

It is also important to note that the highest content preservation is demonstrated by two models from the bottom of the leaderboard, namely, the **Delete** baseline and **anzak** team’s model. Both of them do not generate the output text from scratch but only remove or change individual words. This approach yields sentences which are very similar to the original ones.

The manual scores (see Table 2) provide a completely different result. There, the human references are significantly better than other models, but closely followed by the solution by the **SomethingAwful** team. This team is the only team whose solution succeeded in outperforming the ruT5-based baseline model. The winning team’s model is also based on ruT5 (although they use ruT5-large), but with the additional preprocessing. The model of the **FRC CSC RAS** team, which got the 3rd best result in terms of fluency

¹⁸<https://huggingface.co/sberbank-ai/ruRoberta-large>

¹⁹<https://huggingface.co/sberbank-ai/ruT5-large>

²⁰<https://huggingface.co/sberbank-ai/rugpt3xl>

Team	Method	ACC _a	SIM _a	FL _a	J _a	ChrF
gleb_shnshn	adversarial	0.97	0.94	0.96	0.87	0.53
orzhan	ruT5-finetune	0.98	0.86	0.97	0.82	0.55
FRC CSC RAS	ruT5-large	0.95	0.86	0.97	0.78	0.57
SomethingAwful	ruT5-clean	0.95	0.82	0.91	0.71	0.57
Mindful Squirrel	lewis	0.93	0.80	0.88	0.66	0.56
king_menin	ruGPT3-XL	0.94	0.73	0.89	0.61	0.50
baseline	RuT5	0.80	0.83	0.84	0.56	0.57
ruPrompts-plus	ruGPT-XL+ruprompts	0.80	0.80	0.83	0.54	0.56
baseline	ruPrompts	0.81	0.79	0.80	0.53	0.55
barracudas	ruT5-preproc	0.85	0.76	0.78	0.52	0.53
human references	manual annotation	0.85	0.72	0.78	0.49	0.77
NSU team	ruGPT3-filter	0.83	0.76	0.76	0.48	0.51
anzak	RoBERTa-replace	0.57	0.89	0.91	0.44	0.54
baseline	Delete	0.56	0.89	0.85	0.41	0.53
baseline	Duplicate	0.24	1.00	1.00	0.24	0.56

Table 1: The performance of the participating models in terms of automatic metrics, sorted by J_a metric. The values **in bold** show the highest value of the metric with the significance level of $\alpha = 0.05$.

and content preservation, is also based on ruT5-large model. This confirms that large pretrained models with fine-tuning on parallel data are a very strong baseline which is hard to beat.

Interestingly, the **adversarial** model whose automatic scores are the highest, in fact produces sentences of a very low quality. This shows that automatic metrics can be “fooled” and should not be used as an ultimate evaluation technique.

In terms of the quality of style change, the model of the **Mindful Squirrel** team yielded the best result which was only outperformed by human references. This model uses a word classifier which decides if a word should be changed or left intact during style transfer. This allows to focus on toxic words.

Overall, the evaluation shows that the models based on **ruT5** fine-tuned on parallel data are the most successful. The two teams that used **ruGPT3** could not approach the results of the competitors. The tuning of prompts is still less efficient than tuning of models. The models based on explicit edit operations are only moderately successful.

7.2 Automatic vs Manual Metrics

The automatic and manual metrics (Tables 1 and 2) provide very diverse results. This suggests that they are weakly correlated.

We check this assumption by computing the Spearman ρ correlations for document-level scores of all metrics. We put in bold all high correlations (p -value ≤ 0.05) in Table 3. We clearly see that none of automatic metrics correlate with their manually measured counterparts. On the other hand, manual style and content metrics are correlated with ChrF score. This suggests that ChrF can be used as an automatic evaluation score. On the other hand, ChrF is not sensitive to sentence style, which means that it can be deceived (for example, the trivial Duplicate baseline performs on par with strong T5-based models in terms of ChrF). However, the power of ChrF was also claimed by (Briakou et al., 2021).

The sentence-level correlations show a slightly different picture. The highest correlation is seen for the style metric, the Spearman ρ score of automatic and manual judgments is 0.418 (moderate correlation). The manual and automatic sentence-level similarity, fluency, and joint scores show very weak or no correlation: 0.251, 0.015, and 0.141, respectively.

However, sentence-level correlations between corresponding manual and automatic metrics differ significantly across models (see Figure 2). We see that automatic and manual toxicity scores are much

Team	Method	ACC _m	SIM _m	FL _m	J _m
human references	manual annotation	0.89	0.82	0.89	0.65
SomethingAwful	ruT5-clean	0.79	0.87	0.90	0.63
baseline	RuT5	0.79	0.82	0.92	0.61
FRC CSC RAS	ruT5-large	0.73	0.87	0.92	0.60
Mindful Squirrel	lewis	0.82	0.79	0.85	0.58
ruPrompts-plus	ruGPT-XL+ruprompts	0.78	0.81	0.90	0.57
orzhan	ruT5-finetune	0.80	0.78	0.87	0.56
barracudas	ruT5-preproc	0.79	0.72	0.78	0.51
king_menin	ruGPT3-XL	0.81	0.70	0.90	0.50
baseline	ruPrompts	0.80	0.70	0.87	0.49
NSU team	ruGPT3-filter	0.77	0.72	0.83	0.45
anzak	RoBERTa-replace	0.43	0.62	0.79	0.17
baseline	Delete	0.39	0.71	0.73	0.16
baseline	Duplicate	0.11	1.00	1.00	0.11
gleb_shnshn	adversarial	0.25	0.13	0.24	0.02

Table 2: Manual evaluation of the participating models, the models are sorted by the J_m metric. The values **in bold** show the highest value of the metric with the significance level of $\alpha = 0.05$.

Metric	STA _a	SIM _a	FL _a	J _a	ChrF
STA _m	0.376	-0.776	-0.398	0.278	0.223
SIM _m	-0.046	0.031	0.190	0.000	0.789
FL _m	-0.083	-0.032	0.288	0.070	0.619
J _m	0.326	-0.495	-0.211	0.350	0.735

Table 3: Spearman’s correlation coefficient between automatic VS manual metrics on system level. Bold numbers denote the statistically significant correlation (p -value ≤ 0.05).

better correlated for the **Delete** and **anzak** models, which are the only models to explicitly remove or replace toxic words identified by a classifier or via a manually compiled list of toxic words. These models apparently produce texts which are easy to classify correctly. Conversely, **gleb_shnshn** model and **human references** are the most difficult to classify. The former deliberately “fools” the classifier with artificial examples, while the latter contains non-trivial phrases whose level of toxicity is difficult to grasp automatically.

Analogously, the similarity scores are also better correlated for the **anzak** model which leaves the majority of words intact, so for it similarity boils down to word matching. On the other hand, T5-based models produce non-trivial paraphrases. These T5 outputs are also difficult to correctly classify for fluency, unlike the models based on word replacements (**anzak** and **Delete**). Overall, we see that it is more difficult to correctly classify outputs of *better-performing models* and *models based on large pre-trained language models* than the simple baseline approaches. This suggests that the automatic evaluation might fail exactly where we need it most, i.e. in discriminating between the good models.

8 Conclusions

We organised a competition on text detoxification for the Russian language. To the best of our knowledge, this is the second such competition. This is also the first detoxification challenge that used manual evaluation. For the needs of competition we created the first parallel Russian corpus for detoxification enabling the use of supervised machine translation approaches to this task.

Our analysis of model performances showed that the best result is attained by models based on the pre-trained ruT5 model fine-tuned on our parallel data. This model produces sentences which were evaluated closely to the human references. This shows that pre-trained Transformers are very powerful and are

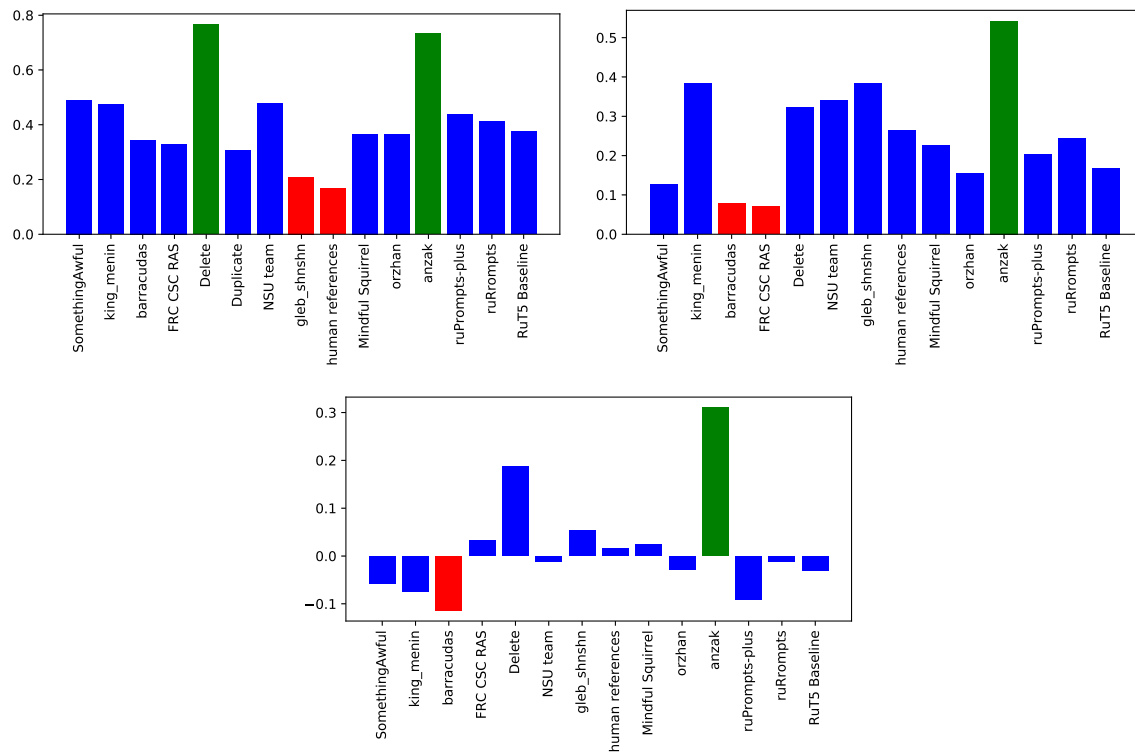


Figure 2: Correlations between automatic and manual metrics at the sentence level for different models: correlation of style accuracy scores (top left), correlation of similarity scores (top right), correlation of fluency scores (bottom). Red and green bars indicate the lowest and the highest values, respectively.

difficult to beat.

We conducted an evaluation of detoxification models for Russian using both automatic and manual metrics. This allowed us to analyse the relationship between the metrics and assess the suitability of automatic metrics for evaluation.

Our analysis shows that the metrics are overall weakly correlated with the human judgements both at the system and the sentence level. We found that ChrF score has a strong correlation with the joint score of style, content, and fluency. Thus, ChrF could be used as a proxy for manual evaluation, but its lack of correlation with the style score makes this metric vulnerable to attacks. We also discovered that the correlation of manual and automatic scores varies for different models. This shows the necessity to consider diverse style transfer models for metrics analysis.

Overall, although the state-of-the-art evaluation setup for style transfer (three parameters and the joint score combined from them) is conceptually correct, the current performance of automatic metrics is insufficient to use it as a replacement for manual evaluation. More research is needed to better fit the quality of manual evaluation.

Acknowledgements

This work was supported by the joint MTS-Skoltech laboratory on AI. The manual evaluation was supported by a Yandex.Toloka research grant.

References

- Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). // *Proceedings of the Sixth Conference on Machine Translation*, P 1–88, Online, November. Association for Computational Linguistics.
- Anatoly Belchikov. 2019. Russian language toxic comments. <https://www.kaggle.com/blackmoon/russian-language-toxic-comments>. Accessed: 2021-07-22.
- Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 1321–1336, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 7979–7996, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- A. P. Dawid and A. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*, 28:20–28.
- Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021a. Methods for detoxification of texts for the russian language. *Multimodal Technologies and Interaction*, 5(9).
- Daryna Dementieva, Sergey Ustyantsev, David Dale, Olga Kozlova, Nikita Semenov, Alexander Panchenko, and Varvara Logacheva. 2021b. Crowdsourcing of parallel corpora: the case of style transfer for detoxification. // *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB)*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.
- Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. *Advances in Neural Information Processing Systems*, 32.

- Zhiqiang Hu, Roy Ka-Wei Lee, Charu C Aggarwal, and Aston Zhang. 2020. Text style transfer: A review and experimental evaluation. *arXiv preprint arXiv:2010.12742*.
- Jigsaw. 2018. Toxic comment classification challenge. <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>. Accessed: 2022-03-16.
- Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification>. Accessed: 2022-03-16.
- Jigsaw. 2020. Jigsaw multilingual toxic comment classification. <https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification>. Accessed: 2022-03-16.
- Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416.
- Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! // *Proceedings of the 22nd Conference on Computational Natural Language Learning*, P 313–323, Brussels, Belgium, October. Association for Computational Linguistics.
- Nikita Konodyuk and Maria Tikhonova. 2021. Continuous prompt tuning for russian: how to learn prompts efficiently with rugpt3? // *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts*.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 737–762, Online, November. Association for Computational Linguistics.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.
- John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. // *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, P 59–69, Online, August. Association for Computational Linguistics.
- 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.
- Machel Reid and Victor Zhong. 2021. LEWIS: Levenshtein editing for unsupervised text style transfer. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 3932–3944, Online, August. Association for Computational Linguistics.
- Yulia Rubtsova. 2012. Avtomaticheskoe postroenie i analiz korpusa korotkih tekstov (postov mikroblogov) dlya zadachi razrabotki i trenirovki tonovogo klassifikatora. // *Inzheneriya znaniya i tekhnologii semanticheskogo veba, T.1*, P 109–116.
- Aleksandr Semiletov. 2020. Toxic russian comments. <https://www.kaggle.com/alexandersemiletov/toxic-russian-comments>. Accessed: 2021-07-22.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment.
- Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Haoran Xu, Sixing Lu, Zhongkai Sun, Chengyuan Ma, and Chenlei Guo. 2021. Vae based text style transfer with pivot words enhancement learning.
- Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14213–14220, May.
- Yandex. 2021. Toxic comment classification challenge. <https://yandex.ru/cup/ml/analysis/#NLP>. Accessed: 2022-03-16.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). // *Proceedings of the 13th International Workshop on Semantic Evaluation*, P 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffenseEval 2020). // *Proceedings of SemEval*.

A Labeling Pipeline Instructions

This appendix contains the illustration of all labeling tasks at Yandex Toloka platform in original Russian language: (i) detoxified paraphrase generation (Figure 3a); (ii) content preservation check (Figure 3b); (iii) toxicity classification (Figure 3c); (iv) fluency check (Figure 3d).

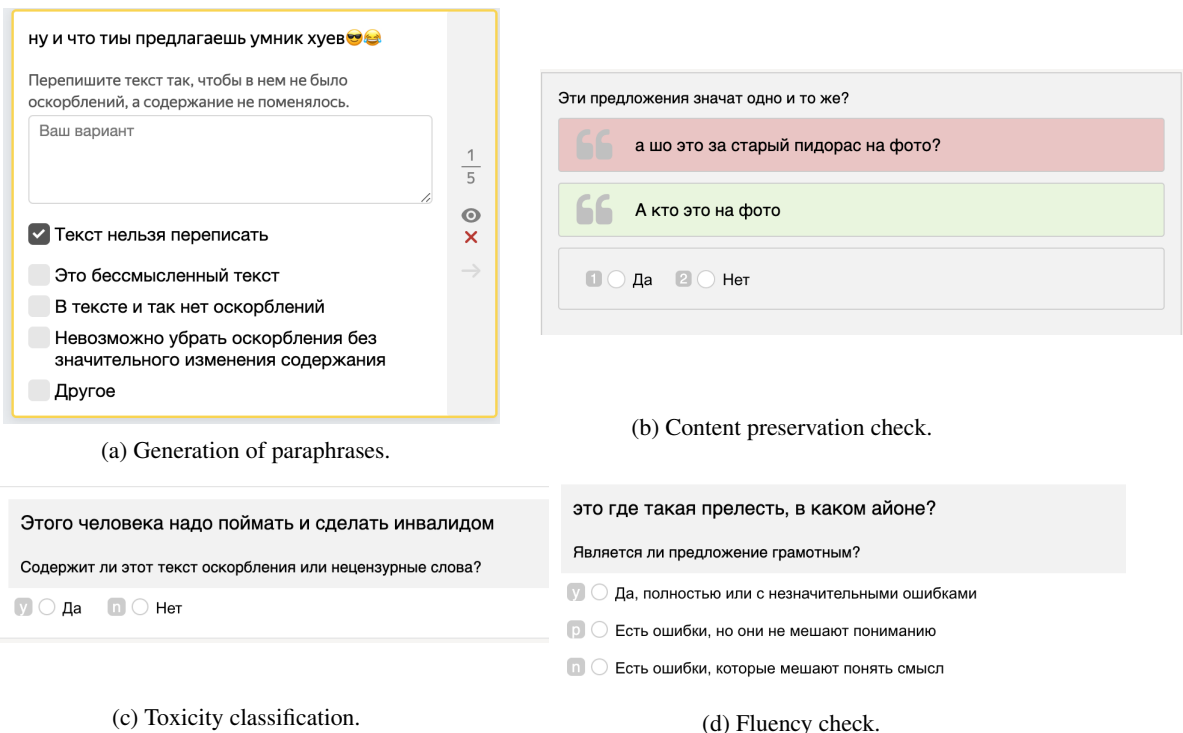


Figure 3: Original Russian interfaces in Yandex Toloka platform for labeling.

B Toxicity Labelling

This appendix contains the additional information on our ways of delivering the notion of toxicity to crowd workers. In section B.1, we provide the full instruction for the toxicity labelling task which was used to label the sentences when collecting the training data and also when evaluating the outputs of models participating in the competition. Section B.2 contains the examples which were used for user training.

B.1 Labelling Instructions

You are asked to read sentences and define if they contain offences, swear or rude words.

Attention! Non-offensive sentences can contain criticism and have a negative sentiment.

Examples

Offensive sentences:

- Интернетная шпана, не тебе меня поучать. (*You internet punks, it's not for you to lecture me*)
- Какая подписка, что ты несёшь, поехавший? (*What kind of subscription, what are you talking about, id**t?*)

- Как минимум два малолетних дегенерата в треде, мда. (*At least two juvenile degenerates in a thread, hmm.*)
- Вообще не вижу смысла спорить по поводу дюймов, х**вый там ips или нет, машина не этим интересна. (*In general, I see no reason to argue about inches, no matter if ips is sh*t or not, it's not the main thing about a car.*)

Neutral (inoffensive) sentences:

- У нас в есть убунты и текникал превью. (*We have ubuntu and technical previews.*)
- он тоже был храбрым! (*he was brave too!*)
- Это бессмысленно, ты же знаешь (*It's pointless, you know*)
- Если он мечтает съесть своих соседей то это плохо. (*If he dreams of eating his neighbors, then this is bad.*)

B.2 Training Examples

- а так реагировать на простое фото это п*здец, по моему фляга у тебя свистит п*здец как (*And such reaction to a simple photo is f*cked up, in my opinion you're retarded*)
 - Label: toxic
 - Hint: Use of obscenity (*f*cked up*) and insult (*retarded*)
- тварина *баная если это ее слова (*F*cking beast if that's her words*)
 - Label: toxic
 - Hint: Use of insult (*f*cking beast*)
- ну ты глупый (*You're stupid*)
 - Label: toxic
 - Hint: Insult of the interlocutor (*stupid*)
- Лжецы. Вот и что с такими делать? (*Liars. So what should we do to them?*)
 - Label: non-toxic
 - Hint: *Liars* is not an insult, but rather a statement of fact and not a swear word.
- очень плохой человек тот кто со своим ружьем такую птицу убивает (*A very bad person is the one who kills such a bird with his gun*)
 - Label: non-toxic
 - Hint: *Bad person* is not an insult.
- Странная, люди просто салатом делятся (*You are strange, people are just sharing their salads*)
 - Label: non-toxic
 - Hint: No offences or swear words.

C Russian Detoxification Parallel Corpus Samples

This appendix presents several samples from collected corpus for the competition to illustrate the nature of the data and purpose of the task.

Original	из-за таких п*доров мы и страдаем (<i>we suffer because of such f*ggots</i>)
Paraphrases	из-за таких плохих людей мы и страдаем (<i>we suffer because of such bad people</i>) Из-за таких людей мы и страдаем (<i>we suffer because of such people</i>) из за таких как он мы и страдаем (<i>we suffer because of people like him</i>)
Original	лучше б руководство выкинули бы на хр*н!!!! (<i>it would be better if the leaders were thrown out to h*ll!!!</i>)
Paraphrases	лучше б руководство выкинули бы (<i>it would be better if the leaders were thrown out</i>) Лучше бы руководство убрали (<i>it would be better if the leaders were removed</i>) Лучше б руководство поменяли бы!!! (<i>it would be better if the leaders were changed!!!</i>)
Original	да вообще пох*й сколько этот д*лбоеб получает (<i>I don't f*cking care how much this motherf*cker gets paid</i>)
Paraphrases	да вообще все равно сколько он получает (<i>I don't much care how much he gets paid</i>) Мне всё равно сколько он получает. (<i>I don't care how much he gets paid</i>)
Original	бл*ть, развели срачь из них*я. (<i>d*mn, they made a f*cking quarrel out of nothing</i>)
Paraphrases	Развели ссору из ничего. (<i>they made a quarrel out of nothing</i>) Развели ругань из нечего (<i>they made a fight out of nothing</i>)
Original	ты мужик ох*ел что пишешь (<i>dude, what the f*ck are you writing</i>)
Paraphrases	Зачем ты такое пишешь (<i>why are you writing this</i>) Ты мужчина обалдел что пишешь (<i>man, are you nuts, what are you writing</i>)
Original	что за п*дик в судейской коллегии???) (<i>what a f*ggot is in the judiciary???)</i>
Paraphrases	кто в судейской коллегии???) (<i>who is in the judiciary???)</i> Что за человек в судейской коллегии? (<i>what a man is in the judiciary?</i>)
Original	дэтих св*лочей надо расстрелять за жестокое обращение с животным! (<i>these b*stards should be shot for cruelty to animals</i>)
Paraphrases	Этих людей нужно наказывать за жестокое обращение с животными! (<i>these people should be punished for cruelty to animals</i>)
Original	на х*я такое выкладывать, это и дети будут смотреть д*лбоебы (<i>what the h*ll do you need to post this, the kids will watch it, motherf*ckers</i>)
Paraphrases	Зачем такое выкладывать, это и дети будут смотреть (<i>Why do you need to post this, the kids will watch it</i>)

Table 4: Examples of detoxified sentences from the collected parallel corpus.

Evidentiality and epistemic modality in the semantics of the German verbs *sollen* and *wollen* (based on the data from the German-Russian parallel corpus)

Dobrovol'skij D. O.

Russian Language Institute of the RAS;
Institute of Linguistics, Moscow, Russia
Stockholm University,
Stockholm, Sweden
dobrovolskij@gmail.com

Zalizniak Anna A.

Institute of Linguistics of the RAS;
Institute of Informatics Problems of the
Federal Research Center "Computer Science
and Control" of the RAS, Moscow, Russia
anna.zalizniak@gmail.com

Abstract

Based on the analysis of the semantics of the German modal verbs *sollen* and *wollen* and their translational equivalents identified in the German-Russian parallel subcorpus of the Russian National Corpus (RNC), the article demonstrates that the categories of evidentiality and epistemic modality should be considered as having an independent status. At the same time, these two linguistic meanings can be expressed simultaneously. We consider possible combinations of the types of evidential and epistemic meaning components expressed by these German verbs. We clarify the classification of indirect evidentiality types by introducing the third intermediate type – reportative-inferential evidentiality, i.e. a conclusion made by the speaker based on the interpretation of someone else's utterance. Addressing the parallel corpus made it possible, on the one hand, to distinguish between the types of evidential meanings expressed by the verbs *sollen* and *wollen*, on the other hand, to clarify the semantics and identify the potential polysemy of the Russian translation equivalents.

Keywords: evidentiality, epistemic modality, modal verbs, German, Russian, parallel corpus, translation

DOI: 10.28995/2075-7182-2022-21-132-140

Эвиденциальность и эпистемическая оценка в значении немецких глаголов *sollen* и *wollen* (по данным немецко-русского параллельного корпуса)

Добровольский Д. О.

Институт русского языка РАН;
Институт языкознания РАН, Москва, Россия
Стокгольмский университет,
Стокгольм, Швеция
dobrovolskij@gmail.com

Зализняк Анна А.

Институт языкознания РАН;
Федеральный исследовательский
центр «Информатика и управление»
РАН, Москва, Россия
anna.zalizniak@gmail.com

Аннотация

В статье на основе анализа семантики немецких модальных глаголов *sollen* и *wollen* и их переводных эквивалентов, выявленных на материале немецко-русского параллельного подкорпуса НКРЯ, демонстрируется, что категории эвиденциальности и эпистемической модальности и должны рассматриваться как имеющие независимый статус; при этом эти два языковых значения могут выражаться одновременно. Рассматриваются возможные комбинации типов эвиденциального и эпистемического компонентов значения, выражаемые данными немецкими глаголами. Предлагается уточнить классификацию типов косвенной эвиденциальности за счет введения третьего промежуточного типа – репортативно-инференциальной эвиденциальности (= вывод, сделанный говорящим на основании интерпретации чужого высказывания). Обращение к параллельному корпусу позволило, с одной стороны, разграничить типы эвиденциальных значений, выражаемых рассматриваемыми немецкими глаголами, с другой – уточнить семантику и выявить потенциальную многозначность единиц русского языка, выступающих в роли их переводных эквивалентов.

Ключевые слова: эвиденциальность, эпистемическая модальность, модальные глаголы, немецкий язык, русский язык, параллельный корпус, перевод

1 Эвиденциальность vs. эпистемическая модальность

Вопросу о соотношении категорий эвиденциальности и эпистемической модальности посвящена огромная литература (см. в частности [van der Auwera, Plungian 1998; Mortelmans 2000b; de Haan 2005; Wiemer 2008; 2018; Cornillie 2009; Leiss 2009; Diewald, Smirnova 2010; 2011; Socka 2011; 2013; Aikhenvald 2004; 2015; Wiemer, Socka 2017; Forker 2018; Guentcheva (ed.) 2018 и др.]); тем не менее, этот вопрос нельзя считать решенным. В нашем исследовании предпринята попытка прояснения этого соотношения на основании анализа переводных соответствий немецких модальных глаголов *sollen* и *wollen*, реально использованных профессиональными переводчиками; материалом исследования послужил немецко-русский параллельный подкорпус НКРЯ (www.ruscorpora.ru).

Под **эвиденциальностью** понимается указание на источник, на который опирается данное высказывание; при этом принято различать прямую и косвенную эвиденциальность. Нас здесь будет интересовать только косвенная эвиденциальность, внутри которой в литературе принято противопоставление *reported evidentiality vs. inferential evidentiality* (см. [Aikhenvald 2004, Plungian 2011: 338-370]). Как мы полагаем, в сфере косвенной эвиденциальности следует различать не две, а три категории:

- 1) говорящий воспроизводит чье-либо высказывание P – репортатив (“reported evidentiality” или “hearsay evidentiality”) [Plungian 2011: 464; Horvath 2013];
- 2) говорящий вывел P исключительно на основании какого-то конкретного чужого высказывания Q – репортативно-инференциальная эвиденциальность;
- 3) говорящий вывел P на основании любых других данных или комплекса данных – собственно инференциальная эвиденциальность.

Поскольку значение 3-го типа глаголами *sollen* и *wollen* не выражается, эту категорию мы здесь не рассматриваем.

Эпистемическая модальность – это оценка говорящим вероятности того, что некоторая ситуация имеет место (ср. [Bybee et al. 1994:179; Nuys 2001: 21; Cornillie 2009: 46] и др.). Эпистемическим называют такое значение пропозиционального оператора P(S), в котором центральным ассертивным компонентом является оценка говорящим вероятности того, что ситуация S имеет/имело/будет иметь место (ср. *Боюсь/похоже/судя по всему, из этой затеи ничего не вышло/не выйдет*). Оценка вероятности находится в диапазоне от нуля до единицы. Значение эпистемической возможности возникает при оценке вероятности как большей нуля, эпистемической необходимости – как приближающейся к единице.

При этом следует сразу отметить, что эпистемическое значение модальных предикатов легче идентифицируется в контексте конкретных ситуаций, относящихся к прошлому или настоящему, ср. *Он должен сейчас быть дома; Он мог потерять ключи*. Когда речь идет о ситуациях, отнесенных к будущему, или об общих суждениях, в значении модального предиката эпистемическая составляющая накладывается на значение «объективной» (“subject-oriented”) модальности. Фразы типа *Он должен вернуться завтра; Он может потерять ключи* наряду с оценкой вероятности включают компонент объективной модальности – деонтической (должен вернуться завтра, чтобы успеть на день рождения матери – я предполагаю, что вернется, поскольку знаю, что он человек долга) или онтологической, т.е. утверждение о свойствах субъекта (может потерять ключи, потому что очень рассеянный, поэтому я оцениваю данное событие как имеющее некоторую вероятность), и т.п.

Если говорящий, сообщая о некотором факте, специально отмечает, что не является его свидетелем, а лишь передает полученную из какого-то источника информацию, он одновременно указывает на отсутствие у него надежных доказательств истинности данного положения вещей (ср. [Diewald 1999: 226]). В этом случае оценка вероятности определяется степенью доверия к источнику информации (ср. [Wiemer 2008: 6]). В статье [Wiemer 2018] обсуждаются возможные способы описания соотношения эпистемического и эвиденциального значений: 1) эпистемическое и эвиденциальное значения существуют независимо, хотя могут как-то влиять друг на друга; 2) эти значения пересекаются; 3) одно входит в другое как его подтип. Самой непротиворечивой, как считает Б. Вимер, оказывается модель, согласно которой эти два значения имеют независимый статус (см. также аргументы в пользу данной точки зрения в [Cornillie 2009]). Наш материал подтверждает эту точку зрения. В том случае, когда объектом эпистемической оценки является

содержание чужого высказывания, эпистемическое и эвиденциальное значения выражаются одновременно. Ниже мы рассмотрим возможные типы комбинаций этих двух значений.

Эпистемическое значение, которое накладывается на эвиденциальное, может быть трех типов:

1) «нулевая» эпистемическая оценка – говорящий не выражает никакого мнения относительно истинности передаваемой информации; в этом случае эпистемическое значение отсутствует;

2) «положительная» эпистемическая оценка – говорящий склоняется к тому, чтобы считать эту информацию истинной;

3) «отрицательная» эпистемическая оценка – говорящий сомневается в истинности данной информации или даже склоняется к тому, чтобы считать эту информацию ложной.

Между первым и вторым случаем нет жесткой границы, что обусловлено прагматически: если говорящий передает полученную от кого-то информацию, эксплицитно не выражая недоверия к ее источнику, это означает, что он до какой-то степени ему доверяет.

Отметим, что когда говорящий передает чье-либо общее суждение, по отношению к которому оценка вероятности невозможна, речь идет лишь о согласии или несогласии говорящего с этим суждением. Такого рода примеры мы из рассмотрения исключаем; ср. (1):

- (1) Frankreich *sollte* ein Staat dritten Ranges werden, riesige Stücke Rußlands müßten annektiert werden, und alle Gegner hätten zu zahlen und Sachwerte abzuliefern bis zum Weißbluten! Das warst du, Heinrich! [Erich Maria Remarque. Der schwarze Obelisk (1956)]

Пусть Франция, *дескать*, станет третьестепенной державой, пусть у России аннексируют громадные территории, пусть все противники платят контрибуцию и отдают реальные ценности, пока их совсем не обескровят! И это говорил ты, Генрих! [Эрих Мария Ремарк. Чёрный обелиск (В. Станевич, 1961)]

2 Немецкие глаголы *sollen* и *wollen*

Значение немецких модальных глаголов в целом достаточно подробно изучено и описано – как в грамматиках и словарях, так и в теоретических работах. Тем не менее, исследование семантики этих глаголов на основе анализа их переводов на русский язык (возможность которого открылась после появления параллельных корпусов текстов) позволяет существенным образом уточнить описание их значений. Мы рассмотрим два немецких модальных глагола, которые могут выражать эвиденциальное значение – *sollen* (с основным значением ‘должен’) и *wollen* (с основным значением ‘хотеть’) – с целью прояснить вопрос о наличии/статусе эпистемического компонента в значении этих глаголов.

2.1 Глагол *sollen*

Вопрос о статусе эпистемического компонента в значении глагола *sollen* широко обсуждается в литературе (см. [Öhlschläger 1989; Diewald 1999; Mortelmans 2000b; Wiemer 2018; Socka 2011] и др.); общепринятым является мнение об отсутствии у *sollen* собственно эпистемического значения. В [van der Auwera, Plungian 1998: 85] *sollen* рассматривается как “hearsay evidential”; например, фраза *Es soll ein guter Film sein* означает ‘говорят, что это хороший фильм’, а не ‘это, должно быть, хороший фильм’; ср. также: *sie soll sehr schön gewesen sein* – *говорят, она была очень красива*; *er soll in Moskau sein* – *я слышал, что он в Москве*; *diese Kräuter sollen gut gegen Grippe sein* – *считается, что эти травы хорошо помогают от гриппа*.

При этом отмечается, что эвиденциальное значение глагола *sollen* может включать «эпистемический оттенок» (“epistemic overtone”, см. [Diewald 1999: 17, 278-283; Socka 2013: 163; Wiemer, Socka 2017; Wiemer 2018]): передавая чужое высказывание, говорящий использует его как основание для собственной эпистемической оценки сообщаемого факта. В статье [Mortelmans 2000b] глагол *sollen* характеризуется как “strongly evidential, weakly epistemic”.

Перечислим комбинации типов эвиденциального значения и типов эпистемической оценки, возможные для глагола *sollen*.

1. Эвиденциальное значение – репортатив, т.е. говорящий передает чужое высказывание.

1+1: репортатив + нулевая эпистемическая оценка. Говорящий передает полученное из какого-то источника сообщение о некотором факте, полностью снимая с себя ответственность за него, т.е. говорящий не выражает никакого своего отношения к его истинности; в русском переводе используются слова: *говорят, по слухам*; а также *будто бы, вроде бы, якобы*¹ и др. Примеры²:

- (2) Übrigens haben wir heute viel über diesen Rainer Engstern gesprochen, er **soll mal was mit einer Schülerin gehabt haben**. [Ingrid Noll. Der Hahn ist tot (1991)]
А вообще мы сегодня долго обсуждали этого Райнера Энгштерна. **Говорят**, у него что-то было с одной ученицей. [Ингрид Нолль. Мертвый петух (О. А. Соколова, 2004)]
- (3) Beizmenne **soll** die aufreizend gelassen an ihrer Anrichte lehrende Katharina nämlich *gefragt haben*: „Hat er dich denn gefickt“, woraufhin Katharina sowohl rot geworden sein wie in stolzem Triumph *gesagt haben soll*: „Nein, ich würde es nicht so nennen.“ [Heinrich Böll. Die verlorene Ehre der Katharina Blum (1974)]
Байцменне **якобы спросил** вызывающе спокойно прислонившуюся к своему серванту Катарину: «А он тебя употребил?», на что Катарина, покраснев, но с гордым торжеством, **будто бы ответила**: «Нет, так я бы это не назвала». [Генрих Бёлль. Потерянная честь Катарины Блум (Е. А. Кацева, 1988)]
- (4) Die Frau **soll** nußgroße Smaragde in ihr Korsett *eingenäht gehabt haben*. [E. M. Remarque. Der Himmel kennt keine Günstlinge (1961)]
Она, **по слухам**, **зашила** в корсет несколько изумрудов, каждый с орех величиной. [Эрих Мария Ремарк. Жизнь взаимы, или У неба любимчиков нет (М. Л. Рудницкий, 2018)]
- (5) Italiener waren aber auch keine da, für sie war es zu früh im Jahr, sie würden erst im Juni zur Jasminernte ins Land kommen, sie konnten's also nicht gewesen sein. Schließlich gerieten die Perückenmacher in Verdacht, bei denen man nach dem Haar des ermordeten Mädchens fahndete. Vergeblich. Dann **sollten es die Juden gewesen sein**, dann die angeblich geilen Mönche des Benediktinerklosters [Patrick Süskind. Das Parfum: Die Geschichte eines Mörders (1985)]
Итальянцев, правда, тоже не было, так рано они не появлялись, они придут заниматься на сбор жасмина только в июне, значит, это не могут быть итальянцы. Наконец, под подозрение попали парикмахеры, которых обыскали, пытаюсь обнаружить волосы убитой де-вухи. Тщетно. Потом **заподозрили евреев**, потом якобы похотливых монахов бенедиктинского монастыря [Патрик Зюскинд. Парфюмер: История одного убийцы (Э. Венгерова, 1992)]

Использованное в переводе примера (5) слово *заподозрили* указывает на чье-то предположение (очевидно, высказанное вслух), возникшее после того, как другие предположения отпали. Субъектом мнения ‘возможно Р’, составляющего ассертивный компонент глагола *заподозрить*, является некоторое третье лицо; говорящий остается нейтрален по отношению к истинности Р.

1+2: репортатив + положительная эпистемическая оценка. В этом случае в переводе используется вводное *говорят*; возможно также *по словам*.

- (6) Ich hoffe, sie ist tot. Sie **soll** meine Mutter *vergiftet haben*. [Cornelia Funke. Tintentod (2007)]
Надеюсь, она погибла. **Говорят**, это она *отравила* мою мать. [Корнелия Функе. Чернильная смерть (М. М. Сокольская, 2012)]

¹ О слове *якобы* в функции нейтральной передачи содержания чужой речи. см. [Апресьян, Шмелев 2017].

² Здесь и далее в примерах *п/ж курсивом* выделен немецкий модальный глагол и его эквивалент в русском переводе; *светлым курсивом* выделен подчиненный модальному глаголу инфинитив (при необходимости – с зависимыми словами) и его эквивалент.

- (7) <...> einen Brief geschrieben, in dem sie sich äußerst entzückt über den Bräutigam ausläßt. Es *soll ein bildschöner Mann sein und von vornehmem Wesen*. Wie glücklich sie sein muß!
[Thomas Mann. Buddenbrooks (1896-1900)]
<...> письмо, в восторженных выражениях рассказывающее о ее женихе. *По словам* Армгард, *он писанный красавец и прекрасный человек*. Как она, верно, счастлива! [Томас Манн. Будденброки (Н. Ман, 1953)]

В примере (6) говорящий оценивает вероятность истинности передаваемого утверждения как достаточно высокую: на этой оценке основана высказанная говорящим надежда («что она погибла»); в (7) фраза «Как она счастлива!» указывает на то, что говорящий оценивает передаваемую информацию как истинную.

1+3: репортатив + отрицательная эпистемическая оценка. В примере (8) слова *дескать*, *вроде бы*, *будто бы* и *мол* – все являются показателями чужого утверждения, с которым говорящий не согласен:

- (8) Jetzt lesen sie aufwieglerische Bücher von Hugenotten oder Engländern. Oder sie schreiben Traktate oder sogenannte wissenschaftliche Großwerke, in denen sie alles und jedes in Frage stellen. *Nichts mehr soll stimmen*, alles soll jetzt plötzlich anders sein. In einem Glas Wassers *sollen* neuerdings ganz kleine Tierchen *schwimmen*, die man früher nicht gesehen hat; die Syphilis *soll eine ganz normale Krankheit sein* und keine Strafe Gottes mehr; Gott *soll* die Welt nicht an sieben Tagen *erschaffen haben*, sondern in Jahrmillionen, wenn er es überhaupt war; [Patrick Süskind. Das Parfum: Die Geschichte eines Mörders (1985)]
Теперь они читают подстрекательские книги гугенотов или англичан. Или пишут трактаты, или так называемые великие научные сочинения, в коих все и вся ставится под вопрос. *Будто бы нет больше ничего достоверного*, и все вдруг изменилось. В стакане воды, *дескать*, *плавают* малюсенькие зверушки, которых раньше никто не видел; сифилис теперь *вроде бы нормальная болезнь*, а не божья кара; Господь, *мол*, *создал* мир не за семь дней, а за миллионы лет, если это вообще был Господь [Патрик Зюскинд. Парфюмер: История одного убийцы (Э. Венгерова, 1992)]

2. В том случае, когда говорящий не просто передает чужое высказывание, а делает из него некоторый вывод, эксплицируя его скрытые смыслы – мы обозначили этот тип эвиденциальности как «репортативно-инференциальный» – говорящий может быть в разной степени уверен в том, что этот вывод сделан правильно; однако он не может ни воздержаться от эпистемической оценки этого вывода, ни считать его неверным. Т.е. здесь возможна лишь комбинация 2+2, ср.:

- (9) „Er ist eben müde geworden. Vielleicht hat ihn der Dieb sogar hypnotisiert. Das *soll vorkommen*“, meinte der Herr. [Erich Kästner. Emil und die Detektive (1929)]
Видно, он устал. Может, даже вор его загипнотизировал. Такие вещи, *говорят*, *бывают*, – сказал господин с газетой. [Эрих Кестнер. Эмиль и сыщики (Л. Лунгина, 1971)]

То, что «такие вещи бывают» – это вывод, который делает говорящий на основании каких-то высказываний третьих лиц (скорее всего, касающихся конкретных фактов подобного рода).

При этом, не сомневаясь в правильности сделанного им вывода из чужих слов, говорящий может быть с ним категорически несогласен по существу, ср.:

- (10) Ich habe ja nicht im entferntesten daran gedacht, Sie oder irgend jemand zu kränken. Sie kennen mich doch schon lange genug, Herr K., um davon überzeugt sein zu können. Sie wissen gar nicht, wie ich die letzten Tage gelitten habe! **Ich sollte meine Mieter verleumden!** [Franz Kafka. Der Prozess (1914)]
 У меня и в мыслях не было обидеть вас или еще кого-нибудь. Ведь вы меня так давно знаете, господин К., вы мне должны поверить. Вы не можете себе представить, как я страдала все эти дни! **Неужели я способна оговорить своих квартирантов!** [Франц Кафка. Процесс (Р. Райт-Ковалева, 1965)]

Пример (10) буквально означает «из того, что они утверждают, следует, что я оговорила своих квартирантов; утверждение ложное (и обидное)». Для реализации данного значения требуется специфический контекст (в частности вопросительное или восклицательное предложение), ср. [Mortelmans 2000b: 136].

2.2 Глагол *wollen*

Главная особенность эвиденциального значения *wollen* состоит в том, что автором воспроизводимого высказывания является сам субъект этого глагола, в отличие от *sollen*, где передается высказывание третьих лиц (ср. [Horvath 2013; Socka 2013: 163; Mortelmans 2000a]). Так, фраза: *Er will es nicht getan haben* означает «он утверждает, что он этого не делал» (*он* – источник информации и субъект действия, обозначенного инфинитивом), а *Er soll sehr krank sein* – «говорят, что он очень больной человек» (*он* – субъект обозначенного инфинитивом состояния, а источник информации – некоторые люди, но только не *он* сам).

Существует две точки зрения: что *wollen* выражает сомнение говорящего в истинности пропозиции (так пишут многие словари в том числе Дуден, см. <https://www.duden.de>; ср. также [Mortelmans, Vanderbiesen 2011]) – и, наоборот, что этот глагол нейтрален относительно оценки истинности вводимого им утверждения [Öhlschläger 1989: 235; Diewald 1999: 225]. Наш материал позволяет разрешить этот спор следующим образом.

В целом для глагола *wollen* характерен «чистый» репортатив, то есть непосредственная передача высказывания другого лица. Что касается эпистемической оценки, она в принципе может быть всех трех типов.

1+1: репортатив + нулевая эпистемическая оценка. Конструкция с глаголом *wollen* в этом значении широко распространена в судебной практике³: любое показание, истинность которого не доказана или не опровергнута, вводится этим глаголом. В примерах такого рода нулевая эпистемическая оценка является как бы условием соблюдения юридических принципов, ср.:

- (11) Patricks Schwester Julia ist verzweifelt. Sie hat einen Mann überfahren. Bei der Obduktion stellt sich heraus, dass der Mann bereits tot war. Er wurde vergiftet und auf die Straße gelegt. Die Bewohner der umliegenden Häuser **wollen nichts bemerkt haben** <...>. (<https://www.fernsehserien.de/soko-leipzig/episodenguide/staffel-8/8561>)
 Сестра Патрика Юлия в отчаянии. Она сбила человека. При вскрытии обнаружилось, что в тот момент он уже был мертв. Он был отравлен и выброшен на улицу. Соседи из ближних домов **говорят, что ничего не видели.**

1+2: репортатив + положительная эпистемическая оценка. Такая комбинация в принципе маловероятна, поскольку если говорящий считает, что сказанное другим человеком скорее всего соответствует действительности, у него нет необходимости вводить его глаголом *wollen*, который подчеркивает заинтересованность – и тем самым возможную необъективность – источника информации. Тем не менее, такие примеры встречаются; ср. (12), где эквивалент для эвиденциального *wollen* в переводе опущен, что свидетельствует об отсутствии у говорящего оснований для сомнения в истинности передаваемой информации:

³ Ср. [Hingson 2021] о роли высказываний категории “hearsay” в судопроизводстве.

- (12) „Denn wenn du *allein mit ihm im Zug gesessen haben willst*, hast du doch keinen einzigen Zeugen“, bemerkte einer der Angestellten. [Erich Kästner. *Emil und die Detektive* (1929)]
– Раз ты *был с ним вдвоем в купе*, значит, у тебя нет свидетелей, – объяснил один из служащих. [Эрих Кестнер. *Эмиль и сыщики* (Л. Лунгина, 1971)]

1+3: репортажив + отрицательная эпистемическая оценка. Такое «скептическое» (ср. „[d]ie skeptische Lesart des reportiven *wollen* [Mortelmans, Vanderbiesen 2011: 86]) эвиденциальное употребление для *wollen* наиболее характерно, и тому имеется как минимум две причины. Первая состоит в том, что эвиденциальное *wollen* сохраняет связь с исходным значением, т.е. буквально означает «X *хочет*, чтобы мы думали, что P» – что естественно порождает сомнения в объективности этой информации. Вторая причина состоит в следующем. В высказываниях с эвиденциальным *sollen* автор передаваемого утверждения (reported speaker) «уходит в тень»⁴, и на первый план выходит само высказывание; сам тот факт, что говорящий его воспроизводит, уже свидетельствует о том, что он в какой-то степени этому утверждению доверяет. А в случае с эвиденциальным *wollen* reported speaker полноценно «присутствует на лингвистической сцене», являясь контрагентом, утверждению которого говорящий может либо доверять, либо не доверять, либо относиться к нему нейтрально – ср. [Mortelmans 2000b: 139]. Поэтому в тех случаях, когда описывается очевидно нереальная ситуация, используется именно глагол *wollen*; ср.:

- (13) Während die Äbtissin, die mein volles Vertrauen genießt und eine Frau von gesundem Menschenverstand zu sein scheint, eine natürliche Erklärung für das Phänomen nicht ausschließt, kommt der Verdacht der Hexerei ausgerechnet vom Vater des Mädchens...
Mit eigenen Augen will er gesehen haben, wie der Teufel in Gestalt eines jungen Mannes das Mädchen im Garten umarmt und sich dann in einer Rauchwolke entmaterialisiert hat und einen leichten Schwefelgeruch hinterließ. [Kerstin Gier. *Saphirblau* (2009)]
Аббатиса монастыря, пользующаяся моим полным доверием, разумная женщина, не исключает естественное объяснение феномена; подозрения же в ведьмовстве высказал отец девушки. Он *уверяет*, что *видел собственными глазами*, как дьявол в виде юноши обнимал девушку в саду, после чего растворился в облаке дыма и оставил после себя слабый запах серы. [Керстин Гир. Сапфировая книга (С. Вольштейн, 2013)]

Встречаются также случаи репортаивно-инференциальной эвиденциальности, когда заключение P делается на основе чьего-то высказывания Q, ср. *Er will es nicht wahr haben* «он не хочет признать, что это правда»: если кто-то не хочет признавать, что это правда, это не значит, что он сказал «это неправда»; то, что он не хочет это признавать – это скорее вывод, сделанный говорящим из каких-то других высказываний обсуждаемого лица. Так, в примере (14) мы имеем дело не с «чистым» репортаивом, т.е. фразу *Er wollte mit ihr unbedingt was gehabt haben* не обязательно понимать в смысле «он говорил: между нами что-то было». Это скорее вывод, который можно – довольно однозначно – сделать из конкретных высказываний персонажа. Эпистемическая оценка говорящего здесь – нулевая.

- (14) *Er wollte mit ihr unbedingt was gehabt haben*: „Mit Tulla, das kannst du ruhig glauben. Sogar bei ihr zu Hause, Elsenstraße. Ihre Mutter guckt weg.“ [Günter Grass. *Katz und Maus* (1961)]
Но он *настаивал, мол, что-то между ними было*. – Насчет Туллы можешь мне поверить. Даже у нее дома на Эльзенштрассе. Мать на это смотрит сквозь пальцы. [Гюнтер Грасс. *Кошки-мышки* (Наталия Ман, 1985)]

⁴ Согласно корпусному исследованию [Mortelmans 2000b] в 60% случаев reported speaker эвиденциального *sollen* является неопределенным. Поэтому не случайно, что самый частотный русский эквивалент эвиденциального *sollen* – безличное *говорят*. А в качестве эквивалента для эвиденциального *wollen* безличное *говорят* вообще не может использоваться.

3 Заключение

Результаты нашей работы могут быть резюмированы следующим образом.

Во-первых, мы уточнили классификацию типов эвиденциальности, введя новый тип косвенной эвиденциальности – «репортативно-инференциальную эвиденциальность». Тем самым, в сфере косвенной эвиденциальности следует выделять не два типа (репортативную и инференциальную эвиденциальность), а три, поскольку между репортативной и инференциальной эвиденциальностью есть еще промежуточный тип, который содержит элементы как репортативности, так и инференциальности. В пользу необходимости выделения этого типа говорит в частности тот факт, что немецкие глаголы *sollen* и *wollen* могут выражать репортативную и репортативно-инференциальную эвиденциальность, а собственно инференциальную эвиденциальность, т. е. вывод, основанный на данных, отличных от высказываний других лиц, ни глагол *sollen*, ни тем более *wollen*, выражать не могут.

Во-вторых, мы показали, что категории эвиденциальности и эпистемической модальности должны рассматриваться как имеющие независимый статус; при этом эти два языковых значения могут выражаться одновременно. Были рассмотрены возможные комбинации типов эвиденциального и эпистемического компонентов значения, выражаемые немецкими глаголами *sollen* и *wollen*.

В-третьих, обращение к параллельному корпусу позволило сопоставить специфику каждого из типов эвиденциального значения *sollen* и *wollen*, выделенных на основе семантического анализа фрагментов оригинального текста, с наборами их русских эквивалентов, реально использованных профессиональными переводчиками. Как показал проведенный анализ, наряду с единицами, которые могут появляться в роли эквивалентов для обоих глаголов (ср. *якобы*, *будто бы*, *вроде бы* и др.), имеются и такие, которые достаточно жестко привязаны лишь к одному из них. Так, безличное *говорят*, а также показатели эпистемической оценки *наверное*, *возможно*, *вероятно* могут быть эквивалентами только для *sollen*; глагол *говорить* в личной форме, а также *уверять*, *настаивать*, *утверждать*, *делать вид* и др. – только для *wollen*. Кроме того, был выявлен ряд специфических средств передачи эвиденциального значения того и другого глагола (ср. глаголы *заподозрить*, *послышаться* и др. для *sollen*; *божиться*, дискурсивное слово *правда-правда* и др. для *wollen*), появление которых обусловлено конкретными семантическими и прагматическими условиями.

Благодарности

Авторы благодарят анонимных рецензентов за конструктивную критику.

References

- [1] Aikhenvald Alexandra Y. Evidentiality. — Oxford: Oxford University Press, 2004.
- [2] Aikhenvald Alexandra Y. Evidentials: Their links with other grammatical categories // Linguistic Typology. — 2015. Access mode: <https://www.researchgate.net/publication/282593953>
- [3] Apresyan Valentina Y., Shmelev Alexei D. “Xeno” markers in the light of the data of parallel corpora and contemporary mass media: the case of the Russian word *jakoby* [“Ksenopokazateli” po dannym parallel'nykh korpusov i sovremennykh SMI: russkoye *jakoby*], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2017” [Komp'yuternaya Lingvistika i Intel'lektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2017”], Moscow: RGGU, 2017. — P. 16–26.
- [4] Bybee Joan, Perkins Revere, Pagliuca William. The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World. — Chicago: University of Chicago Press, 1994.
- [5] Cornillie Bert. Evidentiality and epistemic modality. On the close relationship between two different categories. // Functions of Language 16:1 (2009). — P. 44–62.
- [6] de Haan Ferdinand. Encoding speaker perspective: Evidentials // Frajzyngier Zygmunt, Hodges Adam, Rood, David S. (eds.), Linguistic Diversity and Language Theories. — Amsterdam, Philadelphia: Benjamins, 2005. — P. 379–397.
- [7] Diewald Gabriele. Die Modalverben im Deutschen: Grammatikalisierung und Polyfunktionalität. — Tübingen: Max Niemeyer Verlag, 1999.
- [8] Diewald Gabriele, Smirnova Elena. Evidentiality in German. — Berlin, Boston: De Gruyter Mouton, 2010.

- [9] Diewald Gabriele, Smirnova Elena. Indirekte Rede zwischen Modus, Modalität und Evidentialität // Diewald Gabriele, Smirnova Elena (eds.). *Modalität und Evidentialität / Modality and Evidentiality*. — Trier: Wissenschaftlicher Verlag Trier, 2011. — P. 89–108.
- [10] Forker Diana. Evidentiality and its relations with other verbal categories // Aikhenvald Alexandra Y., (ed.). *The Oxford Handbook of Evidentiality*. — Oxford: Oxford University Press, 2018. — P. 65–84.
- [11] Guentcheva Zlatka (ed.). *Epistemic Modalities and Evidentiality in Cross-linguistic Perspective*. — Berlin: De Gruyter-Mouton, 2018.
- [12] Hingson LaReina. A Preliminary Linguistic Analysis of Hearsay. *Academia Letters*, Article 523. — 2021. Access mode: <https://doi.org/10.20935/AL523>
- [13] Horvath Katalin. Sind reportative Modalverben epistemisch? // Abraham Werner, Leiss Elisabeth (eds.). *Funktionen von Modalität*. — Berlin, Boston: De Gruyter, 2013. — P. 131–155.
- [14] Leiss Elisabeth. Drei Spielarten der Epistemizität, drei Spielarten der Evidentialität und drei Spielarten des Wissens // Abraham Werner, Leiss Elisabeth (eds.). *Modalität: Epistemik und Evidentialität bei Modalverb, Adverb, Modalpartikel und Modus*. — Tübingen: Stauffenburg Verlag, 2009. — P. 3–24.
- [15] Mortelmans Tanja. Konjunktiv II and epistemic modals in German: a division of labour // Foolen Ad, van der Leek Frederike (eds.). *Constructions in Cognitive Linguistics Selected Papers from the Fifth International Cognitive Linguistics Conference, Amsterdam, 1997*. — Amsterdam, Philadelphia: Benjamins, 2000a. — P. 191–215.
- [16] Mortelmans Tanja. On the ‘evidential’ nature of the ‘epistemic’ use of the German modals *müssen* and *sollen* // van der Auwera Johan, Dendale Patrick (eds). *Modal Verbs in Germanic and Romance Languages = Belgian Journal of Linguistics*, 14, 2000b. — P. 131–148.
- [17] Mortelmans Tanja, Vanderbiesen Jeroen. *Dies will ein Parlamentarier „aus zuverlässiger Quelle“ erfahren haben*. Reportives *wollen* zwischen *sollen* und dem Konjunktiv I der indirekten Rede // Diewald Gabriele, Smirnova Elena (eds.). *Modalität und Evidentialität / Modality and Evidentiality*. — Trier: Wissenschaftlicher Verlag Trier, 2011. — P. 69–88.
- [18] Nuyts Jan. *Epistemic Modality, Language, and Conceptualization: A Cognitive-Pragmatic Perspective*. — Amsterdam: Benjamins, 2001.
- [19] Öhlschläger Günther. *Zur Syntax und Semantik der Modalverben des Deutschen*. — Tübingen: Niemeyer, 1989.
- [20] Plungian Vladimir. *Introduction to Grammatical Semantics: Grammatical Meanings and Grammatical Systems of the Languages of the World [Vvedeniye v grammaticheskuyu semantiku: grammaticheskiye znacheniya i grammaticheskiye sistemy yazykov mira]*. — Moscow: RGGU, 2011.
- [21] Socka Anna. Evidentialität und Epistemizität in der Bedeutung reportativer Satzadverbien im Polnischen und Deutschen // Diewald Gabriele, Smirnova Elena (eds.). *Modalität und Evidentialität / Modality and Evidentiality*. — Trier: Wissenschaftlicher Verlag Trier, 2011. — P. 49–68.
- [22] Socka Anna. Skopus reportativer Ausdrücke in Komplementsätzen im Deutschen und Polnischen // Abraham Werner, Leiss Elisabeth (eds.). *Funktionen von Modalität*. — Berlin, Boston: De Gruyter, 2013. — P. 157–185.
- [23] van der Auwera Johan, Plungian Vladimir. Modality’s semantic map // *Linguistic Typology*, 2, 1998. — P. 79–124.
- [24] Wiemer Björn. Lexikalische Markierungen evidenzieller Funktionen: zur Theoriebildung und empirischen Erforschung im Slavischen // Wiemer Björn (ed.). *Lexikalische Evidenzialitäts-Marker in slavischen Sprachen = Wiener Slawistischer Almanach, Sonderband 72*, 2008. — P. 5–49.
- [25] Wiemer Björn. Evidentials and Epistemic Modality // Aikhenvald Alexandra Y., (ed.). *The Oxford Handbook of Evidentiality*. — Oxford: Oxford University Press, 2018. — P. 85–108.
- [26] Wiemer Björn, Socka Anna. How much does pragmatics help to contrast the meaning of hearsay adverbs? (Part 1) // *Studies in Polish Linguistics*. Vol. 12, issue 1, 2017. — P. 27–56. Access mode: <https://www.academia.edu/34803999>

RuCoCo: a new Russian corpus with coreference annotation

Vladimir Dobrovolskii
ABBY

v.dobrovolskii@abbyy.com

Mariia Michurina
MIPT, RSUH

Moscow, Russia

marimitchurina@gmail.com

Alexandra Ivoylova

MIPT, RSUH

Moscow, Russia

a.m.ivoylova@gmail.com

Abstract

We present a new corpus with coreference annotation, Russian Coreference Corpus (RuCoCo). The goal of RuCoCo is to obtain a large number of annotated texts while maintaining high inter-annotator agreement. RuCoCo contains news texts in Russian, part of which were annotated from scratch, and for the rest the machine-generated annotations were refined by human annotators. The size of our corpus is one million words and around 150,000 mentions. We make the corpus publicly available¹.

Keywords: coreference corpus, coreference resolution, anaphora resolution, corpus annotation, Russian language

DOI: 10.28995/2075-7182-2022-21-141-149

RuCoCo: новый русскоязычный корпус кореференции

Добровольский В.А.
ABBY

v.dobrovolskii@abbyy.com

Мичурина М.А.
МФТИ, РГГУ

Москва, Россия

marimitchurina@gmail.com

Ивойлова А.М.

МФТИ, РГГУ

Москва, Россия

a.m.ivoylova@gmail.com

Аннотация

В этой статье мы представляем новый корпус кореференции для русского языка RuCoCo. Цель корпуса RuCoCo - получить большое количество размеченных текстов и одновременно с этим добиться высокого уровня согласия между аннотаторами. RuCoCo состоит из текстов новостей на русском языке, часть из которых была аннотирована с нуля, а для остальных текстов была выполнена машинная разметка и доработана аннотаторами-носителями языка. Размер нашего корпуса составляет один миллион слов и около 150 000 упоминаний. Корпус находится в открытом доступе.

Ключевые слова: корпус кореференции, разрешение кореференции, разрешение анафоры, создание корпуса, русский язык

1 Introduction

The task of coreference resolution was introduced at the Sixth Message Understanding Conference (Grishman and Sundheim, 1996), where the first dataset for coreference resolution task was introduced. The dataset consisted of 25 articles from Wall Street Journal (30,000 words). The annotation scheme was considered a standard until the release of ACE 2005 Multilingual Training Corpus for the 2005

¹<https://github.com/vdobrovolskii/rucoco>

Automatic Content Extraction (ACE) technology evaluation (Doddington et al., 2004). The corpus included texts in English, Chinese and Arabic and contained around 650,000 words in total for the three languages.

The MUC guidelines were domain-oriented, and their definition of a *markable* (mention) was mostly syntactically motivated. But further developments in this area, starting with the ACE initiative, increasingly involved semantic factors, so that recent corpora with coreference annotation define markables based on semantic class restrictions.

Quite a lot of such corpora were created in the last two decades, their primary goals being to increase the size in order to satisfy the requirements of the data-driven approach and to improve inter-annotator agreement which in many cases is too low, especially when a dataset addresses more complex cases of coreference.

The most well-known corpus of this kind is OntoNotes 5.0 (Pradhan et al., 2013). OntoNotes contains texts of various genres in three languages: English, Arabic, and Chinese. The cumulative volume of this corpus is 2.9 million words (about 1.5 million being English). The average annotator agreement for OntoNotes is 91.8% for normal coreference and 94.2% for appositives (Hovy et al., 2006).

The authors of the ARRAU corpus (Poesio et al., 2008; Uryupina et al., 2020) concentrate on "difficult" cases of anaphora: plural anaphora, abstract object anaphora, and ambiguous anaphoric expressions, so the corpus has bridging reference and discourse deixis annotated. It contains only English texts (although there is an Italian analogue LiveMemories (Rodríguez et al., 2010)); its current size is 350,000 tokens. The inter-annotator agreement in ARRAU varies from 67% (annotation of anaphoric ambiguities) to 95% (annotation of complex anaphoric relations).

Thus, most of the largest corpora with coreference annotation contain predominantly English texts; however, with the growing interest in natural language processing of Non-English languages, corpora in other languages are being developed more often. As for the Russian language, there now exist two such datasets, one of them being RuCor (Toldova et al., 2014; Toldova et al., 2015) and the other AnCor (Budnikov et al., 2019).

RuCor contains texts from openly available sources, such as Russian OpenCorpora, Lib.ru and Lenta.ru (156,000 words in total). In this corpus the annotation process was conducted over morpho-syntactically pre-processed texts. The annotation scheme differentiates between primary and secondary markables, according to Potsdam Coreference Scheme (Krasavina and Chiarcos, 2007), where the primary markables are always annotated and represent specific references, while the secondary markables are annotated only if they are antecedents of any of the primary markables. Inter-annotator agreement for RuCor is 66% (Cohen's Kappa) or 85% (Mitkov's metric).

AnCor was created for the Ru-Eval competition in 2019 and contained 523 texts of various genres from Russian OpenCorpora (193,000 words in total). Named entities, common NPs and pronouns were annotated; the inter-annotator agreement for this dataset is 62.7% (75.5% agreement of both annotators and the final version).

As can be seen, although there are plenty of different corpora with coreference annotation, the largest and the most complex ones do not contain texts in Russian, and as for the Russian corpora, they are significantly smaller than the English ones, besides, their inter-annotator agreement is lower.

Therefore our main goals were to create a sufficiently big Russian corpus which would contain annotation of at least some difficult cases of anaphora with the inter-annotator agreement being high enough compared to OntoNotes and ARRAU.

2 RuCoCo: Russian Coreference Corpus

2.1 Data

We utilize the news stories published by NEWSru.com² as our source of text data. The texts were automatically collected and processed in the following way:

1. Any texts containing videos or embedded widgets from other websites were discarded as well as any texts marked as promotions.

²<https://www.newsru.com/>

2. Then the texts were converted to plain text format and cleared of any remaining HTML artifacts.
3. Texts that contained fewer than 20 tokens were also discarded, because they mostly consisted of a heading and a follow-up link only.
4. We then uniformly sampled one million words worth of texts across all text lengths and news categories. The total number of sampled texts is 3075.

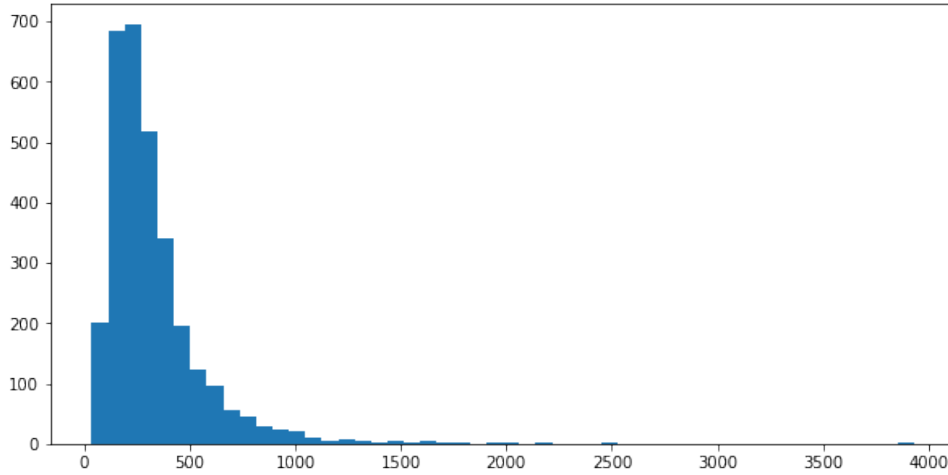


Figure 1: Distribution of text lengths in the sampled data.

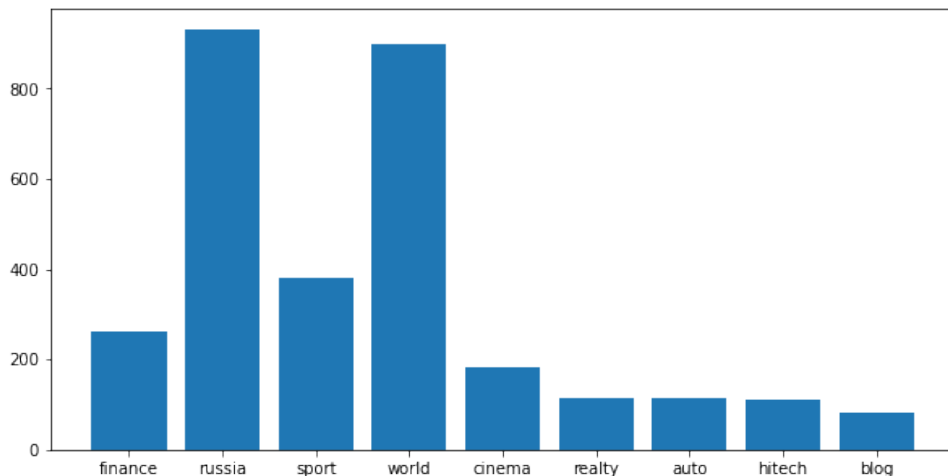


Figure 2: Distribution of news categories in the sampled data.

2.2 Annotation layer

The first release of RuCoCo covers identity (and in some cases, near-identity) coreference of noun phrases and pronouns. We do not annotate singletons, which means that each mention is linked to at least one other mention. We do not assign any attributes to the markables.

Mentions: We treat all noun phrases as potential mentions. Additionally the following types of pronouns are annotated:

- personal, possessive and reflexive pronouns;
- reciprocal pronouns, such as *друг друга* ('each other');
- relative pronouns;
- interrogative pronouns.

However, at this point we do not annotate coreference links with adjectives, clauses and expressions of time, all of which are going to be treated as valid mentions in the second revision of the corpus.

Mention boundaries: In most cases full noun phrases are annotated. To avoid overlapping of mentions referring to the same entity, participle and relative clauses that depend on the mention head are not included in mention boundaries. Therefore, in the following example there is no overlapping: {клиент}₀, {который}₀ хотел пополнить {свой}₀ счет ('{a customer}₀, {who}₀ wanted to top up {their}₀ account'). Parenthesis is not annotated unless it contains an independent clause, in which case it is treated as a regular sentence.

Coreference and anaphora: Coreference is annotated only for mentions of concrete entities. For generic mentions and mentions of abstract entities, events and properties we only annotate anaphora: *Может ли машина действовать разумно? Может ли {машина}₀ обладать сознанием? Может ли {она}₀ чувствовать?* ('Can a machine act intelligently? Can {a machine}₀ have a consciousness? Can {it}₀ feel how things are?'). Here, the first mention of *машина* ('a machine') is not annotated as coreferent with other mentions, because it is a generic mention.

Ellipsis: Mentions with elided heads are not annotated, as it would create ambiguity: *Это твоя сестра или {Даниэля}₀? Это {сестра {Даниэля}₀}₁, {она}₁ приехала на выходные.* ('Is this your sister or {Daniel's}₀? This is {{Daniel's}₀ sister}₁, {she}₁ came for the weekend.'). In the example above, the underlined mention could be recovered as *сестра Даниэля* ('Daniels' sister'), but we do not annotate it as referring to entity #1, because there would be two identical mentions referring to different entities.

Split antecedents: In RuCoCo, we annotate split antecedents as a means of dealing with the following challenges:

- Mentions referring to multiple referents: *{Премьер-министр}₀ и {госпожа Саймондс}₁ поженились вчера днем, небольшая церемония прошла в Вестминстерском соборе. {Пара}_{0,1} отпразднует свадьбу с семьей и друзьями следующим летом.*³ ('{Prime Minister}₀ has married {Carrie Symonds}₁ yesterday afternoon in a "small ceremony" at Westminster Cathedral. {The couple}_{0,1} would celebrate again with family and friends next summer.')
- Coordinate dependents: *{Сборные России и Канады}_{0,1} ранее ни разу не встречались в финалах чемпионатов мира. <..> {Отечественные хоккеисты}₀ победили {канадцев}₁ со счетом 5-3 в Стокгольме в 1989 году.* ('{National teams of Russia and Canada}_{0,1} have not played in IIHF finals before. <..> {The Russian team}₀ defeated {the Canadians}₁ 5-3 in Stockholm in 1989.')

Further in the text we refer to mentions linked to split antecedents as *plural anaphors* and to entities built from such mentions as *plural anaphor entities*. The number of such entities in the corpus can be seen in Table 1.

Category	Words	Mentions	Entities	PA-Entities	APA-Entities
russia	352,672	55,338	13,891	1,083 (7.8%)	2,471 (17.8%)
world	311,445	50,283	12,660	1,045 (8.3%)	2,122 (16.8%)
finance	94,015	11,739	3,020	176 (5.8%)	447 (14.8%)
sport	80,352	11,807	3,331	279 (8.4%)	705 (21.2%)
cinema	53,645	8,003	2,116	167 (7.9%)	431 (20.4%)
realty	34,227	4,509	1,274	72 (5.7%)	184 (14.4%)
hitech	31,365	3,895	1,080	77 (7.1%)	150 (13.9%)
auto	24,735	2,914	881	40 (4.5%)	94 (10.7%)
blog	17,649	1,917	624	39 (6.3%)	94 (15.0%)
Total	1,000,105	150,405	38,877	2,978 (7.7%)	6,698 (17.2%)

Table 1: Number of words, extracted mentions, entities, plural-anaphor (PA) entities and antecedent-of-plural-anaphor (APA) entities across the news categories in RuCoCo.

Metonymy: Linking of metonymies is allowed: *{Лондон}₀ и {Брюссель}₁ официально объявили о соглашении по Brexit. {Евросоюзу}₁ и {Великобритании}₀ удалось выработать*

³<https://www.newsru.com/world/30may2021/bjohnson.html>

соглашение об отношениях после Brexit. ({London}₀ and {Brussels}₁ have announced a Brexit trade deal. {The European Union}₁ and {the United Kingdom}₀ have agreed on a post-Brexit trade deal.').

Corpus format: RuCoCo is distributed as a collection of JSON-formatted files. An entity is represented as a list of character offset pairs. Antecedents of plural anaphor entities are listed in the "includes" section.

```
{
  "entities" : [[[31, 34]], [[39, 42], [100, 103]], [[71, 75]]],
  "includes" : [[], [], [0, 1]],
  "text": "At half-past nine, that night, Tom and Sid were sent to bed, as
  ↪ usual. They said their prayers, and Sid was soon asleep.\n"
}
```

Listing 1: JSON-formatted annotation of the following example: *At half-past nine, that night, {Tom}₀ and {Sid}₁ were sent to bed, as usual. {They}_{0,1} said their prayers, and {Sid}₁ was soon asleep.*

3 Corpus annotation

3.1 Metrics

There exist a number of coreference evaluation metrics, such as *MUC* (Vilain et al., 1995), *B³* (Bagga and Baldwin, 1998), *CEAF* (Luo, 2005), *BLANC* (Recasens and Hovy, 2011) and others. Since the CoNLL-2012 shared task (Pradhan et al., 2012), the average score of *MUC*, *B³* and *CEAF_e*, has become a de-facto standard way to evaluate coreference resolution systems. However, several shortcomings of these three metrics were demonstrated by Moosavi and Strube (2016), who also introduced *LEA*, a coreference evaluation metric designed to overcome those shortcomings. *LEA* of a set of entities *K* is computed as:

$$\frac{\sum_{e_i \in E} (importance(e_i) \times resolutionScore(e_i))}{\sum_{e_j \in E} importance(e_j)} \quad (1)$$

where $importance(e) = |e|$ and the resolution score of entity k_i is calculated against the response set of entities R as follows:

$$resolutionScore(k) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)} \quad (2)$$

Here, $link(e)$ calculates the number of unique coreference link within e : $link(e) = |e| \times (|e| - 1) / 2$.

We adopt *LEA* as our primary metric for measuring inter-annotator agreement and evaluating the neural coreference resolution model. As *LEA* does not support split antecedents out of the box, we modify the metric in the following way: for each plural anaphor entity we additionally calculate the scores of a special dummy entity with *importance* set to be the number of antecedent entities and *resolutionScore* computed based on the directed links between the plural anaphor entity and its antecedent entities.

The corpus was annotated by a team of 20 students of General Linguistics. The annotators were chosen based on trials that involved annotating documents of up to 1500 words. Each of the resulting documents was compared to the gold annotation using the *LEA* metric. The passing score was set to 0.9; the passing rate was 67%. Five of the annotators with the highest annotation quality were later appointed as moderators.

3.2 Neural pre-annotator

To speed up the annotation process, we developed a neural coreference resolution model to pre-annotate the texts. The model is based on the architecture proposed by Lee et al. (2018) and improved by Joshi et al. (2019) with the following differences:

- We use the Russian version of RoBERTa (Liu et al., 2019) pretrained by Sber AI⁴.
- We replace the neural mention extraction module with a rule-based syntactic mention extractor built on top of spaCy (Honnibal and Montani, 2017). This allows us to explicitly define what a mention is instead of relying on neural networks for mention extraction.
- Following Dobrovolskii (2021), we represent mentions using only weighted sums of the subtoken embeddings that constitute the mention.

To train the model, we used the automatically merged annotations obtained during the early phases of annotation. We ignored plural anaphors and used the original *LEA* to evaluate the pre-annotation quality. The model performed at 0.62 F1 after being trained on 100,000 words, at 0.68 F1 after being trained on 400,000 words and at 0.73 F1 after training on the whole dataset of 1,000,000 words.

3.3 Annotation process

The annotation process consisted of two steps: the first 100,000 words were annotated from scratch, i.e. the task was to identify and link all coreferent mentions in raw texts; the remaining 900,000 words were first pre-annotated by a neural coreference resolution model and the annotators were asked to correct the resulting documents.

Each text in the corpus was annotated by two annotators and then finalized by a moderator who received an automatic merge of the two versions with differences highlighted. Additionally, 3500 words of each annotator were manually checked by the authors of the markup scheme to provide feedback on an early stage.

3.4 Inter-annotator agreement

We measured the inter-annotator agreement and found it to be 0.759 F1. Because the annotators do not have a closed set of mentions to link, we suspect that some of the differences between annotations can be attributed to lack of attention. To eliminate this factor, we conducted the following experiment on a subset of the data approximately 50,000 words in size: each annotator was given back their own annotations automatically merged with the other annotation versions. The annotators were asked to independently correct the documents. The resulting inter-annotator agreement was 0.890 F1.

3.5 Disagreement analysis

We analysed discrepancies of the two phases of corpus annotation: 1) from scratch (50 random texts, about 16,000 words examined) and 2) pre-tagged annotation (158 random texts, 50,000 words examined). Discrepancies were divided into several categories:

- missing/redundant coreference cluster;
- missing/redundant markable;
- missing/redundant anaphoric chain;
- plural anaphors with split antecedents;
- mentions referred to different entities;
- NP borders.

To make the comparison more informative, we carried out the error analysis of the neural model used for pre-tagging, although we need to keep in mind that after the first 100,000 words were checked, we made a number of minor clarifications and changes in the guidelines to facilitate the work of our annotation team. See the comparison of discrepancies in annotation from scratch, model errors and pre-tagged texts in Figure 3.

By **missing/redundant coreference cluster** we mean all cases when one of the two annotators skipped the whole cluster or marked up an unnecessary entity. It is the most frequent type when annotators disagree (about 39% for both annotation stages). There was no closed set of entities, moreover, for abstract and generic entities, events or referents denoting open sets (so-called non-concrete entities) only anaphora must be annotated. Thus, annotators should decide whether the entity is concrete or non-concrete. They disagree on the following examples: locations without proper names like *кризисный*

⁴<https://github.com/sberbank-ai/model-zoo>

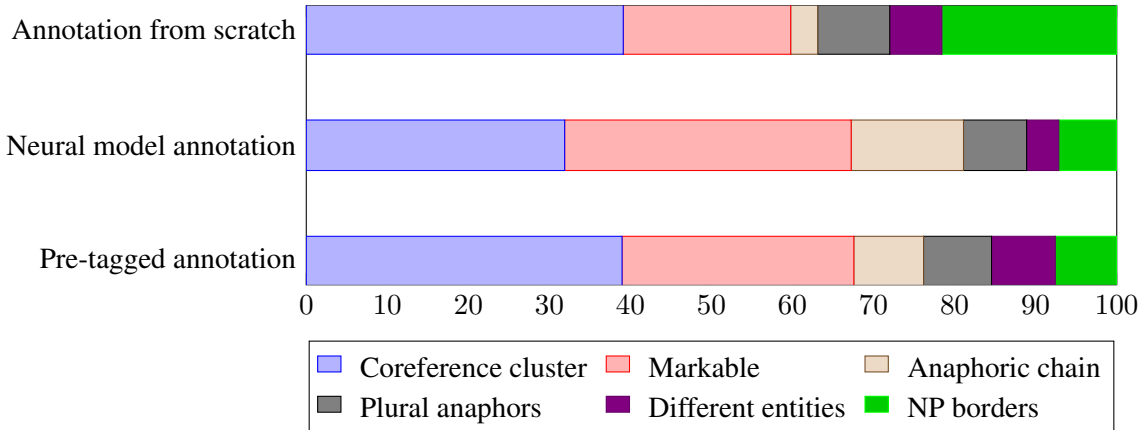


Figure 3: Comparison of discrepancies on all annotation steps, %

регион ('crisis area'), *жилой квартал* ('residential area'), and also when locations are nested in an organisation name: *Россия* ('Russia') in *Министр транспорта России* ('Minister of Transport of Russia') or in *Верховном суде РФ* ('Supreme Court of the Russian Federation'). Some other popular types are events like *концерт в Москве* ('the concert in Moscow'), *чемпионат России по хоккею* ('Russian ice hockey championship') and some abstract entities that are very similar to events as they have participants like *контракт* ('contract'), *уголовное дело* ('criminal case').

Missing/redundant markable (about 20% and 28% respectively) is the case when an annotator missed one or several mentions, although the coreference cluster is there in both annotation versions. For these cases we examined types of NPs missed by one annotator in the annotation from scratch stage, having preserved the taxonomy as in (Toldova et al., 2015) in order to compare them. See Table 2 to check numbers. We can observe that both annotation groups of students tend to miss noun groups (i.e. noun phrases headed by a noun) more than any other NP type.

NP Type	Our Data, %	Toldova et al., 2015, %
Reflexive pronouns	4.73	3.76
Relative pronouns	1.77	6.20
Anaphoric pronouns	4.73	12.47
Possessive pronouns	2.37	6.48
Noun groups	85.2	71.08
Adverbs (here/there)	1.18	0.00

Table 2: Types of missed NPs

As for **missing/redundant anaphoric chain** (3.3% and 8.6%) i.e. chains with abstract or generic entities where only anaphora resolution was performed, annotators mostly missed chains containing a relative pronoun *который* ('which/that') as an anaphoric element e.g. *срок, до которого* ('the deadline by which'), *той политической линии, которую* ('the policy that').

In **plural anaphors with split antecedents** (9% out of all discrepancies, both stages), the most common discrepancy is a missing relation between a person and a group of people: a son and a family, *Кондолиза Райс* ('Condoleezza Rice'), *сенатор Хиллари Клинтон* ('Senator Hillary Clinton') and *политики* ('politicians'). Less frequent cases of disagreement are the following: part-whole relations (which are not annotated as split anaphora) and entities denoting several items with part of these items as split antecedents: *50 терактов* ('50 terror attacks') and *20 терактов* ('20 terror attacks').

Mentions referred to different entities (6.4% and 7.9%) include cases where one or several mentions were assigned to different clusters by annotators in some confusing contexts (e.g. pronouns) or one annotator labelled some mentions in one and the same chain while the other one has divided it into

several chains e.g. cases with metonymy like *Пхеньян* ('Pyongyang') and *КНДР* ('North Korea'), *Израиль* ('Israel') and *Израильская армия* ('Israel Defense Forces').

Disagreement on **NP borders** covers 21% of discrepancies in the first stage and substantially less on the pre-tagged stage (7.5 %). We may assume that it may be due to the ability of our model to find correct borders or that it is due to the clarified guidelines of syntactic ambiguities we made before the second annotation stage: we have highlighted that in all such cases the maximum NP border must be annotated. This category presupposes cases where annotators excluded modifiers as in *изменения* ('changes') vs. *самые существенные изменения* ('the most significant changes'), complements e.g. *Банк* ('the Bank') vs. *Банк России* ('the Bank of Russia') and less often appositives: *Берт Ньюборн* ('Burt Neuborne') vs. *Берт Ньюборн, профессор права Университета Нью-Йорка* ('Burt Neuborne Professor of Civil Liberties at New York University').

This analysis was presented to the moderators so that they would know what to pay attention to. Despite all these discrepancies, the resulting inter-annotator agreement is still 0.890 F1 and all the disagreements were resolved by our moderators.

4 Conclusion

The result of our work is the Russian Coreference Corpus, which is the largest corpus with coreference annotation for Russian so far. We managed to achieve almost 90% inter-annotator agreement; we also analyzed the most common disagreements between our annotators so that we know what issues are to be solved. Further developments will include annotating more difficult cases of anaphora as well as increasing the size and genre diversity of the corpus.

Acknowledgements

We are grateful to our annotation team from General Linguistics Department of RSUH for their hard work, attentive approach to the project and immense help in discussions. We would also like to thank Prof. Svetlana Toldova and Evgeniya Inshakova for their useful observations and helpful advice.

References

- Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. // *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, P 79–85, Montreal, Quebec, Canada, August. Association for Computational Linguistics.
- A. E. Budnikov, S. Yu. Toldova, D.S. Zvereva, D. M. Maximova, and M. I. Ionov. 2019. Ru-eval-2019: Evaluating anaphora and coreference resolution for russian. // *Computational Linguistics and Intellectual Technologies - Supplementary Volume*.
- 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.
- George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. // *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).
- Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. // *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
- Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.
- Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. // *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, P 57–60.

- Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and 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 5803–5808, Hong Kong, China, November. Association for Computational Linguistics.
- Olga Krasavina and Christian Chiarcos. 2007. Pocos-potsdam coreference scheme. // *Proceedings of the Linguistic Annotation Workshop*, P 156–163.
- 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.
- 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*, abs/1907.11692.
- Xiaoqiang Luo. 2005. On coreference resolution performance metrics. // *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, P 25–32, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.
- 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.
- Massimo Poesio, Ron Artstein, et al. 2008. Anaphoric annotation in the arrau corpus. // *LREC*.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. // *Joint Conference on EMNLP and CoNLL - Shared Task*, P 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. // *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, P 143–152, Sofia, Bulgaria, August. Association for Computational Linguistics.
- M. Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17:485 – 510, 10.
- Kepa Joseba Rodríguez, Francesca Delogu, Yannick Versley, Egon Stemle, and Massimo Poesio. 2010. Anaphoric annotation of wikipedia and blogs in the live memories corpus. // *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.
- S. Toldova, A. Roytberg, A. Ladygina, M. Vasilyeva, I. Azerkovich, M. Kurzukov, G. Sim, D. Gorshkov, A. Ivanova, A. Nedoluzhko, and Grishina Y. 2014. Ru-eval-2014: Evaluating anaphora and coreference resolution for russian. // *Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue"*, P 681–694.
- Svetlana Toldova, Ilya Azerkovich, Yulia Grishina, Alina Ladygina, Olga Lyashevskaya, Anna Roytberg, Galina Sim, and Maria Vasilieva. 2015. Pre-experiments on annotation of russian coreference corpus. *Higher School of Economics Research Paper No. WP BRP*, 35.
- Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.
- Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. // *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Prosody and polysemy in Russian discourse formulae

Duryagin P. V.

Ca' Foscari University of Venice

/ Venice, Italy

pavel.duryagin@unive.it

Abstract

The paper presents the first descriptive approach to the prosody of Russian discourse formulae using the methods of experimental phonetics. The prosody of a frequent idiomatic formula *da nu* was studied. The analysis of pitch contours revealed that this unit can be marked by two tonal configurations: the falling one that can be identified as the IK-2 of the traditional holistic approach, and the rising one that is distinct from IK-3 and apparently contains a unique high tonal target followed by an irregularly truncated low boundary tone. The usage of these configurations is distributed unevenly and depends on the additional pragmatical meanings set by the dialogical context. In addition, the data indicates that some pragmatic meanings can be disambiguated by segmental duration. The subjects used significantly longer vowels in *da nu* to express surprise, compared to the rejection of new information. Meanwhile, the expression of doubt takes an intermediate position in this pragmatic continuum, marked by a lengthened pre-stressed and a shortened stressed vowel.

Keywords: phonetics; prosody; intonation; discourse formula

DOI: 10.28995/2075-7182-2022-21-150-158

Просодия и многозначность в русских дискурсивных формулах

Дурягин П. В.

Университет Венеции Ка' Фоскари

/ Венеция, Италия

pavel.duryagin@unive.it

Аннотация

В статье описан первый опыт использования методов экспериментальной фонетики для описания просодии многозначных дискурсивных формул русского языка. В качестве объекта была выбрана одна из наиболее частотных единиц такого типа – дискурсивная формула *да ну*. Анализ контуров ЧОТ показал, что эта идиоматическая единица может оформляться двумя тональными конфигурациями: нисходящей, которая может быть идентифицирована как ИК-2 системы интонационных конструкций, и восходящей, которая существенно отличается от ИК-3 и, вероятно, представляет собой одиночную «высокую» тональную цель в сочетании с нерегулярно усекаемым низким пограничным тоном. Употребления этих конфигураций по-разному распределены в зависимости от прагматических оттенков значения, задаваемых диалогическим контекстом. Кроме этого, согласно данным эксперимента, прагматические оттенки значения рассмотренной дискурсивной формулы могут маркироваться длительностью гласных. При выражении удивления испытуемые использовали более долгие гласные, чем при неприятии новой информации; при этом *да ну* при выражении недоверия занимает промежуточное положение и характеризуется продленным предударным гласным в сочетании с кратким ударным.

Ключевые слова: фонетика; просодия; интонация; дискурсивная формула

1 Введение

Традиционным предметом изучения интонологии является просодическое оформление высказываний, обладающих четко определенной синтаксической структурой: повествовательных предложений разной степени завершенности, общих и частных вопросов, различных типов речевых актов (апеллятивов, директивов и т. п.). Кроме этого, в исследованиях на материале русского языка неоднократно выдвигались предложения включать данные о просодических

характеристиках отдельных «дискурсивных слов» в словарные статьи (см. обзор литературы в [1] и реализацию идеи в [2]), а также предпринимались попытки определения того, как то или иное служебное слово способно модифицировать весь «интонационный контур» фразы (см., например, опыт описания «лексикализованной» просодии фраз, содержащих частицы *-то* и *вот* в [3]). При этом значительно меньше внимания в литературе уделялось просодии автономных идиоматических дискурсивных единиц, которые в результате диахронических процессов утратили синтаксическую структуру, не имеют в своем составе свободных слотов и используются только для выражения положительной или отрицательной реакции на реплику собеседника: *Да ну! Как же! Вот еще! А то! Как скажешь*. В этой работе вслед за [4] мы будем называть такие прагматикализованные неоднословные высказывания дискурсивными формулами (далее – ДФ).

В настоящее время системное описание инвентаря ДФ на материале русского языка осуществляется в рамках проекта «Русский прагматикон» [5]. Эта база данных содержит словарные статьи для более чем 600 единиц с указанием их структуры, значения, примеров употребления, а также типичной сопровождающей их жестикуляции и интонации. Опыт описания русских ДФ показывает, что практически все они «выражают положительную или отрицательную реакцию на реплику собеседника, которая может быть осложнена дополнительными семантическими или прагматическими компонентами» [4: 12]. Инструментами маркирования этих дополнительных компонентов значения являются жестикуляция и просодическое оформление высказывания. В то время как первые подходы к систематизации жестовой составляющей уже осуществляются на материале разметки мультимедийного подкорпуса НКРЯ [6], имеющиеся данные о просодии ДФ ограничиваются краткими ссылками на типы интонационных конструкций (далее – ИК) в «Прагматиконе».

Нам представляется, что при описании просодического оформления многозначных ДФ возникают следующие основные вопросы:

- 1) Какими просодическими средствами маркируются дополнительные прагматические компоненты значения в анализируемой ДФ?
- 2) Насколько последовательно носители русского языка используют эти просодические средства для маркирования многозначности анализируемой ДФ?
- 3) С какими единицами, используемыми для описания интонации синтаксических структур (ИК / тональными конфигурациями автосегментной метрической модели; далее – АМ-модели), могут быть соотнесены тональные контуры анализируемой ДФ?

В настоящей статье представлена попытка дать ответ на эти вопросы на примере одной многозначной ДФ – *да ну*. Выбор объекта обусловлен, во-первых, тем, что данные об этой ДФ не ограничиваются статьей «Русского прагматикона»: ранее семантика *да ну* и сопровождающая ее жестикуляция отдельно рассматривалась в [6], а пилотный эксперимент по изучению просодии был описан в [7]. Во-вторых, отсутствие в составе *да ну* глухих сегментов делает эту единицу удобным объектом для измерений значений частоты основного тона (далее – ЧОТ). Наконец, в-третьих, эта формула широко используется носителями русского языка в разговорной речи, что способствует успешной элицитации естественных диалогических данных в формате фонетического эксперимента.

2 Эксперимент

2.1 Стимулы

Материалом исследования послужили 12 коротких диалогов на бытовые темы, состоящих из 3–5 реплик. Одна из заключительных реплик каждого из этих «целевых» стимулов содержала ДФ *да ну*. В соответствии с существующими описаниями многозначности этой единицы было составлено по два диалога на следующие коммуникативные ситуации:

а) недоверие в ответ на сообщение факта, который говорящий не готов принять немедленно, так как располагает иной информацией (далее в тексте тип контекста условно обозначен как *Недоверие*);

б) удивление в ответ на сообщение факта, который говорящий ранее считал маловероятным, но оценивает как положительную новость и принимает (далее – *Удивление*);

в) отрицание в ответ на сообщение факта, который говорящий не принимает и продолжает считать ложным (далее – *Ложный факт*);

г) отказ в ответ на предложение совершить некоторое действие (далее – *Предложение*)

д) отрицание в ответ на предупреждение об опасности или о возможных нежелательных последствиях текущей ситуации (далее – *Опасность*);

е) отрицание в ответ на комплимент со стороны собеседника (далее – *Комплимент*)

Ниже в качестве примера приведен один из диалогов (тип ситуации – *Недоверие*, ДФ выделена жирным шрифтом):

(Испытуемый): *Ты Алену давно видел? Что-то она мне не пишет, не звонит...*

(Экспериментатор): *Так она же на стажировку в Германию уехала.*

(Испытуемый): ***Да ну?*** *Правда что ли? Она же немецкого не знает!*

(Экспериментатор): *Вот так вот. Выучила, наверное!*

Статья «Русского прагматикона» для *да ну* включает краткое описание просодии этой ДФ в терминах модели ИК Е. А. Брызгуновой [8]: для ДФ в функциях (*a, б, в*) постулируется контур ИК-2, а в функциях (*г, д, е*) – ИК-7. Для уточнения этих данных и ответа на вопрос о том, с какими единицами описания просодии традиционных синтаксических структур могут соотноситься контуры ДФ в диалогической речи, в материал эксперимента были дополнительно включены по три диалога, содержащих близкие *да ну* по сегментному составу однословные высказывания в контекстах, считающихся прототипическими для ИК-2 (частный вопрос ‘*Кому?*’, обращение к собаке ‘*Балу!*’, однословное восклицание ‘*Камю!*’) и ИК-3 (общие вопросы ‘*Тому?*’, ‘*Тону?*’, ‘*Балу?*’)¹. Помимо этого, в эксперимент были добавлены 12 диалогов-филлеров, содержащих иные частотные ДФ. Полный список 30 стимулов представлен в онлайн-приложении, доступном по адресу: <https://osf.io/3xmnu>.

2.2 Процедура и испытуемые

Интерфейс для эксперимента был создан в программе *PsychoPy* [9]. Диалоги были представлены на экране компьютера на отдельных слайдах, перемешанных в случайном порядке, индивидуально для каждого информанта. Реплики, предназначенные для испытуемого и для экспериментатора, были оформлены разными шрифтами. Впервые увидев диалог, испытуемый должен был прочитать его про себя. После этого он сообщал о своей готовности экспериментатору, и они вместе читали диалог по ролям.

В эксперименте приняли участие 32 испытуемых (23 женщины и 9 мужчин), средний возраст – 22,3 года, $SD = 2,3$ года. Все они на момент записи постоянно проживали в Москве и являлись студентами, аспирантами или недавними выпускниками столичных вузов. Половина из них родились и выросли в Москве или Московской области, остальные переехали из других регионов после окончания школы. Все испытуемые, по оценке автора, являются носителями русского литературного произношения без ярко выраженных диалектных черт.

2.3 Анализ данных

В рамках эксперимента было записано и проанализировано 384 произнесения ДФ *да ну* и 177 произнесений похожих по сегментному составу слов в «прототипических» контекстах ИК-3 и ИК-2². Все записи были вручную сегментированы и размечены в программе *Praat* [10]. В соответствии с практикой автосегментной метрической модели просодическая разметка включала в себя обнаружение в контуре «низких» и «высоких» тональных целей, соответствующих локальным минимумам и максимумам ЧОТ [11].

¹ Выбор ИК-3 вместо ИК-7 объясняется трудностью эlicitации конструкции седьмого типа в двусложных однословных высказываниях; кроме того, мы опираемся на анализ Е. А. Брызгуновой, которая в своих работах указывает, что эти единицы имеют одинаковый контур и различаются только наличием гортанной смычки [8: 118].

² Из анализа были исключены 15 единиц, в которых испытуемые использовали в «прототипических» контекстах контуры, отличные от ИК-2 и ИК-3; в большинстве случаев это было употребление особого «звательного» контура вместо ИК-2 при обращении к собаке: «Балу-у!».

На основании разметки были получены измерения длительности всех сегментов, а также при помощи скрипта *Praat* [12] были извлечены нормализованные по времени тональные контуры всех произнесений. Нормализация по времени представляет собой такое измерение ЧОТ, при котором значения нулевой форманты измеряются фиксированное количество раз через равные промежутки времени в пределах каждого из сегментов (например, в настоящей работе ЧОТ измерялась на 5%, 15%, 25%... 85%, 95% длительности обоих гласных и второго (сонорного) согласного в *да ну* и других стимулах). Получаемые в результате такого преобразования графики позволяют визуально сравнивать форму большого числа контуров, основываясь на тайминге тональных событий и «отфильтровывая» индивидуальные различия в темпе речи, а также эффекты типичного для ДФ эмфатического увеличения длительности сегментов. Для такой визуализации использовался пакет *R ggplot2* [13]. Объем публикации не позволяет привести в качестве иллюстрации все полученные графики, однако читатель может ознакомиться с ними в онлайн-приложении.

Статистический анализ результатов измерения ЧОТ и длительности сегментов проводился построением моделей линейной регрессии со смешанными эффектами с использованием пакетов *R lme4* [14] и *effects* [15]. Таблицы регрессионного анализа, созданные при помощи пакета *sjPlot* [16], также доступны в онлайн-приложении.

3 Результаты и обсуждение

3.1 Тайминг и направление движений ЧОТ

На основании визуального анализа тайминга и направлений движения тона на нормализованных тонограммах можно выделить два основных типа контуров, которые участники эксперимента использовали для маркирования прагматических оттенков в ДФ *да ну*:

- 1) «нисходящая» тональная конфигурация (далее для краткости – Контур 1): начинается на среднем или низком для говорящего уровне, достигает пика в конце предупредительного гласного либо в сонорной инициали ударного слога, завершается падением до значения на 1–2 стандартных отклонения ниже среднего уровня тона говорящего; см. Рис. 1(а);
- 2) «восходящая» тональная конфигурация (далее – Контур 2): начинается на уровне ниже среднего и достигает пика (на 1–4 стандартных отклонения выше среднего) либо в центре, либо в конце ударного гласного; см. Рис. 1(б). В случае, если максимум ЧОТ достигается ранее конца ударного гласного, за ним следует падение тона, как правило, незначительное, редко достигающее среднего для говорящего уровня.

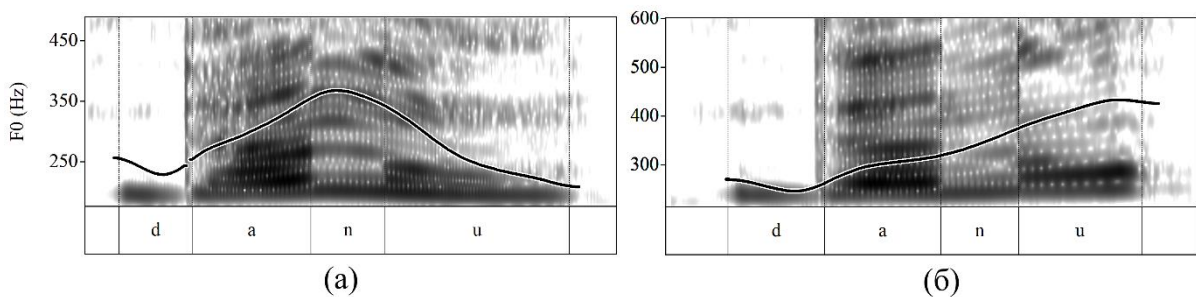


Рис. 1: Спектрограммы и оригинальные (не нормализованные) тональные контуры произнесений информантом женского пола ДФ *да ну* в контекстах «Удивление» (а) и «Опасность» (б)

Полученные экспериментальные данные позволяют внести ряд уточнений в описание, представленное в статье «Русского прагматикона».

Во-первых, обращает на себя внимание неодинаковое распределение типов контура по контекстам (Табл. 1). В то время как Контур 1 был практически обязателен для маркирования недоверия, удивления и отрицания ложного (с точки зрения говорящего) факта, использование Контура 2

является факультативным. «Восходящая» конфигурация использовалась для реакции на предложение, сообщение об опасности и комплимент примерно в половине записанных диалогов, при этом не всеми испытуемыми и непоследовательно.

Тип ситуации	Контур 1	Контур 2	Тип ситуации	Контур 1	Контур 2
<i>Недоверие</i>	63	1	<i>Предложение</i>	26	38
<i>Удивление</i>	61	3	<i>Опасность</i>	42	22
<i>Ложный факт</i>	58	6	<i>Комплимент</i>	27	37
Всего, в проц.	94,8	5,2	Всего, в проц.	49,5	50,5

Табл. 1: Результаты анализа тональных конфигураций в ДФ *да ну*; количество употреблений испытуемыми «нисходящего» Контура 1 и «восходящего» Контура 2

Во-вторых, формат эксперимента позволяет напрямую сравнить (как визуально, так и с помощью статистических методов) обнаруженные Контур 1 и 2 с ИК-2 и ИК-3 соответственно. Возможности визуального сопоставления нормализованных по времени контуров демонстрирует Рис. 2, содержащий графики произнесений стимулов одной из информанток, в речи которой Контур 1 и 2 были последовательно распределены по контекстам в соответствии с тенденцией, представленной в Табл. 1³.

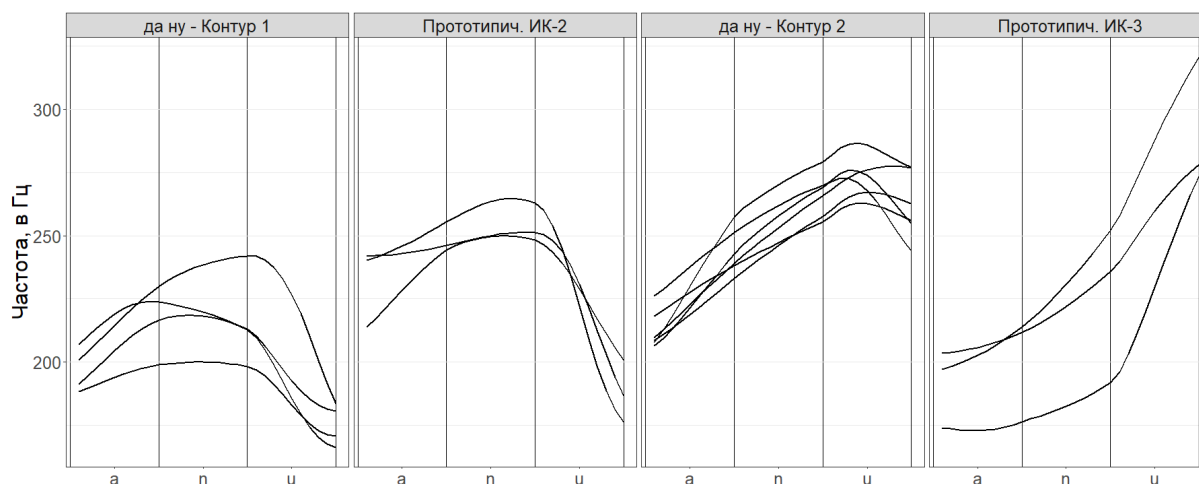


Рис. 2: Нормализованные по времени тональные контуры стимулов в произнесении испытуемой женского пола (измерения ЧОТ на звонкой смычке и взрыве [д] опущены)

При визуальном сравнении Контура 1 и «прототипических» произнесений ИК-2 обращает на себя внимание бóльшая стабильность тайминга и больший диапазон падения тона, характерные для единиц второго типа. Визуальная оценка подтверждается результатами статистического анализа: смешанная модель линейной регрессии, построенная для нормализованных по времени измерений ЧОТ (тип контекста как фиксированный эффект; случайные свободные члены для испытуемого, порядкового номера места измерения ЧОТ и лексического наполнения стимула) подтверждает, что испытуемые произносили стимулы «прототипического» ИК-2 на более высоком уровне тона, чем различные варианты *да ну* ($p < 0,001$). Представляется, однако, что эта вариативность имеет «паралингвистический» характер (см. подробнее о лингвистических и паралингвистических элементах интонации в [17: 49–70]), то есть, отражает не категорическое противопоставление двух разных тональных конфигураций, а градуальное варьирование, используемое говорящим для маркирования разной степени эмфазы в пределах одного и того же тонального

³ Исключение составили два стимула типа «Ложный факт»; в обоих участница эксперимента не отделяла ДФ от последующей фразы просодически («Да ну ты перепутал!»); эти контуры (вернее, «предъядерные» элементы более крупных тональных конфигураций) опущены на Рис. 2.

акцента. Например, особо сильное удивление может маркироваться увеличенным диапазоном движения тона и повышенным «стартовым» уровнем ЧОТ во фразе, так же как более настойчивая попытка позвать кого-либо по имени, более строгая просьба или приказ и т. п. Вероятно, обнаруженные значимые различия в уровне ЧОТ объясняются тем, что все контексты ИК-2 (говорящий подзывает к себе собаку, произносит имя писателя после долгих попыток вспомнить его, в возмущении задает риторический вопрос *Кому (жаловаться)?*) в большей степени требовали от испытуемых выражения эмфазы, в то время как использованные в эксперименте контексты употребления *да ну* оставляли возможность для относительно сдержанной реакции на речевой акт собеседника.

Что же касается вариативности тайминга ИК-2, то это явление подробно описано в литературе о т. н. «заносе»: известно, что факультативное передвижение пика нисходящего контура на первый предупредительный слог в целом характерно для современного литературного произношения и, в частности, для московского варианта нормы [18]. Кроме того, в некоторых случаях в наших данных наблюдаемая на нормализованных графиках вариативность тайминга пика Контур 1 является артефактом нормализации по времени и связана с существенной степенью варьирования длительности сегментов – просодического средства, о котором пойдет речь в разделе 3.2.

Таким образом, в соответствии с существующими описаниями мы предлагаем предварительно идентифицировать Контур 1 как ИК-2 системы Брызгуновой; в транскрипции АМ-модели эта конфигурация может быть обозначена как $H^*+L L\%$ или $H+L^* L\%$ (дискуссия о том, какому из двух «ядерных» тонов должен быть присвоен астериск, выходит за рамки настоящего исследования).

При сравнении Контур 2 и «прототипических» контуров ИК-3 обращают на себя внимание два основных различия. Во-первых, на предупредительном слоге Контур 2 отсутствует «нырок» ЧОТ, обязательный для контура общего вопроса в русском языке [19] и соответствующий низкой тональной цели L в «ядерном» тональном акценте L^*+H либо $L+H^*$ [20, 21], соответствующем центру ИК-3. Это заметно при визуальном сравнении контуров различие подтверждается статистическим анализом: смешанная регрессионная модель для измерений ЧОТ на первом предупредительном гласном (тип контекста как фиксированный эффект; случайные свободные члены для испытуемого, порядкового номера места измерения ЧОТ и лексического наполнения стимула) указывает на то, что первый слог ДФ, оформленный Контуром 2, произносится на значительно более высоком уровне, чем первый слог аналогичных по сегментному составу общих вопросов, оформленных ИК-3 ($p < 0,001$). Вторым важным различием является то, что трункция пограничного тона $L\%$, наблюдаемая в русском общем вопросе (при отсутствии ударных слогов контур ИК-3 обязательно «обрывается» на пике [21]), при маркировании ДФ *да ну* Контуром 2 имеет факультативный характер.

Таким образом, представленные данные позволяют сделать вывод о том, что Контур 2 ДФ *да ну* не может быть идентифицирован как ИК-3. Мы предлагаем предварительно интерпретировать его в рамках АМ-модели как $H^* (L\%)$ – монотональный «ядерный» тональный акцент, высокая тональная цель которого ассоциирована с центром ударного гласного, а низкий пограничный тон, как правило, усекается и может сохраняться только за счет сверхдолгого конечного гласного. Будущим исследованиям предстоит выяснить, представлен ли такой контур в высказываниях с «регулярной» синтаксической структурой (ср., например, сомнение-предположение: «*Какого персонажа он может озвучить... Разве что Балу?..*»). Альтернативой такому поиску является присвоение дискурсивным формулам статуса «частных иллокуций», имеющих уникальное просодическое оформление, «хранящихся в памяти говорящих в готовом виде» [22: 98] и требующих «индивидуальной фиксации в иллокутивном словаре» [Там же: 127].

3.2 Длительность сегментов и тип фонации

Описывая просодические средства маркирования высказываний, следует помнить о том, что просодия включает в себя не только интонацию (контролируемые говорящим движения частоты основного тона), но и другие суперсегментные средства, а именно изменение длительности сегментов (ритмическая структура слова) и качества голоса (тип фонации).

Судя по полученным данным, именно ритмика слова может играть особую роль при маркировании прагматических компонентов значения в ДФ *да ну* в контекстах, регулярно оформлявшихся

ИК-2. Статистический анализ длительностей гласных (были построены отдельные смешанные модели линейной регрессии для длительности гласных [a] и [y] со случайными свободными членами для испытуемых и фиксированной переменной для типа контекста) показал, что первый предупредительный гласный [a] был значительно более кратким в контекстах типа *Ложный факт* по сравнению с контекстами типа *Недоверие* и *Удивление*. При этом длительность гласного [y] была значительно выше в только контексте *Удивление* и не различалась между контекстами *Ложный факт* и *Недоверие* (см. Рис. 3). Таким образом, имеет место закономерность: вероятно, степень принятия новой информации в ДФ *да ну* маркируется долготой сегментов. Выражая удивление и приятие в ответ на хорошие новости, испытуемые регулярно использовали продленные гласные как в первом предупредительном, так и в ударном слогах. Напротив, категорически отрицая ложную информацию, участники эксперимента использовали более краткие гласные в обоих слогах *да ну*. Интересно, что в этом отношении ритмическая структура ДФ в контекстах недоверия занимает промежуточное положение. Вероятно, последовательное продление предупредительного и сокращение ударного гласного при выражении недоверия может быть интерпретировано следующим образом: «я принимаю к сведению твою информацию» (продленный первый слог, как при удивлении), «но не могу согласиться с ее достоверностью, так как обладаю противоречащими ей данными» (сокращенный второй слог, как при отрицании ложного факта).

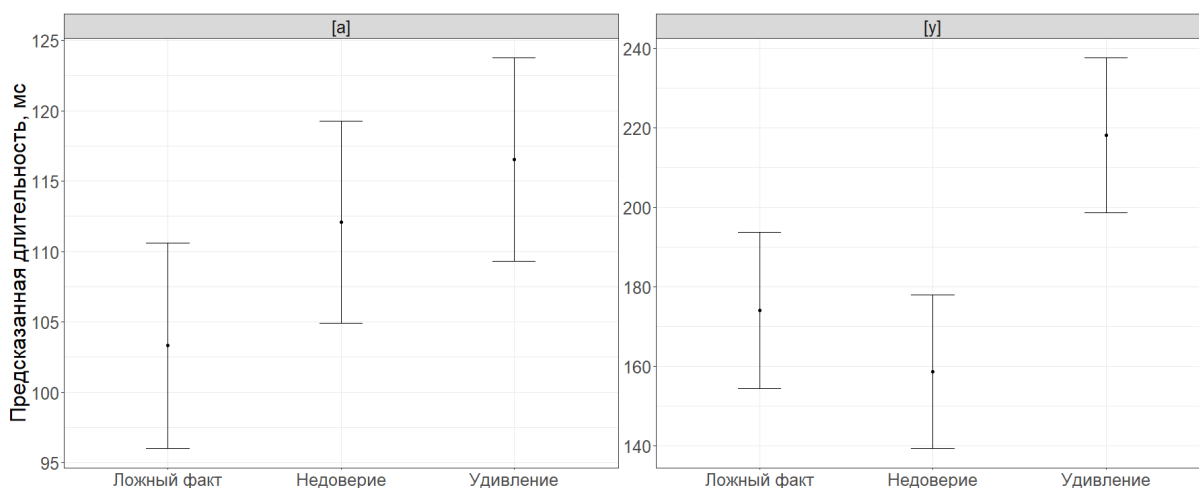


Рис. 3: Предсказанные регрессионными моделями (ожидаемые) значения и доверительные интервалы длительности гласных в ДФ *да ну*, оформленной ИК-2, в зависимости от контекста

Наконец, что касается фонации, то в наших данных информанты спорадически использовали аспирацию для маркирования «приятного удивления». Подобные аффективные функции типа голоса, называемого в англоязычной литературе *breathy voice* (голос «с придыханием»), ранее неоднократно описывалась как в литературе, посвященной невербальной коммуникации [23], так и в лингвистических исследованиях, в частности, на материале русского языка [24: 27–28, 143, 225]. Нам представляется, что подобные данные также должны находить место в системном описании просодии ДФ.

4 Выводы

Проведенное исследование просодии ДФ *да ну* позволяет внести ряд важных уточнений в существующие описания. Экспериментальные данные подтверждают, что многозначность этой единицы маркируется двумя разными тональными конфигурациями. «Нисходящий» контур, характеризующийся высокой вариативностью тайминга пика и диапазона падения ЧОТ, был идентифицирован как ИК-2 ($H^*+L L\%$ или $H+L^* L\%$ АМ-модели). При этом альтернативная «восходящая» конфигурация определенно не может быть идентифицирована как ИК-3, поскольку содержит единственную «ядерную» тональную цель H^* , за которой следует нерегулярно усекаемый низкий пограничный тон $L\%$. Особого внимания заслуживают данные о дистрибуции двух

контуров: было показано, что в зависимости от контекста и индивидуальных предпочтений говорящего та или иная конфигурация может быть как обязательной, так и факультативной. Наконец, нам удалось получить подтверждение того, что маркирование неоднозначности ДФ не ограничивается интонацией: в контекстах употребления *да ну*, обязательно оформляемых ИК-2, участники эксперимента использовали длительность гласных просодического ядра для маркирования континуума *Принятие – Недоверие – Отрицание* в ответ на новую информацию, полученную от собеседника.

Предложенная в настоящей работе методология сбора и анализа данных может быть использована для дальнейшего пополнения базы ДФ; в частности, особый интерес представляет изучение механизма маркирования многозначности в тех редких формулах, которые способны выражать как согласие с собеседником, так и отрицание (например, *Как же!*). По нашему мнению, систематизация данных о просодии ДФ может послужить решению ряда нетривиальных прикладных задач, таких как преподавание прагматики неродного языка, а также автоматическое распознавание эмоций в звучащей речи.

References

- [1] Kobozeva I.M., Zaharov L.M. (2004), Why do we need an audio dictionary of discourse words? [Dlja čego nužen zvučaščij slovar' diskursivnyx slov russkogo jazyka], Proceedings of the International Seminar "Dialog 2004" [Trudy Meždunarodnogo seminar "Dialog-2004"], Moscow, pp. 292–297.
- [2] Active dictionary of the Russian language [Aktivnyj slovar' russkogo jazyka] / Editor-in-Chief: Yu. D. Apresjan. — Moscow: Jazyki russkoj kultury, 2014. — Vol. 1.
- [3] Levontina I.B. (2016), Lexicalized prosody and the polysemy of discourse markers, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2016"], Bekasovo, pp.369–381.
- [4] Rakhilina E.V., Bychkova P.A., Zhukova, S.Yu. (2021), Speech acts as a linguistic category: The case of discourse formulae [Rečevye akty kak lingvističeskaja kategorija: diskursivnye formuly], Topics in the Study of Language [Voprosy Jazykoznanija], №2, pp. 7–27.
- [5] Yaskevich A., Bychkova P., Koziuk E., Rakhilina E., Slepak E., Utkina A., Zhukova S., Zotova T. (2021) The Russian Pragmaticon. An electronic database of the Russian pragmatic constructions. Available at <https://pragmaticon.ruscorpora.ru/>.
- [6] Bychkova P.A., Rakhilina E.V., Slepak E.A. (2019), Discourse formulae, polysemy and gesture marking [Diskursivnye formuly, polisemija i žestovoe markirovanie], Proceedings of the V.V. Vinogradov Russian Language Institute [Trudy Instituta Russkogo Yazyka imeni V.V. Vinogradova], Vol. 21, pp. 256–283.
- [7] Duryagin P.V., Rakhilina E.V. Prosodic means for marking of polysemy in discourse formula da nu [Prosodičeskije sredstva markirovanija polisemii diskursivnoj formuly da nu], Abstracts of the 6th International scientific conference "Russian linguistic culture" [Tezisy VI Meždunarodnoj naučnoj konferencii "Kul'tura russkoj reči"], Moscow. Available at <https://drive.google.com/file/d/1eXDZzrSpHqhYU3fPqVk0KMSPGnKkTXuM>.
- [8] Bryzgunova E.A., Intonation [Intonacija], Russian Grammar [Russkaja grammatika], Vol. 1, pp. 96–120.
- [9] Peirce J. W., Gray J. R., Simpson S., MacAskill M. R., Höchenberger R., Sogo H., Kastman E., Lindeløv J. (2019), PsychoPy2: experiments in behavior made easy, Behavior Research Methods, Vol. 51, pp. 195–203.
- [10] Boersma P., Weenink D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.09, retrieved 15 February 2022 from <http://www.praat.org/>.
- [11] Arvaniti A. (2020), The Phonetics of Prosody, Oxford Research Encyclopedia of Linguistics. Available at <https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-411>.
- [12] Arantes P. (2021). parantes/time-normalized-f0: Version 2.0 of time-normalized-f0 Praat script (v2.0). Available at <https://doi.org/10.5281/zenodo.4489748>.
- [13] Wickham H. ggplot2: Elegant Graphics for Data Analysis. — Springer-Verlag New York, 2016.
- [14] Bates D., Mächler M., Bolker B., Walker S. (2015), Fitting Linear Mixed-Effects Models Using lme4, Journal of Statistical Software, Vol. 67(1), pp. 1–48.
- [15] Fox J., Weisberg S. (2019). An R Companion to Applied Regression, 3rd edition. — Sage, Thousand Oaks CA, 2019.
- [16] Lüdecke D (2021). sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.10. Available at <https://CRAN.R-project.org/package=sjPlot>.
- [17] Gussenhoven C. The Phonology of Tone and Intonation. — Cambridge: Cambridge University Press, 2004.

- [18] Kasatkina R.F. (2005), The Moscow akanije in the light of dialectal data [Moskovskoe akan'e v svete nekotoryx dialektnyx dannyx], Topics in the Study of Language [Voprosy Jazykoznanija], №2, pp. 29–45.
- [19] Rathcke T. (2006), A perceptual study on Russian questions and statements, Arbeitsberichte Des Instituts Für Phonetik Und Digitale Sprachverarbeitung Der Universität Kiel (AIPUK), Vol. 37, pp. 51–62.
- [20] Igarashi Y. (2009). Russian Interrogatives and Intonational Categories, The Discourse Potential of Under-specified Structures, pp. 227–270.
- [21] Rathcke T. (2017), How Truncating Are “Truncating Languages”? Evidence from Russian and German, *Phonetica*, Vol. 73(3–4), pp. 194–228.
- [22] Yanko T. E. Intonation strategies of Russian speech in a comparative aspect [Intonacionnye strategii russkoj reči v sopostavitel'nom aspekte]. — Moscow: Jazyki slavjanskix kul'tur, 2008.
- [23] Anikin A. (2020), A Moan of Pleasure Should Be Breathly: The Effect of Voice Quality on the Meaning of Human Nonverbal Vocalizations, *Phonetica*, Vol. 77(5), 327–349.
- [24] Kodzasov S.V. Studies in Russian prosody [Issledovanija v oblasti russkoj prosodii]. — Moscow: Jazyki slavjanskix kul'tur, 2009.

Motion verbs in multimodal communication

Evdokimova Alexandra
Institute of Linguistics, Russian
Academy of Sciences, Moscow, Russia
arochka@gmail.com

Nikolaeva Yulia
Moscow State University,
Moscow, Russia
julianikk@gmail.com

Budennaya Evgeniya
Higher School of Economics,
Institute of Linguistics, Russian Academy of Sciences,
Moscow, Russia
jane.sdrv@gmail.com

Abstract

The article explores correlations between motion verbs and head and hands gestures using the RUPEX corpus. The verbs are divided into four groups based on their meanings. Monological and dialogical parts of the recordings are compared along with the speaker's role and viewpoint in gestures. The pilot analysis of motion verbs in the multimodal corpus showed that the relationships between verb type, non-verbal behavior and speaker's role depend on a complex set of factors and manifests itself in different ways in different channels. In the verbal channel no direct relationship between the semantic type of the verb and the speaker's role was detected; however, the narrators and commentators who have seen the film used more affectional vocabulary than the reteller while the latter tended to use more vector-prefixed verbs. In manual channel prefixes or their absence do not influence the use of hand gestures. Transitive verbs meaning manipulations of different items are more probable to be illustrated by depictive gestures. Predictably, motion verbs in the strict sense are more prone to be supported by observer viewpoint (O-VPT) gestures, while verbs of manipulation are usually used with C-VPT gestures. In cephalic channel motion verbs in the strict sense (relocation of a character) are usually illustrated by O-VPT depictive gestures, and manipulation verbs are more probably supported by pantomime C-VPT gestures similar to manual channel. In some head gestures the viewpoint is combined. If the verb is repeated by the same or another speaker the gestures differ in both manual and cephalic channels. Cephalic gesture clusters on motion verbs have mostly a depictive function, which may be considered a gestural illustration.

Keywords: verbs of motion, hand gestures, head gestures, multimodal communication

DOI: 10.28995/2075-7182-2022-21-159-175

Глаголы движения во взаимодействии с жестами рук и головы

Евдокимова А.
Институт языкознания РАН,
Москва, Россия
arochka@gmail.com

Николаева Ю.
Московский государственный
университет имени
М. В. Ломоносова, Москва, Россия
julianikk@gmail.com

Буденная Е.
Высшая школа экономики,
Институт языкознания РАН,
Москва, Россия
jane.sdrv@gmail.com

Аннотация

В статье на материале корпуса RUPEX анализируется влияние семантики глаголов, на примере глаголов движения, на выбор жеста головы или рук. Выделенные 4 семантические группы глаголов движения и жестовый материал из двух кинетических каналов (мануального и цефалического) рассматривается с учетом типа коммуникации (монолог vs. диалог), роли говорящего (рассказчик/комментатор/пересказчик), позиции наблюдателя или персонажа. Было обнаружено, что в вербальном канале нет строгой связи между семантическим типом глагола и ролью говорящего, однако рассказчики и комментаторы, непосредственно видевшие фильм, используют больше оценочной лексики, по сравнению с пересказчиком. В мануальном канале на глаголах движения преобладают жесты с точки зрения персонажа, особенно в монологе. С приставочными и бесприставочными глаголами разница в использовании жестов незначима. Глаголы, обозначающие действия участников — манипуляции с другими предметами — чаще будут сопровождаться изобразительными жестами по сравнению с глаголами перемещения. Глаголы, описывающие перемещение референтов, гораздо чаще сопровождаются жестами наблюдателя, а упоминания действий участников — жестами персонажа. В цефалическом канале на глаголах собственно движения преобладают изобразительные жесты, чаще всего показывающие направление и роль наблюдателя (O-VPT). На глаголах с семантикой движения рук увеличивается процент пантомимных жестов, что указывает на роль персонажа (C-VPT), в некоторых жестах эти роли накладываются. При повторах глаголов, занимающих ту же позицию в композиции рассказа, самим говорящим или его слушателями наблюдается тенденция несовпадения жестов. Кластеры жестов из двух каналов чаще изобразительные, что указывает на использование говорящими объединения двух каналов для иллюстрации глаголов движения как прием.

Ключевые слова: глаголы движения, жесты рук, жесты головы, мультимодальная коммуникация

1 Introduction. Motion verbs and accompanying gestures

Recently multimodal studies in linguistics have addressed some new research topics. One promising area of studies can be exploring phenomena considering type and structure of communication (e.g. monologue vs. dialogue), speakers' roles and their stances.

Some studies based on multimodal corpora showed how these factors are connected to verbal, prosodic and kinetic behavior [4, 13, 21, 24]. This article continues the perspective and explores correlations between the use of motion verbs and accompanying gestures.

According to classical works on lexical semantics, motion verbs [1, 15, 18, 31] are those which describe relocation of the subject moving from start to endpoint. C.Fillmore divides them into Source-oriented and Goal-oriented verbs [9] considering which of the points (starting or ending) the verb is oriented at. Motion verbs can be also divided into allative vs. ablativ [8], lative vs. elative [24], centripetal vs. centrifugal [the term suggested by I.A. Sternin [23]]. E.V. Rakhilina [25] suggested a classification of Russian motion verbs; in [17] T. A. Maisak and E. V. Rakhilina examined in detail the verb *idti* ('to go').

Gesture accompaniment of Russian motion verbs was studied in [5; 20; 34], among others. Our work is based on the RUPEX corpus (www.multidiscourse.ru and [12]), namely reference subcorpus including recordings #04, #22, #23, a total duration of about 1 hour, each recording consisted of three individual videos and one video from a wide-angle camera. The corpus includes vocal annotation and annotations of three kinetic channels: oculomotor, manual (hand gestures) and cephalic (head gestures). Each recording has four stages and four participants with fixed roles: 1) Narrator (N) and Commentator (C) watch "The Pear Film" [5]; 2) N retells the story to Reteller (R), who has not seen the film (first monologue); 3) C can add to what was told by N and R can ask any questions about the film to N and C (dialogue stage); 4) Listener (L) comes and R retells him the story (second monologue).

We studied the retellings in the RUPEX and discovered that motion verbs mark key points of the story. In this article we study the verbs and accompanying hand and head gestures considering the stage of the recording and the speaker's role. We considered functional types of gestures [14] and the speaker's viewpoint in gesture [16]. Since Russian is a satellite-framed language [30] (the verb describes manner of movement, and prefixes, prepositions and adverbs show direction and path [10]), we expected character viewpoint (C-VPT, the speaker's gesturing as a character in the story) to be more often with unprefix verbs, and observer viewpoint (O-VPT, the speaker's gesturing as someone watching the scene and not participating in it) to be used with prefixed verbs, when path and trajectory are highlighted.

We examined core motion verbs and compared them to three other groups, verbs describing: 1) hand movements (*sobral* 'gathered', *brosil* 'threw', *dal* 'gave' etc.), body movements (*ogljansja* 'looked back', *osmotrel* 'inspected', *povernulsja* 'turned around' etc.), or inanimate referents movements (*razletelis* 'flew apart', *sletela* 'flew off' etc.). Adding three new groups was aimed at studying if choice

of gesture type is influenced by verb meaning (see discussion in [20] and [10]). E. A. Grishina pointed to the fact that prefix type (manner or path) plays a role in gestures illustrating motion verbs; one of key features in gestures in the context of motion verbs is viewpoint or perspective. We tested this hypothesis for head gestures and analyzed types of head gestures accompanying the motion verbs.

The article is organized as follows: part 2 describes annotation methods for each of three channels (vocal, manual and cephalic). Part 3 presents results for the vocal channel and relates them to head and hand gestures considering different types of verbs. Part 4 discusses general findings of the research.

2 Data and annotation methods

2.1 Motion verbs

Motion verbs were chosen from all the elementary discursive units (EDUs) as described in chapter [1]. Table 1 shows numbers of motion verbs and their percentage to the total number of EDUs (the top line of the table lists participants according to their role and number of recording).

	04N	04C	04R	22N	22C	22R	23N	23C	23R
Number of verbs	82	38	126	63	83	60	72	41	99
Number of EDUs	385	246	512	339	275	263	402	231	326
Percentage of motion verbs to EDUs	21.3%	15.5%	24.6%	18.6%	30.2%	22.8%	17.9%	17.8%	30.4%

Table 1: Number of motion verbs and EDUs for each participant

Further, the verbs were divided into four semantic groups: 1) relocation (core motion verbs), 2) hand movements, 3) body movements and 4) inanimate referents movements. Additionally, they were marked for recording stage (monologue or dialogue), being prefixed or unprefixed, and the speaker's role (Narrator, Reteller, or Commentator).

2.2 Matching head and hand gestures to motion verbs

Gestures were divided into four functional types: depictive, pragmatic, pointing and beats [14]. Since there were few beat gestures, they were combined with pragmatics. For head gestures, there was added regulator as the fifth type [6, 7].

Manual gestures were regarded as corresponding to the motion verb if they were synchronized with the word. For hand gestures the key criterion was overlapping of stroke or hold phase with the word; the gesture was marked as matching if the verb appeared on preparation or retraction of the gesture, and the gesture stroke overlapped with actants of the verb. Gesturer's viewpoint was annotated only for depictive and pointing gestures, and only for those of them which illustrated the story itself [27, 19] (and not the speaker's stance or the process of communication).

In the cephalic channel, gestures were also considered if they (at least partially) coincided with the verb. Depictive gestures were additionally marked as pantomime (C-VPT) or showing movement direction (O-VPT) [10, 21]. There can be more than one cephalic gesture per word, so single gestures and combinations of 2–3 gestures per verb were considered separately.

All gestures were annotated using ELAN software (<https://archive.mpi.nl/tla/elan>).

Fig. [1] shows a C-VPT (character viewpoint) depictive manual gesture and cephalic depictive pantomime, and fig. [2] presents O-VPT depictive manual gesture, cephalic depictive direction-related.



Fig. 1: C-VPT gesture: *Beret sebe i stavit vsju korzinu* ‘He takes the whole basket and puts it’



Fig. 2: O-VPT gesture: *Po etoj lestnitse lazit fermer* ‘The farmer goes up and down the ladder’

3 Results

3.1 Vocal channel and motion verbs

3.1.1 General results for semantic groups of verbs

As shown by the analysis of the verb distribution, narrators use mostly movement verbs in the monologue and much less of them appear in the dialogue. The Retellers in recordings #4 and #23 actually have two retellings: a short one in the dialogue part of the recording (trying to memorize the story), and a more elaborated one in their retelling itself (see Fig. 3). Commentators take part only in dialogue.

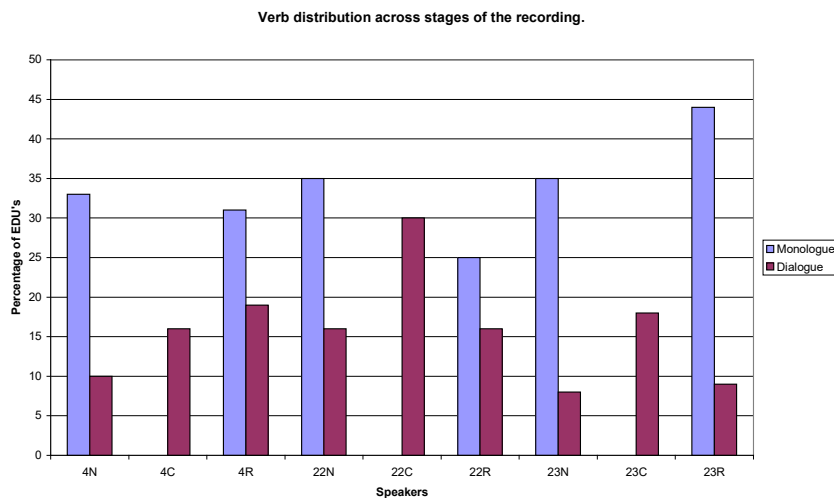


Fig. 3: Verb distribution across stages of the recording

The difference reveals distinction between monologue and dialogue: the verbs are mentioned when the referents are discussed and not to describe the plot.

The distribution of verbs into four semantic groups is shown in table 2:

Type of verb / Speakers	4N	4C	4R	22N	22C	22R	23N	23C	23R
Relocation verbs	46	21	77	36	49	36	39	16	56
Hand movements	30	12	44	22	33	21	27	24	32
Body movements	6	2	2	2	0	0	1	0	5
Inanimate object movements	3	3	5	1	4	3	5	2	6

Table 2: Distribution of verbs by semantic groups

As shown in table 2, every speaker used relocation verbs significantly more often than other motion verbs, which can be explained by the original design of the study (the story describes a lot of the characters' relocations). Hand movement verbs are the second most common: the characters manipulate baskets, pears, a racket etc. There are few verbs in the last two groups, so for further analysis we will focus on relocation and hand movement verbs.

The distribution of prefixed and non-prefixed verbs across all recordings is presented in Fig. 4.

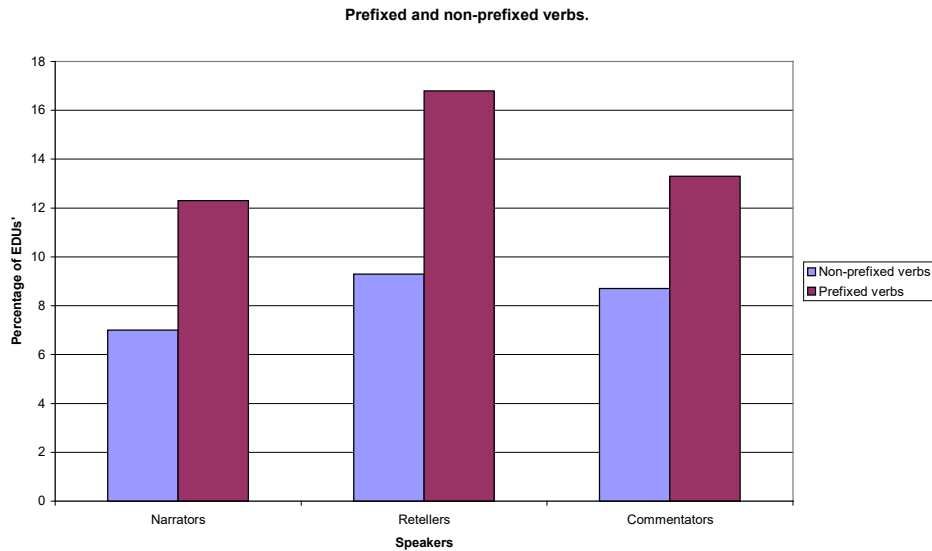


Fig. 4. Prefixed and non-prefixed verbs

Fig. 4 shows that there was no difference between the participants and part of the recording considering prefixed vs non prefixed verbs.

Prefixes were divided according to their type into vector (*v-*, *do-*, *pod-*, etc.) and route (*raz-*) ones.

3.1.2 Participants' roles and their gestures

In general, the Reteller tends to repeat verbs used by the Narrator and the Commentator, but there were also some differences.

So, in all the analyzed recordings, the Retellers did not use expressive, judgemental or colloquial words talking about the film, while those who saw the film (Narrator and Commentator), sometimes described the story in a colloquial way (using words as *ulepětyvaet* 'sneaks away', *umatyvaet* 'winds up' instead of neutral *uezzhaet* 'goes away'), see the same episode from N's and R's monologues in (1, 2) and (3, 4), respectively:

- (1) 22N, extract

N-vE055	On stavit tuda jetot /bagazhnik 'He puts there this rack'
N-vE056	i prosto ulepětyvaet! , 'And just sneaks away!'

- (2) 22R, extract

R-vE129	znachit on-n (ə 0.22) stavit sebe-e — 'So, he puts himself'
R-vE130	(ʔ 0.44) (pered –ruľēm, 'in front of the handlebar'
R-vE131	ja tak –ponjalā,) 'as I understood'
R-vE132	— etu –korzinu-u (ʏ 0.36) s= polnuju /gruʃh, 'This basket full of pears'
R-vN034	(ʏ 0.32)
R-vE133	(ə 0.28) ʌʔi-i (ʔ 0.13) (0.14) uezzhaet znachit ot etogo v-v= /sadovnika, 'and goes away from this gardener'

(3) 23N, extract

N-vE071	on \stavit eti/↑ gr <u>u</u> shi, ‘he puts these pears’
	...
N-vE073	i u= == ‘and...’
pN-035	(0.26)
N-vE074	i \u matyvaet ochen’ bystro. ‘and winds up very fast’

(4) 23R, extract

R-vE205	— on etu korzину tuda /→st <u>a</u> vit,, ‘He puts there this basket’
	...
R-vE210	i koroche \u ezzhaet . ‘and goes away, anyway’

This can be explained by the fact that N and C saw the film and had their personal assessment of events, while R, who did not see the film, consistently remained neutral and cautious.

Those of the verbs used by R and not repeating N and C, were mostly with vector prefixes *pod-*, *s-*, *po-*. These R’s verbs were either synonyms for N’s and/or C’s in the same context (for example, in recording 22, R says about a falling hat *sletaet* ‘flies off’, while N said *padaet* ‘falls down’), or interim actions which R explicitly reconstructed by himself, and N did not find it necessary to mention them having seen the film. Thus, in #4 R mentions several times that the boy *podkhodit* ‘goes up’ to the basket before taking the pears, while in a similar episode (5) N does not consider it necessary to specify where the boy is going but just says that he takes the whole basket, which is more relevant for the narrative).

(5) 4N

N-vE082	snachala mal'chik hochet vzjat' /odnu ↑grushu, ‘first, the boy wants to take one pear’
N-vN023	(η 0.49)
N-vE083	/potom-m ponimaet chto-o (? 0.45) nichto emu ne /groz <u>i</u> t, ‘then understands that he is perfectly safe’
N-vE084	dovol'no bospalevno berjot celuju –↑korz <u>i</u> nu, ‘quite unobtrusively takes the whole basket’

(6) 4R

R-vE372	podkhodit == ‘goes up’
pR-281	(0.13)
	...
R-vE374	on / podkhodit (η 0.22) (0.13) (ə 0.31) k /korz <u>i</u> ne, ‘he goes up to the basket’
R-vE375	berët ottuda odnu /gr <u>u</u> shu, ‘takes one pear from there’
pR-282	(0.06)
R-vE376	potom /pered <u>u</u> myvaet, ‘then changes his mind’
R-vN048	(η 0.38)
R-vE377	kladët eë /obratno, ‘puts it back’
R-vE378	berët tseluju \korz <u>i</u> nu. ‘takes the whole basket’

3.2 Manual gestures

3.2.1 Types of manual gestures in monologues and dialogues

Although there are some differences between the participants, in dialogues (hereinafter d) compared to monologues (hereinafter m), the number of gestures with motion verbs decreases (χ -square, $p < 0.001$), see Fig. 5. The difference between the stages of recording can be attributed to the fact that the participants were asked to describe the film in as much detail as possible, so they used as many depictive devices including gestures as they could, while in dialogue they were more involved in interaction with others. It can be also interpreted as less involvement in the description of events or the lack of opportunities for coherent and detailed gestural illustrations that are in the monologue.

The differences between the participants are statistically insignificant.

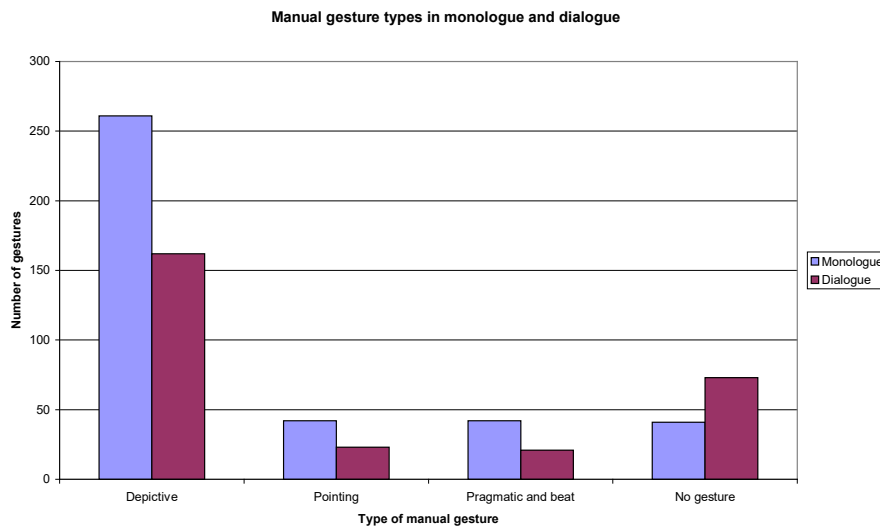


Fig. 5: Manual gesture types in monologue and dialogue

Additionally, Narrator used more C-VPT gestures than Reteller (see Fig. 6), $p < 0.05$. This fact has already been observed in [6] and is explained by Narrator’s personal experience of the film compared to Reteller’s who has only heard about the story.

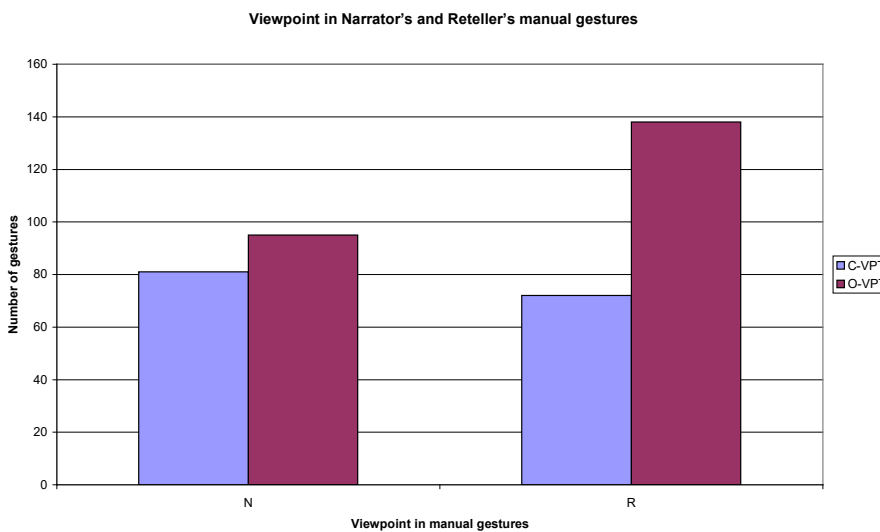


Fig. 6: Viewpoint in Narrator’s and Reteller’s manual gestures

There was no significant difference in the viewpoint depending on the stage of the recording.

3.2.2 Verb semantic type and manual gestures

Fig. 7 shows that there is a tendency for depictive gestures to be used with hand movement descriptions ($p < 0.05$) compared to core relocation verbs that are more often combined with pointing and pragmatic gestures.

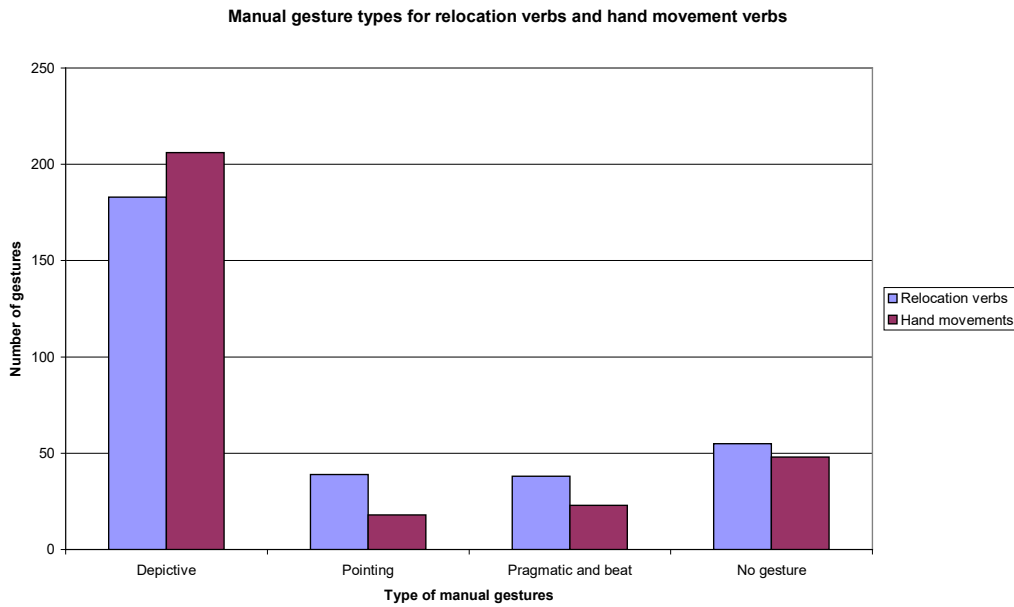


Fig. 7. Manual gesture types for relocation verbs and hand movement verbs

As follows from Fig. 8, for depictive and pointing gestures with motion verbs, the observer’s point of view (O-VPT) will occur more often, and for descriptions of hand movements C-VPT is prevalent ($p < 0.05$). However, there are examples when the motion verb is supported with pantomime using manual gestures; other factors can influence viewpoint too. Further analysis showed that it is more noticeable in monologue parts.

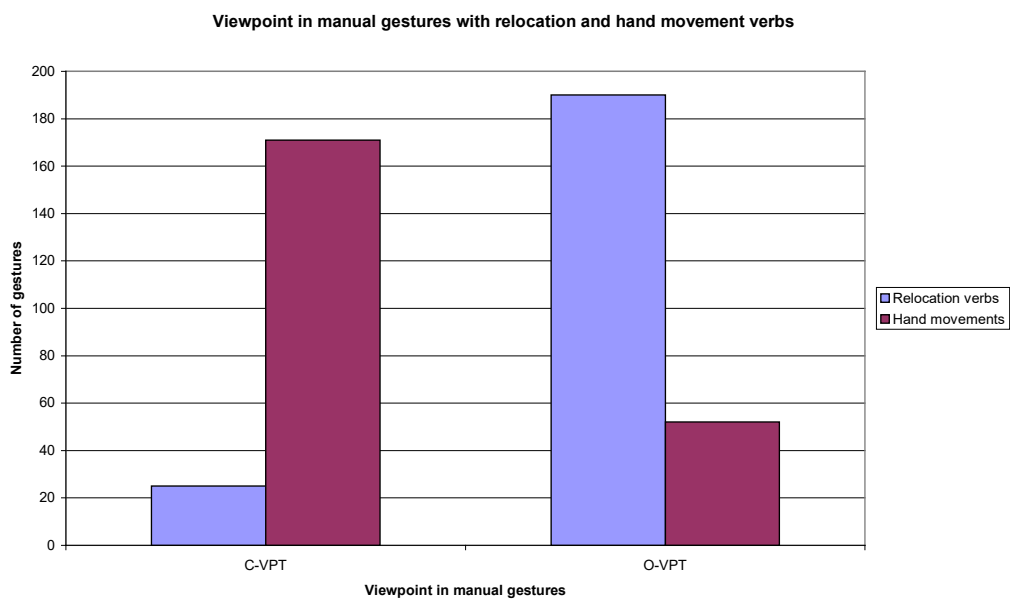


Fig. 8: Viewpoint in manual gestures with relocation and hand movement verbs

For prefixed and non prefixed verbs there was no significant difference in manual gestures.

3.2.3 Semantic type of the verb and manual gestures

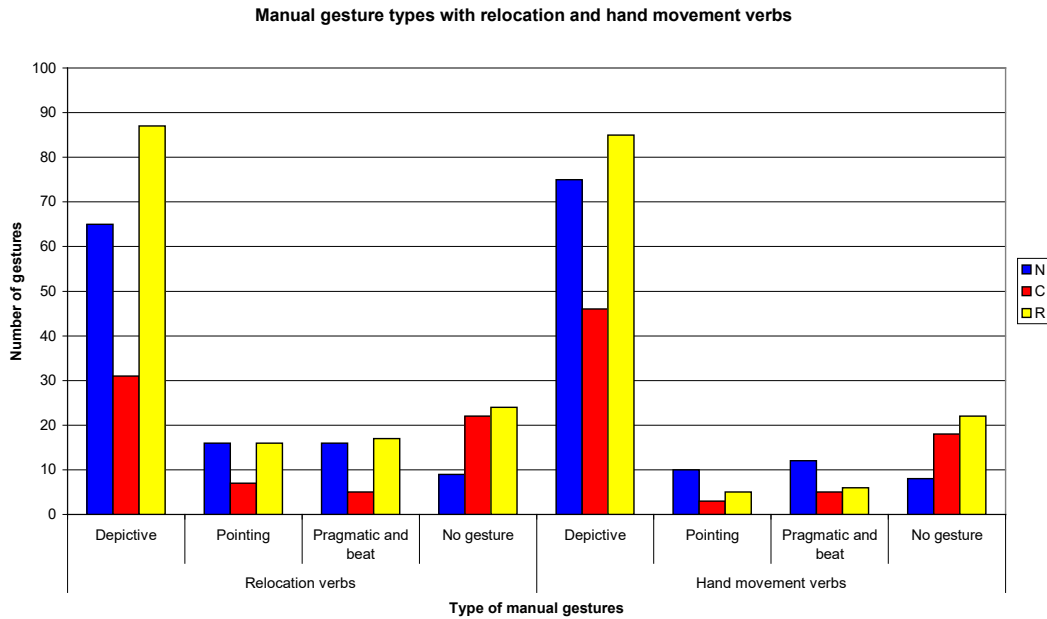


Fig. 9: Manual gesture types with relocation and hand movement verbs considering speaker’s role

As Fig. 9 shows, both the speaker's role and verb type influence hand gestures. Narrator and Reteller who have monological parts use more pointing and pragmatic gestures depicting relocations than Commentator does (χ -square, $p < 0.05$). This can be interpreted as follows: monological parts involve diverse gesticulation, and for motion verbs it is less important if the speaker has personally witnessed the story. However, hand movements descriptions are supposed to be easily illustrated by depictive gestures, and this can be done even for short discourse segments (as those which Commentator has), while for visual tracking of referents’ position monologue parts are more convenient.

Additional analysis showed that there is a clear tendency to change either the word or the gesture when the event description is repeated.

3.3 Head gestures

3.3.1 General distribution of gesture types

	4N/ m	4N/ d	22N/ m	22N/ d	23N/ m	23N/ d	4R/ m	4R/ d	22R/ m	22R/ d	23R/ m	23R/ d	4C/ d	22C/ d	23C/ d
Single gestures, %	78	94	78	93	83	82	91	89	91	77	73	55	90	87	75
Gesture combinations, %	22	6	22	7	17	18	9	11	9	23	27	45	10	13	25
Percentage of EDUs with motion verbs to all EDUs	33	10	35	6	35	8	31	19	25	16	44	9	16	30	18

Table 3: Head gesture across the speakers

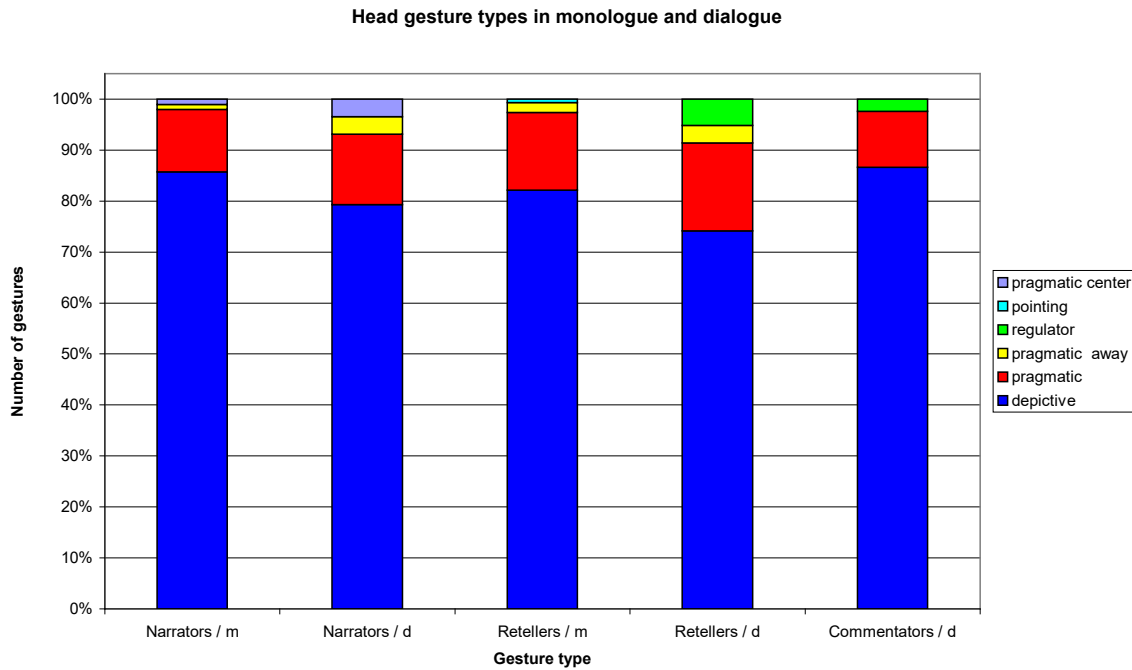


Fig. 10: Single gesture types

The difference between the participants in the number of gestures in monologue and dialogue (Fig. 10) depends on the number of EDUs at these stages of the recording (see Table 3). In dialogue, both Narrators and Retellers have their depictive gesture percentage reduced in favor of other gesture types.

Considering separately the distribution of types of depictive gestures (Fig. 11), we single out: 1) depictive showing the direction of movement, 2) depictive showing manner of movement, 3) combinations of these two types, where the first gesture is more prominent 4) depictive, where the type cannot be determined.

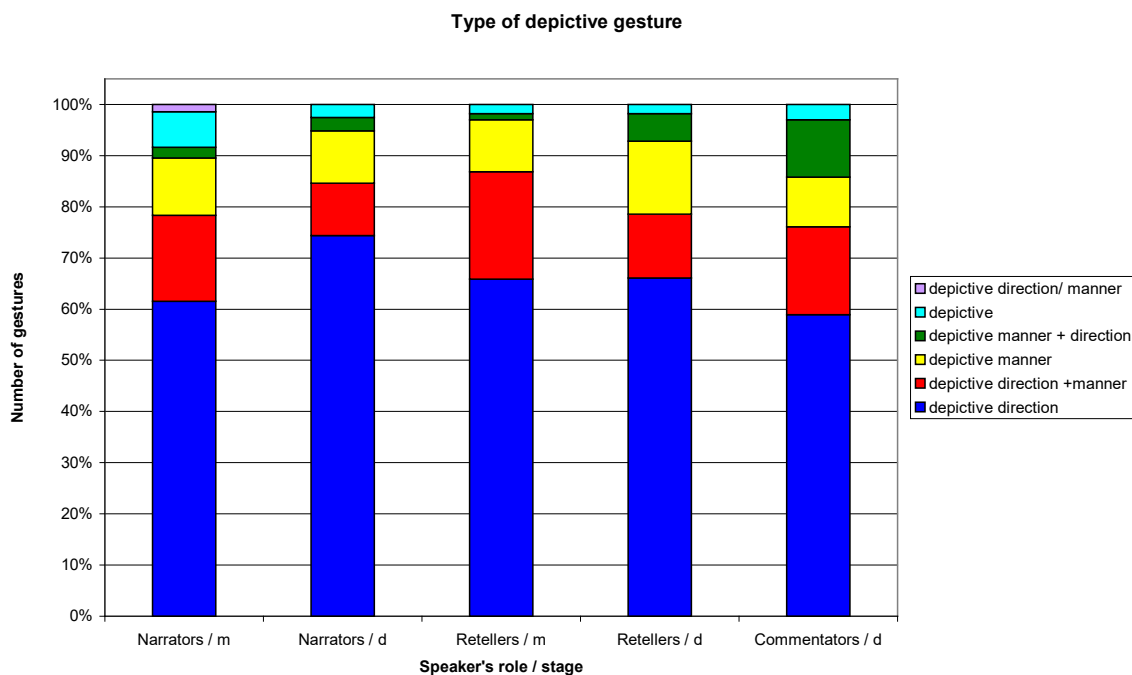


Fig. 11: Types of depictive head gestures

Depictive gestures showing the direction predominate in monologues and dialogues (see Fig. 11). They indicate the O-VPT, according to Grishina [10]. In monologues, narrators have more variability in the types of depictive gestures than in dialogues.

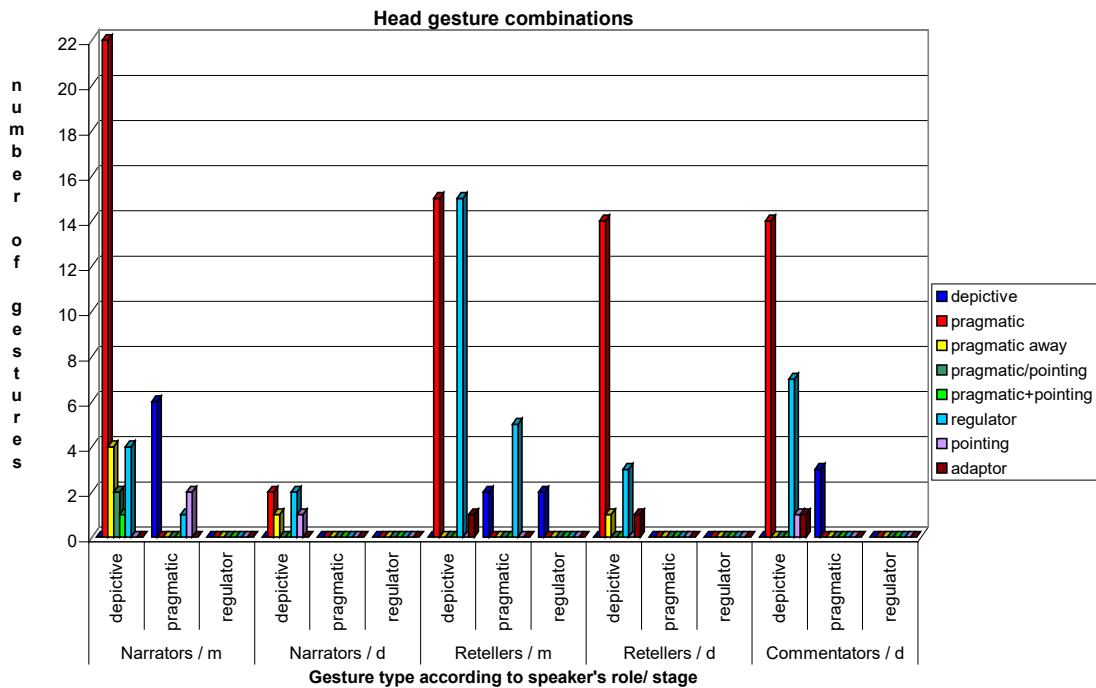


Fig. 12: Head gesture combinations and gesture types

All participants in the monologue used predominantly a cluster combining a depictive gesture with a pragmatic one (Fig. 12). In the dialogue, this type of combination becomes the main one. Such a cluster means that with the help of his pragmatic gesture, the speaker draws the listener's attention to the depictive one. The second most frequent is a cluster of depictive and regulator gestures, where the speaker uses a regulator one to test the listener's reaction. That's the way the Narrator checks if the Commentator agrees with his version of events, and the Reteller verifies if his telling is understandable for the Listener. The use of gesture combination types by Narrators differs from that of participants in other roles (Fig. 12): 1) the depictive gestures are halved in favor of pragmatic ones, and 2) pointing and regulator gestures appear.

3.3.2 Distribution of gesture types with different semantic groups of verbs

According to the same parameters as in the previous section, we describe two semantic groups separately: 1) verbs of relocation, and 2) verbs with hand movement semantics. In each category, we analyze verbs with and without prefixes.

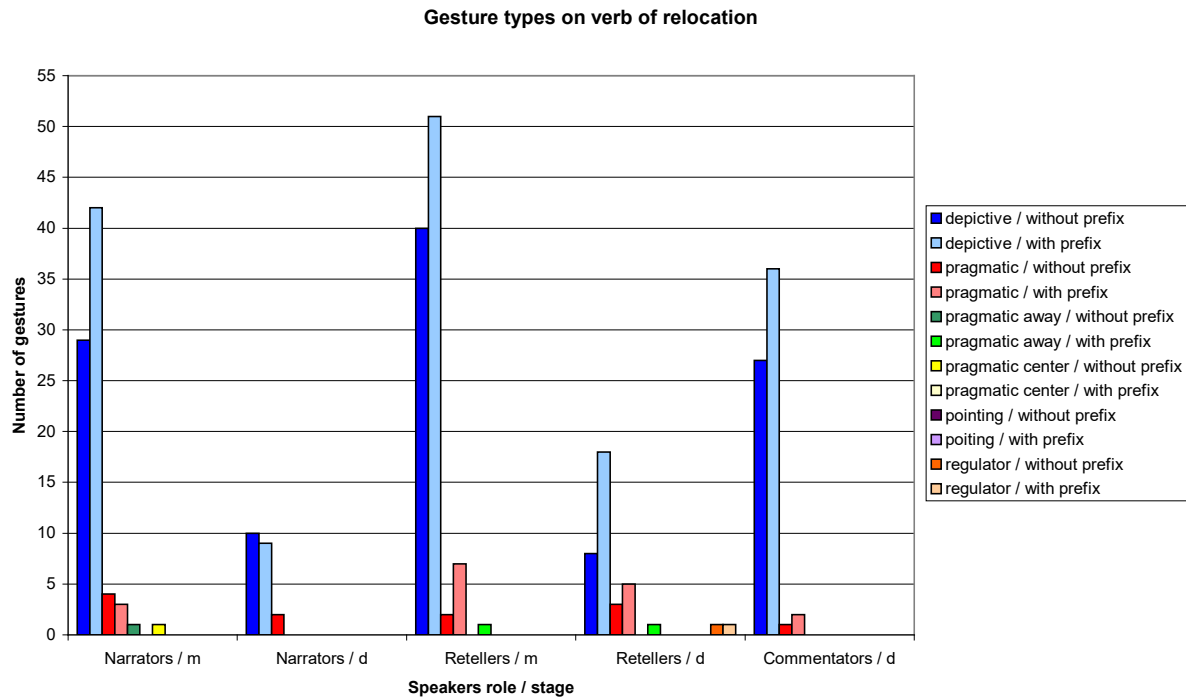


Fig. 13: Gesture types on relocation verbs

Both for monologues and dialogues (Fig. 13) with prefixed and non-prefixed relocation verbs, depictive gestures, which are similar to hand gestures, prevail (χ -square, $p < 0.001$), and pragmatic gestures are much less frequent.

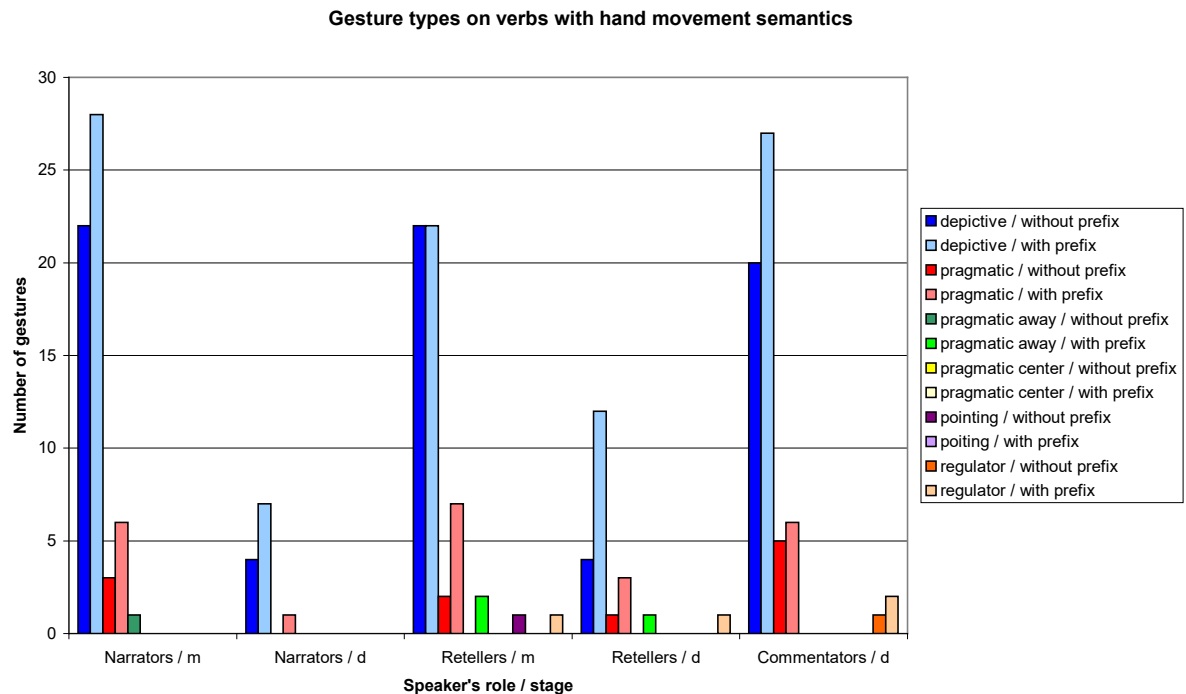


Fig. 14: Gesture types on verbs with hand movement semantics

With the same predominance of depictive gestures (Fig. 14), for prefixed verbs the number of pragmatic and regulator gestures increases (χ -square, $p < 0.005$), which draws the listener's attention to their hands.

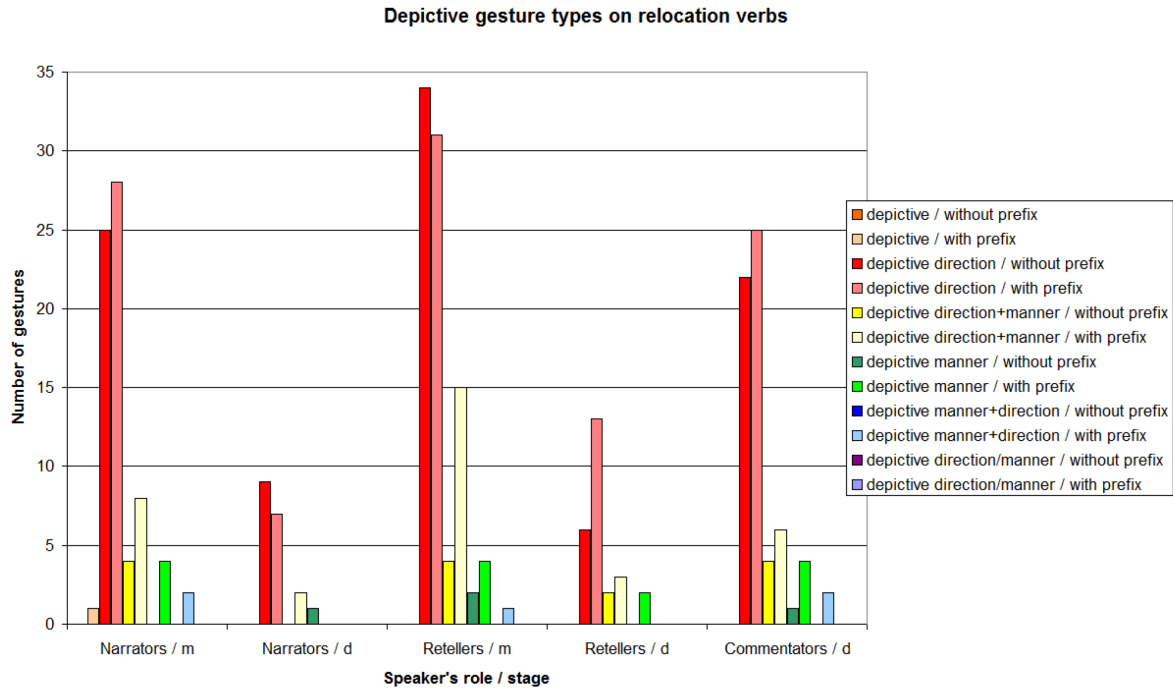


Fig. 15: Depictive gesture types on relocation verbs

Regardless of the prefixes, the type of depictive showing direction prevails (Fig. 15), which indicates the O-VPT. However, in this semantic group on verbs with prefixes, the number of depictive gestures increases, combining both functions: direction and pantomime (manner), as well as depictive type, from which we can conclude that the semantics of prefixes affects the change in the depictive gesture type (χ -square, $p < 0.002$), and also changes the O-VPT to the C-VPT, more often in a monologue, but the C in the dialogue.

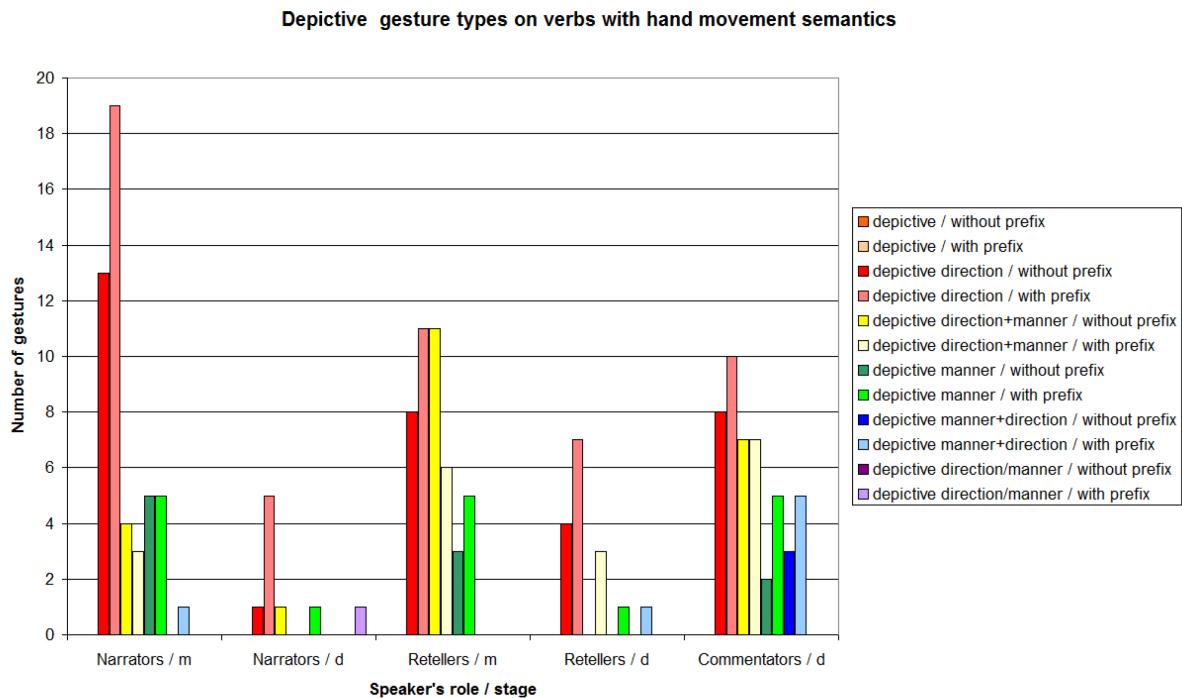


Fig. 16: Depictive gesture types on verbs with hand movement semantics

Verbs with the semantics of hand movement increase (χ -square, $p < 0.001$) in the number of depictive gestures of the movement manner and combining the function of direction and manner (Fig. 16). For some participants, the number of depictive gestures showing direction is reduced so much that they are no longer the predominant type. There is more variety in prefixed verbs.

3.4 Gesture clusters in the hands and head

In multichannel analysis, we use clustering as a method [6], for which in this research a cephalic channel is selected as the main one. After this selection it is sequentially checked whether the gesture forms a cluster with manual channel according to time in direction (Fig. 17a) or by functions (Fig. 17 b):



Fig. 17 a, b: Examples of depictive clusters
 a) 23N — *zabiraetsya* (climbs) b) 4R — *povorachivaet* (turns)

As we see (Fig. 14), Narrators in monologue and dialogue more often use the separate gestures in each channels, but Retellers in monologue use the clusters of two channels:

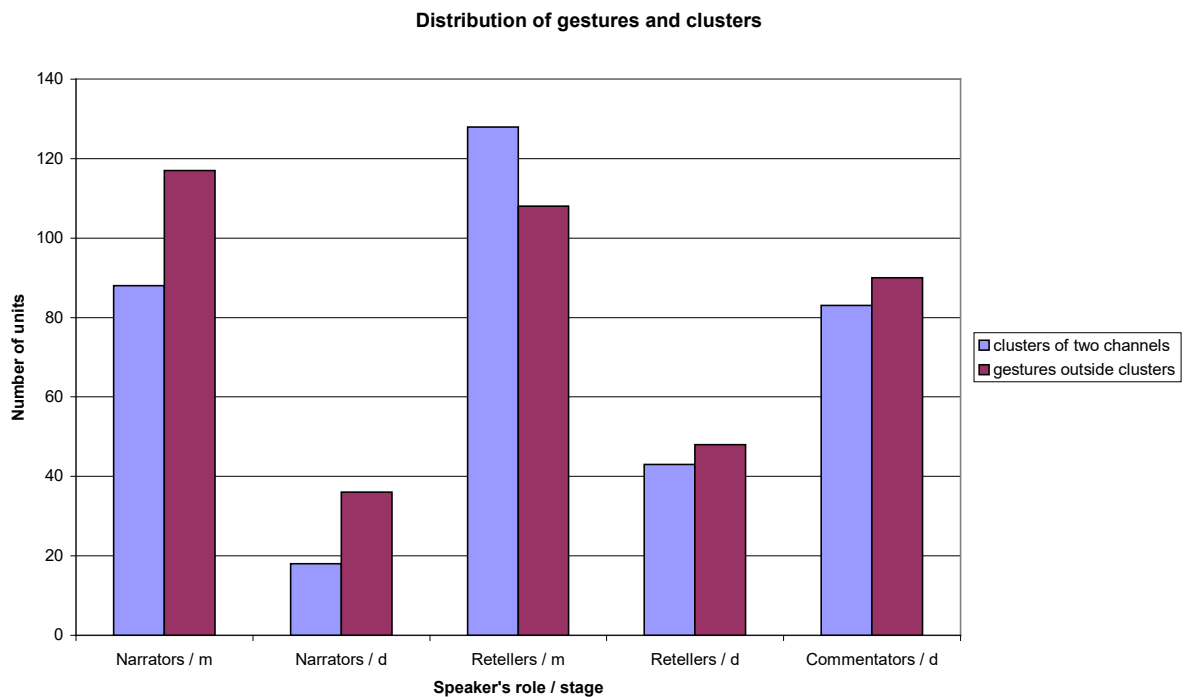


Fig. 18: Distribution of clusters and gestures

When applying this method to the entire set of considered verbs, we revealed the following patterns (Fig. 14):

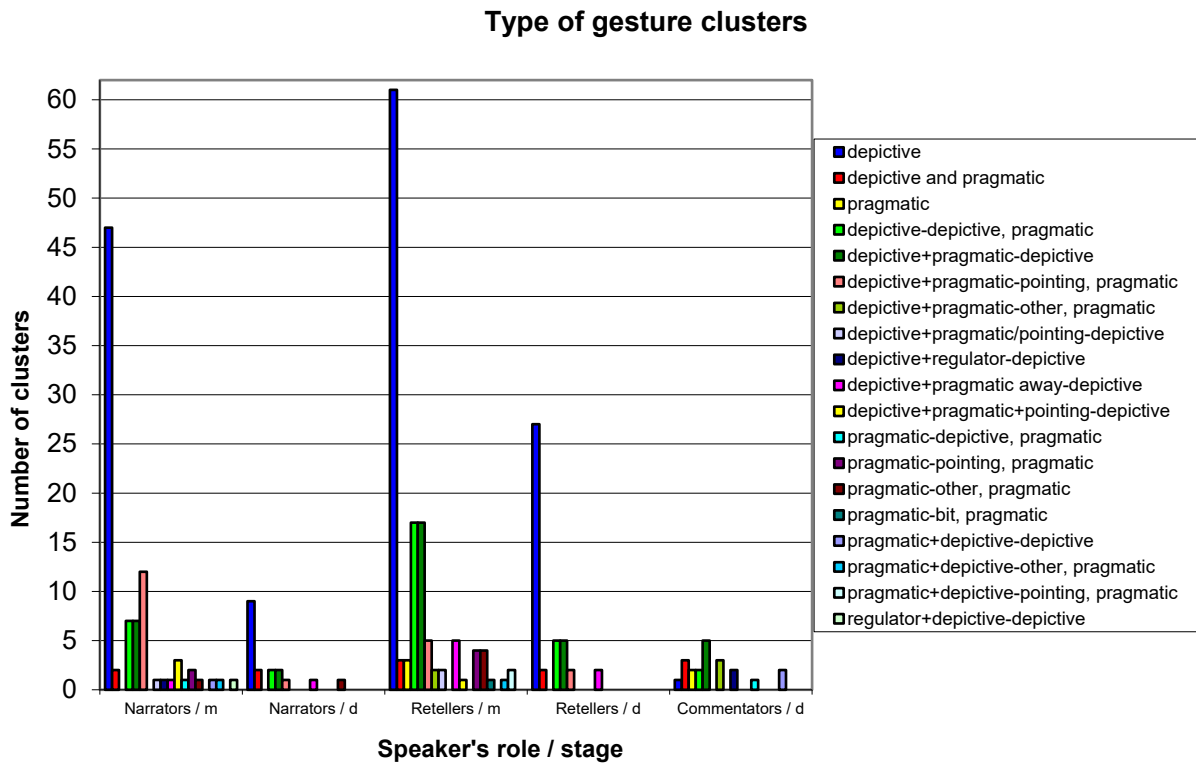


Fig. 19: Cluster types of hand and head gestures

In the monologues of Narrators (Fig. 19), depictive clusters predominate (more than 50%). For Retellers, clusters of depictives combined with pragmatic gestures increase in number. It can be explained by the fact that they have not seen the film.

In narrator dialogues, the percentage of clusters with pragmatic and depictive gestures increases so much that Narrator #22 has no depictive gestures. This is because the role of the Narrator is yet realized, and he can afford to relax and participate in communication without describing anything else and illustrating it by depictive gestures. Regardless of the role in the recording, everyone has a greater variety of gesture clusters in monologues than in dialogues.

4 Conclusion

Results show that choice in verbal and kinetic channels differ depending on the speaker's role and type of discourse.

In the verbal channel: Narrators and Commentators who saw the film at the preliminary stage, used more expressive and judgment verbs compared to Retellers. In addition, Retellers completed the narration with verbs with vector prefixes which were not used in the original story, meanwhile explicating the whole chain of events for himself.

In the manual channel: Relocation verbs are more often illustrated by observer viewpoint gestures compared to hand movement verbs, which attract character viewpoint gestures; it is even more prominent for monologues.

Hand movements are also more often illustrated by depictive gestures, but the tendency is less pronounced. It means that manual gestures clearly depend on the meaning of the word they accompany.

Pointing and pragmatic gestures can be used to mark character's relocation in longer parts of discourse by both who witnessed the story and who knew it second-hand, but referring to specific actions in a story needs personal experience and is more expected in monologues.

Verb prefixes do not seem to influence choice of manual gestures.

In the cephalic channel: on the verbs of motion, depictive gestures predominate, most often showing the direction and expressing the point of view of the observer (O-VPT) (χ -square, $p < 0.001$). In the relocation verb semantic group on verbs with prefixes, the number of depictive gestures, combining both functions: direction and manner of movement, increases (χ -square, $p < 0.002$), as well as depictive showing manner of motion, from which we can conclude that the semantics of prefixes affects the change in the depictive gesture type. On verbs with the semantics of hand movements, the depictive gestures' percentage of the (pantomime) type (χ -square, $p < 0.001$) increases, both in combination with direction gestures and on their own, which indicates the role of the character (C-VPT), in some gestures points of view overlap. As a result of the analysis, it was revealed that Narrators and Retellers repeated the same motion verbs when talking about the same episodes but used a different concomitant head gesture (either of a different functional type, or performing a different direction, amplitude and/or movement type (tilt, turn, etc.). Thus, when the Reteller repeated the verbs previously used by the Narrator and the Commentator, the Reteller's accompanying gestural behavior differed from the first two speakers.

In gesture clusters from two channels, depictive ones prevail (χ -square, $p < 0.001$). Narrators (three times more in 22 entries) use the gesture clusters more often in monologue than in dialogue, and Narrator in 22 entries, on the contrary, 1.5 times in dialogue. Narrators more often resort to clusters with pragmatic gestures combined with depictive instead of purely depictive ones (χ -square, $p < 0.001$) since they have not seen the film and are not so confident in their story.

The overall analysis showed that the relationship between semantic type of the verb, non-verbal behavior and the speaker's role depends on a complex set of factors. In the verbal channel, verbs tend to be repeated among all participants, however, Narrators more often used verbs with vector prefixes, completing the sequence of events by themselves. In non-verbal channels, the prefixes did not affect the choice of gesture; the latter was largely guided by the stage (monologue vs. dialogue), the speaker's role and the semantic type of the aligned verb (the Narrators more often accompanied hand movements descriptions by pointing and pragmatic gestures). The semantics of the verb and the situation (manipulation vs. displacement) also turns out to be tied to the choice of point of view (C-VPT vs. O-VPT), but to a lesser extent to the choice of gesture functional type.

References

- [1] Apresyan Ju. D. Selected works. Lexical semantics. Synonymous means of language (2nd ed.) [Izbrannye trudy. Lexicheskaja semantika. Sinonimicheskie sredstva yazyka (2-ye izd.)]. — Moscow: Nauka, 1995. — Vol. 1. M., 1995.
- [2] Apresyan Ju. D. (1990), Lexicographical portrait of the verb VYJTI [Lexikograficheskij portret glagola VYJTI], Topics in cybernetics. The language of logics and the logics of language [Voprosy kibernetiki. Jazyk logiki i logika jazyka], pp. 70–96.
- [3] Bourdin Philippe, On Goal-bias across languages: modal, configurational and orientational parameters // Proceedings of LP'96. Typology: prototypes, item orderings and universals. — Karolinum — Prague: Charles University Press, 1997. — P. 185–218.
- [4] Budennaya E. V., Nikolaeva Ju. V., Evdokimova A. A., Sukhova N. V. (2020), Referential phenomena in speaker's kinetic channels, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2020" [Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii "Dialog 2020"], RGGU, pp. 133-146.
- [5] Chafe W. L. (ed.). The pear stories: Cognitive, cultural, and linguistic aspects of narrative production (Advances in Discourse Processes. — Vol. III). — Norwood, N.J.: Ablex, 1980.
- [6] Evdokimova A. A., Nikolaeva Ju. V. (2021), Character and observer viewpoint: their projection in manual and cephalic gestures [Tochka zreniya personazha i tochka zreniya nabludatelja: ikh realizatsija v zhestakh golovy i ruk], Cognitive studies [Kognitivnye issledovaniya yazyka], Vol. 1 (44), pp.182–189.
- [7] Evdokimova A.A. (2020), New types of pragmatic head gestures - Pragmatic center and Pragmatic away [Novye tipy pragmaticheskikh zhestov golovy — Pragmatic center i Pragmatic away], Linguistics and methods of teaching foreign languages [Lingvistika i metodika prepodavanija inostrannykh jazykov], Vol. 1 (12), pp. 136–148.
- [8] Ferm, L. (1990) Expression of direction with prefixed verbs of displacement in modern Russian. On the issue of prefix-prepositional determinism [Vyrazhenie napravleniya pri pristavochnyh glagolah peremetsheniya v sovremennom ruskom yazyke. K voprosu prefixalno-predlozhnogo determinizma]. Uppsala.
- [9] Fillmore Ch., Deictic categories in the semantics of come // Foundations of language. — 1966. — Vol. 2. — P. 219–227.

- [10] Grishina E. A. (2017), Russian gesticulation from a linguistic perspective (corpus studies) [Russkaja gestikuljatsija s lingvisticheskoj točki zrenija (korpusnyje issledovanija)]. — Moscow: LRC Publishing House.
- [11] Ikegami Yoshihiko., ‘Source’ vs. ‘goal’: A case of linguistic dissymmetry // Concept of case. — Tübingen: Gunter Narr Verlag, 1987. — P. 122–146.
- [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. — 2018. — Vol. 15(2). — P. 191–200.
- [13] Korotaev N. A. (2019), Hesitation pauses in narratives and conversations: A quantitative comparison [Pausy xesitacii v rasskaze i v razgovore], Proceedings of the Corpus Linguistics Conference 2019 [Trudy mezhdunarodnoj konferencii “Korpusnaja lingvistika-2019”], Saint Petersburg, pp. 48–54.
- [14] Litvinenko A. O., Kibrik A. A., Fedorova O. V., Nikolaeva J. V., Annotating hand movements in multichannel discourse: Gestures, adaptors and manual postures. // Russian Journal of Cognitive Science. — 2018. — Vol. 5 (2). — P. 4–17.
- [15] Lyons John. Semantics. — Cambridge: Cambridge University Press, 1977. — Vol.2.
- [16] Maysak T.A (1999), Asymmetry of valencies of Russian motion verbs: “Russian option” [Assimetrija valentnostej u glagolov dvizhenija: “russkiy variant”], Computational Linguistics and Intellectual Technologies; Proceedings of the international workshop “Dialog 1999” [Komp’yuternaya Lingvistika i Intellekтуal’nye Tekhnologii: Trudy Mezhdunarodnogo Seminara “Dialog 1999” po kompyuternoy lingvistike i eye prilozheniyam], Tarusa, pp. 172–185.
- [17] Maysak T. A., Rakhilina E.V. (1999), Semantics and statistics: the verb TO GO (IDTI) amid other motion verbs [Semantika i statistika: glagol IDTI na fone drugih glagolov dvizheniya], Logical analysis of language: languages of dynamical world [Logichesky analiz yazyka: Yazyki dinamicheskogo mira], Dubna, pp. 53–66.
- [18] Miller George A., Johnson-Laird Philip N. Language and perception. — Cambridge: Cambridge University Press, 1976.
- [19] Nicoladis Elena, Shirazi Safi, Hand Preference in Adults’ Referential Gestures during Storytelling: Testing for Effects of Bilingualism, Language Ability, Sex and Age // Symmetry. — 2021. — Vol. 13(10):1776.
- [20] Nikolaeva Ju. V. (2013) Illustrative gestures in Russian discourse [Illyustrativnye gesty v russkom diskurse]. (Dissertation... candidate of philological science) [(Diss. ... kand. filol. nauk)]. M.
- [21] Nikolaeva J., Budennaya E., Evdokimova A., Evidential components in multimodal communication // CEUR Workshop Proceedings. — 2021. — Vol. 2852 of Linguistic Forum 2020: Language and Artificial Intelligence.
- [22] Paducheva, E.V. (1996) Research on semantics [Semanticheskie issledovaniya]. M.
- [23] Paducheva E.V., Krylov S.A. (1983), Deixis: general theoretical and practical aspects [Deixis: obtsetoreticheskiye i pragmaticheskiye aspekty], Language production in the aspect of linguistic pragmatics. [Jazykovaja dejatel’nost v aspekte lingvisticheskoj pragmatiki], Moscow, pp. 25–96.
- [24] Podlesskaja V. I., Korotaev N. A., Mazurina S. I. (2019), A corpus study of self-repairs in Russian monologues and dialogues [Samoispravlenija govornjashchego v russkom monologicheskom i dialogicheskom diskurse: opyt korpusnogo issledovanija], Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference “Dialog 2020” [Komp’yuternaya Lingvistika i Intellekтуal’nye Tekhnologii: Trudy Mezhdunarodnoj Konferencii “Dialog 2020”], RGGU, pp. 508–522.
- [25] Rakhilina, E.V. (1990) Semantics or syntax? To the analysis of general questions in Russian [Semantika ili sintaxis? K analizu chastnyh voprosov v russkom yazyke]. Munchen: Sagner.
- [26] Rakhilina E.V. (1998), GO or COME [IDTI ili PROHODIT’?], Computational Linguistics and Intellectual Technologies; Proceedings of the international workshop “Dialog 1998” [Komp’yuternaya Lingvistika i Intellekтуal’nye Tekhnologii: Trudy Mezhdunarodnogo Seminara “Dialog 1998” po kompyuternoy lingvistike i eye prilozheniyam], Kazan, pp. 303–307.
- [27] Rohrer Patrick L., Florit-Pons Júlia et al., Children Use Non-referential Gestures in Narrative Speech to Mark Discourse Elements Which Update Common Ground. // Frontiers in Psychology. — 2022. — Vol. 12:661339.
- [28] Rozhansky, Ph.I. (1992) Verbal constructions with spatial arguments [Glagolnye konstrukcii s prostranstvennymi argumentami]. (Dissertation... candidate of philological science) [(Diss. ... kand. filol. nauk.)] M.
- [29] Stefanowitsch Anatol, Rohde Ada. The goal-bias in the encoding of motion events. // Paper presented at the International Workshop “Motivation in grammar”. — University of Hamburg, July 7-9, 1999.
- [30] Talmy Leonard, Semantics and syntax of motion // Syntax and semantics. — New York: Academic Press, 1975. — Vol. 4. — P. 181–238.
- [31] Talmy Leonard. (1985). Lexicalization patterns: semantic structure in lexical forms // Language typology and syntactic description. — Cambridge: Cambridge University Press, 1985. — Vol. 3. — P. 36–149.
- [32] Ungerer Friedrich, Schmidt Hans-Jörg. An introduction to cognitive linguistics. — London: Longman, 1996.
- [33] Verspoor Marjolijn, Dirven René, Putting concepts together: Syntax // Cognitive exploration into language and linguistics. — Amsterdam and Philadelphia: Benjamins, 1999. — P. 75–100.
- [34] Zhukova M. V. (2020), Co-speech gesticulation accompanying verbs of motion in Russian. — Moscow, National Research University Higher School of Economics. Student thesis.

Lightweight and accurate system for entity extraction and linking

Evseev D. A.

Moscow Institute of Physics and Technology
Dolgoprudny, Russia
dmitrij.euseew@yandex.ru

Abstract

Entity extraction and linking components in dialogue assistants should meet the requirements of low resource consumption and high accuracy. In this paper we present lightweight system which extracts entity mentions from the text and finds corresponding Wikidata ids and Wikipedia pages links. Entity extraction and linking is performed into the following steps: extraction of entity substrings from the text, retrieval of candidate entities from Wikidata knowledge base and entity disambiguation. Entity extraction is based on RoBERTa-tiny model for token classification. Extracted substrings are classified into 42 fine-grained tags for filtering of candidate entities. Candidate entities are ranked by number of connections of candidate entities in the text in Wikidata knowledge graph. The proposed system outperforms on WNED-WIKI other lightweight solutions, such as REL and OpenTapioca. The system supports easy adding new Wikidata entities to the database and using other knowledge bases for entity linking.

Keywords: entity extraction, entity linking, entity disambiguation, knowledge base
DOI: 10.28995/2075-7182-2022-21-176-184

Легкая и точная система для извлечения сущностей и связывания с базой знаний

Евсеев Д. А.

Московский физико-технический институт
Долгопрудный, Россия
dmitrij.euseew@yandex.ru

Аннотация

Компоненты для извлечения сущностей и связывания с базой знаний в диалоговом ассистенте должны отвечать таким требованиям, как низкое потребление памяти, а также высокая точность. В данной статье описывается система, которая извлекает сущности из текста и находит для них соответствующие ids в Wikidata и ссылки на страницы Википедии. Извлечение и связывание сущностей происходит в несколько этапов: извлечение подстрок с сущностями из текста, извлечение возможных сущностей из базы знаний Wikidata и устранение неоднозначности сущностей. Компонент для извлечения сущностей основан на RoBERTa-small для классификации токенов. Извлеченные подстроки классифицируются на 42 класса для фильтрации возможных сущностей. Возможные сущности в тексте сортируются по числу связей с использованием графа знаний Wikidata. Предлагаемая система превосходит на датасете WNED-WIKI другие системы с низким потреблением ресурсов, такие как REL и OpenTapioca. Система поддерживает добавление новых сущностей Wikidata в базу данных, а также использование других баз знаний для связывания сущностей.

Ключевые слова: извлечение сущностей, связывание сущностей с базой знаний, устранение неоднозначности сущностей, база знаний

1 Introduction

Entity Linking is the task of identifying an entity mention in unstructured text and establishing a link to an entry in a knowledge base (Sevgili et al., 2021). In dialogue assistants entity linking is a key

component for natural language understanding, because entities in the utterance can help to detect user’s intention to change the topic and facts from the knowledge base extracted for detected entities can be used for generation of meaningful response.

For parallel dialogue interaction with multiple users entity linking system in a dialogue assistant should be deployed in many replicas, so one of the requirements to EL system is low resource consumption. State-of-the-art entity linking systems are based on large pretrained Transformers (De Cao et al., 2020) or store entities inverted index in RAM (Wu et al., 2019). Lightweight solutions, which store entity embeddings in SQLite database (van Hulst et al., 2020) or use Wikidata (Vrandečić and Krötzsch, 2014) knowledge graph for entity disambiguation (Delpeuch, 2019), stored in Solr¹ index, show low accuracy of entity linking.

In this paper we present lightweight (which can be deployed on an average laptop or desktop machine and does not need much RAM and GPU) and fast entity linking system which can be used in dialogue assistants. The system consists of the following components: identifying entity mention in text, retrieve of candidate entities from the knowledge base, entity mention classifier by types and entity disambiguation using Wikidata knowledge graph and Wikipedia hyperlinks graph. RoBERTa-tiny (Liu et al., 2019) model is used for token classification into three classes: beginning of the entity mention, inside the entity mention and tokens which do not belong to any entity. Detected mentions are classified into 42 tags according to Wikidata entity types with another RoBERTa-tiny model. Candidate entities for the mentions are retrieved from the inverted index in SQLite database with FTS5 extension which supports full text search by entity mentions. For training of RoBERTa model we preprocessed Wikipedia pages with hyperlinks to obtain a dataset of paragraphs annotated with entity mentions and corresponding classes. After filtering we find connections of candidate entities for a mention with candidate entities for other mentions using the knowledge graph. The knowledge graph is stored in the same SQLite database as inverted index which is not loaded into RAM. The proposed system outperforms on WNED-WIKI (Petroni et al., 2020) OpenTapioca and REL. The system does not need pretrained entity embeddings which results in easy adding of new Wikidata entities into the database without need to retrain the models. The system supports entity linking over other knowledge bases provided that the tags of entity type classification model were mapped to knowledge base types.

2 Related work

TagME (Ferragina and Scaiella, 2011) is one of the first entity linking systems, which finds Wikipedia page links for entity mentions in text and uses Wikipedia hyperlinks graph for entity disambiguation. Further improvement of entity linking systems was connected with neural network architectures. In the work of (Ganea and Hofmann, 2017) candidate entities are ranked by bilinear form of entity embedding x_e and embeddings of tokens x_w of K-word local context $c = \{w_1, \dots, w_K\}$ (1):

$$\psi(e, c) = \sum_{w \in c} \beta(w) e_w^T B x_w, \quad (1)$$

Global disambiguation, exploiting document-level coherence of entities is performed with CRF-based model. In the system (Le and Titov, 2018) bilinear form is calculated between embeddings of pairs of entities for global disambiguation. In (Le and Titov, 2019) the dataset for training of the model (Le and Titov, 2018) was extended with unlabeled texts with extracted mentions. Candidate entities for the mentions were scored by collective agreement using Wikipedia hyperlinks graph and the entity with the highest score was considered as an answer. In REL (van Hulst et al., 2020) entity disambiguation is based on calculation of bilinear form between entity and context embeddings and entity embeddings for different mentions, the same as in (Le and Titov, 2018). REL system is lightweight because it uses SQLite database for storing entity embeddings. In the approach of (Martins et al., 2019) LSTM is used to extract entity mentions and obtain context embeddings.

In (Kolitsas et al., 2018) all possible n-grams in the sentence were considered as mentions. Entity disambiguation is performed by dot products of candidate entity embeddings and mention embeddings,

¹<https://solr.apache.org/>

obtained with LSTM with attention.

Every entity in the knowledge base has the type, (in Wikidata it is defined with the relation P31, "instance of", for example, <Moscow, instance of, city>). In (Raiman and Raiman, 2018) entity types are used for filtering of candidate entities. The document tokens are fed into BiLSTM to obtain mention embeddings, which are fed into dense layer for classification into classes corresponding to types.

In OpenTapioca system (Delpeuch, 2019) candidate entities are ranked by the popularity which is calculated by a log-linear combination of number of statements n_e of entity e , site links s_e and its PageRank $r(e)$. Global disambiguation is performed with similarity metrics $s(e, e')$ (the probability that two such one-step random walks starting from e and e' end up on the same item), which are combined using the Markov chain to obtain the score for each entity.

BLINK (Ledell Wu, 2020) retrieves candidate entities from Faiss index of description embeddings. Top N candidate entities descriptions are re-ranked with cross-encoder: the text with entity mention and description of every entity, separated with [SEP]-token, are fed into BERT and dense layer on top of [CLS] hidden state is used for classification into two classes: 1 - entity description corresponds to the mention, 0 - otherwise.

GENRE entity linking system (De Cao et al., 2020) is based on generative model (pretrained BART (Lewis et al., 2019)). GENRE can function in two modes: entity disambiguation, when the text is fed into the model and it generates the text annotated with Wikipedia page links in place of entity mentions, and entity linking, when the entity mention is marked with special token and the model generates the page title.

ExtEnD (Barba et al., 2022) system solves entity disambiguation task the same way as extractive question answering systems. ExtEnD is based on Longformer (Beltagy et al., 2020) which takes as input text with entity mention, marked with special tokens, and candidate Wikipedia page titles, separated with special tokens. The model is trained to find spans of the correct page title.

3 System for entity extraction and linking

The proposed entity linking system consists of the following components: identifying entity mentions in text, classification of entity mentions by types, retrieval of candidate entities from the database, disambiguation of candidate entities using Wikipedia hyperlinks graph.

3.1 Entity recognition

Entity recognition is implemented as classification of text tokens into three classes: "B-ENT" for beginning of the entity mention, "I-ENT" for inner part of the mention and "O" for other tokens. Text tokens are fed into pretrained Transformer (RoBERTa-tiny), Transformer hidden states are fed into dense layer for token classification.

We trained the model on the dataset of preprocessed Wikipedia pages. The process of page annotation includes the following steps:

1. we extracted all hyperlinks from the page with the corresponding mentions m_1^h, \dots, m_N^h ;
2. for the page and every hyperlink h_i on the page we extracted all Wikipedia surface forms $m_{i1}^s, \dots, m_{iK}^s$ using the anchor dictionary (the dictionary where a key is a page title and a value is the list of mentions of the page in Wikipedia);
3. we annotate the tokens of hyperlink mentions m_1^h, \dots, m_N^h with BIO-markup;
4. we find substrings which correspond to surface forms $m_{11}^s, \dots, m_{1K}^s, \dots, m_{N1}^s, \dots, m_{NK}^s$ and annotate with BIO-markup.

The dataset contains 130K samples in train set and 2K samples in valid set. RoBERTa-tiny, trained on the dataset, achieves F1=83.2 on valid set and F1=82.6 on test set.

Extraction of more or less entities from the text can be controlled with a threshold in token classification model (A.1).

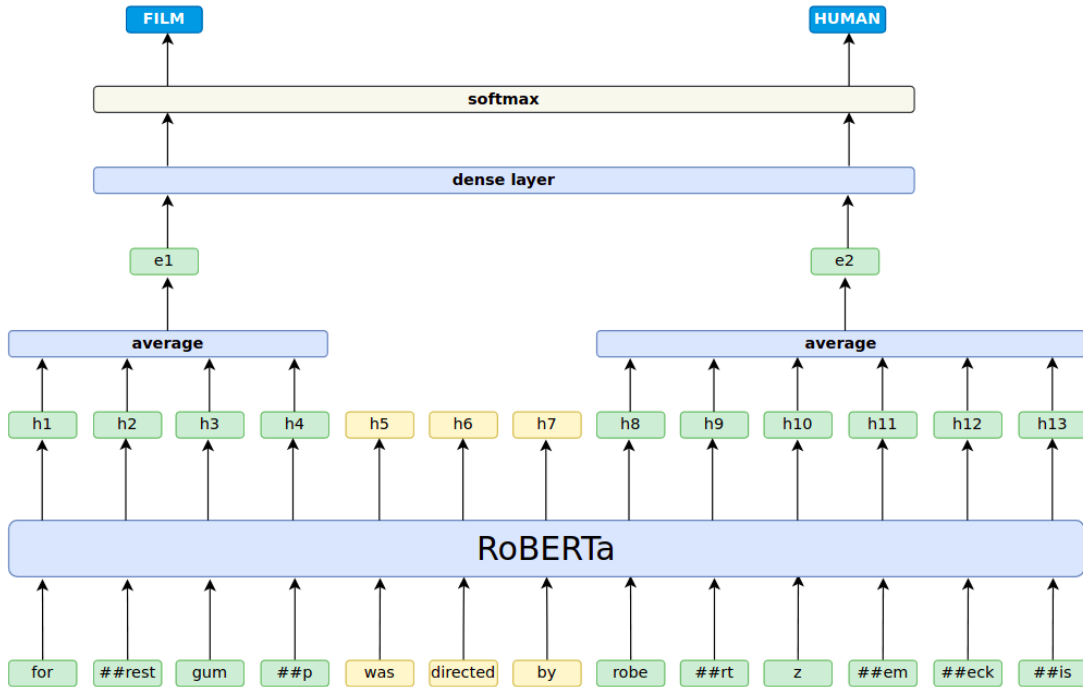


Figure 1: Type classification

3.2 Classification of entity mentions by types

Every entity in Wikidata has the relation P31 ("instance of") or P279 ("subclass of"), for example, <"Forrest Gump", "instance of", "film">. Entity types are useful for entity disambiguation. For example, in the sentence "Forrest Gump was directed by Robert Zemeckis." the type "film" of the mention "Forrest Gump" helps to choose the entity Q134773 ("Forrest Gump", film) instead of entities Q552213 ("Forrest Gump", novel) and Q3077690 ("Forrest Gump", fictional character).

Wikidata contains about 35K types (objects in triplets <entity, P31, type>). We united Wikidata types into 43 types (A.2), for example, Wikidata types "film", "television series", "animated feature film", "feature film", "animated film", "television program" we merged into the type "FILM". All Wikidata entities and corresponding Wikipedia page titles we annotated with these 43 tags.

For classification of entity mentions by types we feed text tokens into Transformer encoder (RoBERTa in our case). Mention embeddings are obtained by averaging of Transformer hidden states for mention tokens. Mention embeddings are fed into dense layer for classification into 42 classes corresponding to types (Figure 1).

For training of the model we processed paragraphs from Wikipedia pages with hyperlinks. For every hyperlink in the paragraph we found mention spans and the type for the hyperlink page title. We cut long paragraphs to the maximum length of 512 RoBERTa subtokens and left only paragraphs with at least two hyperlinks. The dataset contains 100K in train set and 2K in valid set. The trained model achieves F1=79.6 on WNED-WIKI dataset.

3.3 Entity disambiguation with Wikidata graph

In some cases correct entities for the mention are hard to disambiguate based on types. For example, in the sentence "Barcelona defeated Napoli with the score 4:2." the mention "Barcelona" corresponds to the entity Q7156 (FC Barcelona) and in the sentence "Barcelona defeated Valencia BC in the last match." "Barcelona" is Q54893 (FC Barcelona Basquet). We use connections between candidate entities for

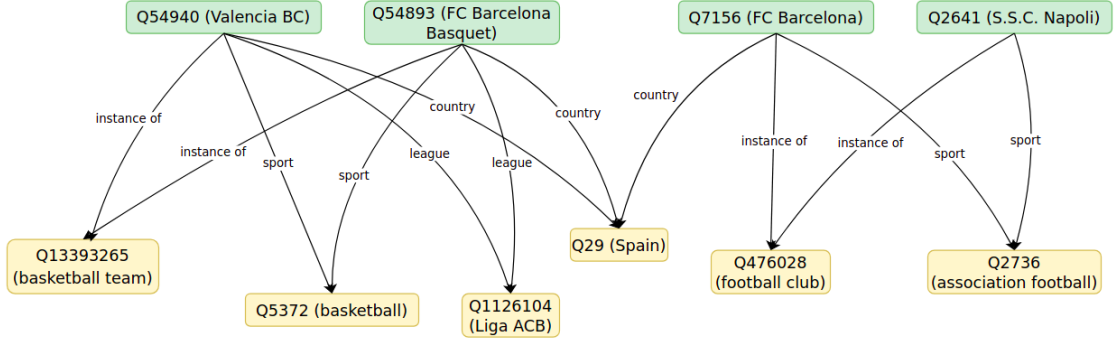


Figure 2: Global disambiguation

different mentions in the text in Wikidata knowledge graph and Wikipedia hyperlinks graph (Figure 2). Entities in Wikidata are mapped to corresponding Wikipedia pages, so we used both relations between entities in Wikidata and edges between pages in Wikipedia in hyperlinks graph. In the former sentence FC Barcelona and "Napoli" (Q2641 (S.S.C. Napoli) are connected with the edge "P31" (instance of) and the node Q476028 (association football club) and with the edge "P641" (sport) and the node Q2736 (association football).

Disambiguation of entities for mention m in text (for example, entities {"Q7156", "Q54893"} for mention "Barcelona" in the sentence "Barcelona defeated Napoli with the score 4:2.") was inspired by (Usbeck et al., 2014). For each entity $e_j^i \in C^i = \{e_1^i, \dots, e_{N_i}^i\}$ for mention m^i we find all entities in Wikidata connected with e_j^i with outgoing edges (the edges in directed graph that begins in e_j^i). In Figure 2 the edge, outgoing from the entity Q7156, connects Q7156 and Q29. We build a graph $G_k = (V_k, E_k)$, where $V_0 = \{C^1, \dots, C^N\}$ (candidate entities), $E_0 = \emptyset$, E_1 are edges outgoing from the nodes V_0 , V_1 are found as follows (2):

$$V_1 = V_0 \cup \{y : \exists x \in V_0 \wedge (x, y) \in E_1\} \quad (2)$$

All nodes $x, y \in V_1$ we initialize with authoritative values $x_a = \frac{1}{|V_1|}$ and hub values $x_h = \frac{1}{|V_1|}$ and iterate k times (3):

$$x_a \leftarrow \sum_{(y,x) \in E_1} y_h, y_h \leftarrow \sum_{(y,x) \in E_1} x_a \quad (3)$$

After k iterations all candidate entities $e_j^i \in C^i$ for mention m^i have corresponding values $x_{a_j}^i$, candidate entities are sorted by $x_{a_j}^i$.

4 Evaluation

The proposed entity extraction and linking system was tested on WNED-WIKI dataset. The system outputs three confidences: the Levenshtein distance between the mention (entity substring in text) and Wikidata entity title, the confidence of entity type classification model (Section 3.2) and the score of proximity with other mentions in Wikidata graph (Section 3.3). The final confidence was obtained as linear combination of these confidences and if the confidence is lower than the threshold, the entity mention was considered as not found in Wikidata.

WNED-WIKI dataset contains 6.8K samples with mentions from Wikipedia paragraphs and corresponding page titles. The proposed system outperforms REL (van Hulst et al., 2020) and OpenTapioca (Delpuch, 2019) on WNED-WIKI (Table 1). OpenTapioca disambiguates candidate entities by the number of connections between entities for different mentions in Wikidata graph. REL is based on ranking of candidate entities by dot products of entity and context embeddings. Global disambiguation in REL is performed by calculation of dot products of candidate entity embeddings for different mentions, but

the system does not use explicit information about connections between entities in Wikidata knowledge graph. Our system performs both local disambiguation (filtering of candidate entities by types obtained from type classification model) and global disambiguation by proximity of candidate entities in Wikidata.

GENRE, ExtEnD and BLINK systems achieve high F1 because they are based on powerful methods of page title generation (GENRE), extraction of page title span from the list of candidate titles (ExtEnD) and cross-attention between text and candidate entity description (BLINK) with large pretrained Transformers. GENRE is an encoder-decoder model with two modes:

- taking text with entity mention marked with special tokens as input and generating the page title;
- taking text as input and generating the same text where entity mentions are replaced with page titles.

Generation of page titles in autoregressive way, token-by-token, allows to learn relations between context and entity name.

The main component of ExtEnD system is a Longformer which receives the text where the entity mention is marked with special tokens, and the list of candidate pages titles. The model is trained to extract the span of correct page title the same way as extractive question answering models. Longformer hidden states are fed into two dense layers, the first defines the probability of the token to be the span start, the second - the span end. Cross-attention in Transformer architecture between page title, entity mention and text tokens leads to effective learning of relationship between page title and context.

BLINK system consists of two components: extraction of candidate entities from Faiss index and re-ranking of entities. At re-ranking step the text with entity mention replaced with special token and candidate entity description are fed into BERT and dense layer on top of CLS-token hidden state is used for classification of the description into two classes: 1 - if the description corresponds to the context, 0 - otherwise.

Large pretrained Transformers in GENRE and ExtEnD result in high quality, but using Longformer in ExtEnD leads to low inference speed. In GENRE prefix tree of 6M Wikipedia pages is loaded to RAM and requires 6.1 Gb. Also, GENRE and ExtEnD does not support zero-shot transfer to other knowledge bases. BLINK system is zero-shot: the entity is defined only by short text description, but the entities index (5.3 M) is loaded into RAM which requires 37.5 Gb. Cross-encoding of text and entity descriptions in BLINK is slower compared with other methods (Table 1) because the input text should be fed into BERT the number of times equal to the number of candidate entities. To obtain memory requirements of the models we launched each of the models on Nvidia DGX-1 server with Tesla P100 GPUs and inferred on WNED-WIKI dataset.

The proposed system shows lower F1 than GENRE, BLINK and ExtEnD on WNED-WIKI, but is fast and much more lightweight and can be used on an average laptop or desktop computer. Our system is based on RoBERTa-tiny for entity extraction and type classification and stores entity inverted index and Wikidata graph in SQLite database (2.5 Gb on disk, 42.9 M rows) which is not loaded into RAM (??). Moreover, our system does not need pretraining of entity embeddings and therefore supports easy adding of new Wikidata entity (with one insert query to SQLite database) and transfer to other knowledge bases, provided that the types of entities in the knowledge base were mapped to tags of entity type classification model.

Model	RAM, Gb	GPU, Gb	WNED, micro F1	Inference time, per 1 sample
Our system	1.9	1.4	68.2	0.15
GENRE	9.7	2.8	87.4	0.15
BLINK	37.5	1.1	75.5	0.61
ExtEnD	4.5	2.5	88.8	1.1
REL	2.0	0.95	41.4	0.17
OpenTapioca	4.4	0	26.8	0.21

Table 1: Comparison of the proposed entity linking system with other solutions

To define the contribution of entity linking system components into the metrics, we tested entity linking system on WNED-WIKI in two settings:

- using only entity type classification component for entity disambiguation;
- using both entity type classification and entity disambiguation with Wikidata graph.

In the former setting we achieved micro F1 of 49.8 on WNED-WIKI, in the latter setting - 68.2. The results indicate that connections in Wikidata and Wikipedia between entities in text for different mentions are significant for entity disambiguation and improve the metrics relative to using only entity type classification by about 18 points. For example, in the sample from WNED-WIKI "Towns within the division include Pipers River, Scottsdale, Evandale, Swansea, ..." for the mention "Swansea" the system in setting with using for disambiguation only entity types chooses the wrong entity Q23051 ("Swansea"). Wikidata graph helps to define to correct entity Q986654 ("Swansea, Tasmania"), because most of the locations in the sample text are connected with the entity Q34366 ("Tasmania").

5 Conclusion

In this work, we have described the system for entity extraction and linking. The system performs detection of entity mentions in the text, candidate entities retrieval, entity classification by types with RoBERTa-based model and entity disambiguation using Wikidata knowledge graph. The system is lightweight: entity extraction and type classification components are based on RoBERTa-tiny model, entities inverted index and Wikidata are stored in SQLite database, which is not loaded into RAM. Our system outperforms other lightweight solutions on WNED-WIKI dataset due to combination of local disambiguation based on filtering of candidate entities with type classification component and global disambiguation by proximity of candidate entities in Wikidata knowledge graph.

References

- Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2022. ExtEnD: Extractive entity disambiguation. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Online and Dublin, Ireland, May. Association for Computational Linguistics.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.
- Antonin Delpuech. 2019. Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131*.
- Paolo Ferragina and Ugo Scaiella. 2011. Fast and accurate annotation of short texts with wikipedia pages. *IEEE software*, 29(1):70–75.
- Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920*.
- Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*.
- Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. *arXiv preprint arXiv:1804.10637*.
- Phong Le and Ivan Titov. 2019. Boosting entity linking performance by leveraging unlabeled documents. *arXiv preprint arXiv:1906.01250*.
- Martin Josifoski Sebastian Riedel Luke Zettlemoyer Ledell Wu, Fabio Petroni. 2020. Zero-shot entity linking with dense entity retrieval. // *EMNLP*.
- 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*.
- 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*.

- Pedro Henrique Martins, Zita Marinho, and André FT Martins. 2019. Joint learning of named entity recognition and entity linking. *arXiv preprint arXiv:1907.08243*.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*.
- Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. // *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Ozge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2021. Neural entity linking: A survey of models based on deep learning.
- Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. 2014. Agdistis-graph-based disambiguation of named entities using linked data. // *International semantic web conference*, P 457–471. Springer.
- Johannes M van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P de Vries. 2020. Rel: An entity linker standing on the shoulders of giants. // *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, P 2197–2200.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

A Appendix

A.1 Confidence threshold for tags in token classification model

The softmax layer in token classification model outputs confidences of every class for the token (B-ENT, I-ENT, 0). We do not follow the strategy of choosing the label L_{ij} with maximal confidence p_{ij} for the token t_i . Instead, we set a threshold and choose the maximum of B-ENT and I-ENT confidences (if it is below the threshold) and O-tag otherwise (4):

$$L_i = \begin{cases} \text{B-ENT}, & p_{i,b-ent} > p_{i,i-ent} \ \&\& \ p_{i,b-ent} > thres \\ \text{I-ENT}, & p_{i,i-ent} > p_{i,b-ent} \ \&\& \ p_{i,i-ent} > thres \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The example of regulation of entity extraction with the threshold (the sample from WNED-WIKI dataset): if the threshold of B-ENT and I-ENT is 0.7, the model extracts from the sentence "Noel Mary Purcell was an Irish rugby union and water polo player." substrings "Noel Mary Purcell" and "Irish", if the threshold is 0.1, the substrings "rugby union" and "water polo" are also extracted as entities.

A.2 Tags for entity classification

Tags of entity type classification model were mapped with types of entities in Wikidata (types are defined with the relation P31 ("instance of"), for example, <"Forrest Gump", "instance of", "film">). The table 2 contains entity tags and corresponding entity types. For example, the tag "RIVER" is mapped to the type Q4022 ("river").

Using this mapping, we found tags for all Wikidata entities. The search of types was recursive (if the entity has the type which does not correspond to any tags, we found the types of the type, and so on till the one of the types matched any tag, the recursion depth was constrained to 10 steps). If no tag was found, the entity was assigned to "MISC" ("miscellaneous") tag.

Entity tag	Wikidata types	Entity tag	Wikidata types
film	Q11424, Q5398426, Q29168811 Q24869, Q202866, Q15416	work of art	Q838948, Q17537576
song	Q482994, Q55850593, Q7302866 Q105543609, Q134556	academic discipline	Q11862829
literary work	Q7725634	type of sport	Q31629
animal	Q729, Q7377, Q57814795, Q39201	music genre	Q188451
sport team	Q847017, Q12973014	sports season	Q27020041
food	Q2095, Q19861951	sports event	Q13406554, Q18608583
city	Q7930989	county	Q28575
country	Q7275, Q6256	politician	Q82955
fac	Q12280, Q811979, Q12819564 Q41176, Q1248784 Q34442, Q25631158	actor	Q33999
event	Q1656682, Q108586636, Q16510064	writer	Q36180, Q28389, Q49757
product	Q431289, Q167270, Q2424752	musician	Q639669, Q177220, Q36834 Q753110, Q488205
law	Q3150005, Q93288, Q1864008	athlete	Q2066131, Q18536342
language	Q20829075, Q20162172 Q34770, Q33742	national sports team	Q1194951
nation	Q6266, Q41710, Q81058955 Q33829, Q231002	river	Q4022
norp	Q4392985, Q9174, Q110401282 Q5390013, Q7257, Q49447, Q82821	road	Q34442
per	Q5	business	Q4830453, Q891723 Q6881511, Q783794
loc	Q1048835, Q15642541, Q486972 Q82794, Q618123	occupation	Q4164871, Q12737077, Q28640
org	Q43229	chemical element	Q11344, Q11173
		sports league	Q623109
		political party	Q7278
		us state	Q35657
		association football club	Q476028
		championship	Q1344963, Q500834, Q1079023
		sports venue	Q1076486

Table 2: Mapping of entity classification tags and Wikidata entity types

A.3 Candidate entities retrieval

Index of entities with corresponding Wikipedia page titles and Wikidata triplets is stored in SQLite database with FTS5 extension. The row in the table with entities contains entity title, entity id in Wikidata, Wikipedia page title, entity tag and string with Wikipedia triplets (in which the entity is the subject) and hyperlinks on corresponding Wikipedia page, separated with tabulation. The size of database is 2.5 Gb on disk, the database contains 42.9 M rows.

For retrieval of candidate entities we execute a query to the database which contains entity substring and top-3 tags, detected with entity type classification model. If the confidence of top-1 tag is lower than the threshold ($thres = 0.4$), "MISC" tag ("miscellaneous") is added to the set of tags in the query.

Controllable Multi-attribute Dialog Generation with PALs and Grounding Knowledge

Evseev D. A.

Moscow Institute of Physics
and Technology
Dolgoprudny, Russia
dmitrij.euseew@yandex.ru

Nagovitsin M. S.

Moscow Institute of Physics
and Technology
Dolgoprudny, Russia
nagovitsin.ms@phystech.edu

Kuznetsov D. P.

Moscow Institute of Physics and Technology
Dolgoprudny, Russia
kuznetsov.den.p@gmail.com

Abstract

Today, neural language models are commonly employed for generation of natural like responses in dialog systems. The main issue that limits wide adoption of neural generation is related to poor predictability of responses in terms of content, as well as dialog attributes such as dialog acts and sentiment. In this paper we propose a method based on projected attention layers (PALs) for controllable multi-attribute knowledge grounded dialog generation. We compared a number of methods for training and blending representations produced by PALs combined with DialoGPT base model. Results of our experiments demonstrate that separate pre-training of PAL branches for different attributes followed by transfer and fine-tuning of dense blending layer gives the highest accuracy of control of a generated response for fewer trainable parameters per an attribute. Furthermore, we applied our approach for controllable multi-attribute generation with grounding knowledge to Blenderbot model. Our solution outperforms the baseline Blenderbot and CRAYON model in control accuracy of dialog acts and sentiment on Daily Dialog as well demonstrates a comparable overall quality of dialog generation given grounding knowledge on Wizard of Wikipedia.

Keywords: controllable dialog generation, knowledge grounded dialog generation, projected attention layers.

DOI: 10.28995/2075-7182-2022-21-185-202

Управляемые диалоговые генеративные модели с параметризацией атрибутов и дополнительным контекстом генерируемых реплик на основе PALs

Евсеев Д. А.

Московский физико-технический
институт
Долгопрудный, Россия
dmitrij.euseew@yandex.ru

Наговицин М. С.

Московский физико-технический
институт
Долгопрудный, Россия
nagovitsin.ms@phystech.edu

Кузнецов Д. П.

Московский физико-технический институт
Долгопрудный, Россия
kuznetsov.den.p@gmail.com

Аннотация

В настоящее время нейронные языковые модели широко используются для генерации реплик в диалоговых системах. Основным недостатком нейронной генерации связан с непредсказуемостью содержания реплик. В данной статье мы предлагаем метод, основанный на проекционных слоях внимания (PAL) для управления несколькими параметрами генерируемой реплики. Мы сравниваем несколько методов обучения и смешивания представлений, полученных с помощью PALs, добавленных к DialoGPT. В ходе экспериментов было установлено, что обучение PALs отдельно

для каждой из веток управляемых параметров и последующее дообучение полносвязного смешивающего слоя приводит к наибольшей точности управляемой генерации, при этом используя меньшее число обучаемых параметров. Также мы применили наш подход для генерации с контролем атрибутов опираясь на внешние знания, с помощью модели Blenderbot. Наше решение превосходит базовую версию Blenderbot и модель CRAYON по точности предсказания диалогового акта и сентимента генерируемых реплик на датасете Daily Dialog, и при этом показывает сравнимое качество генерации реплик с использованием контекста на датасете Wizard of Wikipedia.

Ключевые слова: управляемая генерация реплик в диалоге, генерация реплик по контексту, проекционные слои внимания.

1 Introduction

Majority of open-domain dialog systems use hand-crafted finite state machines for response generation (Larsson and Traum, 2000), (Bocklisch et al., 2017), (Finch and Choi, 2020). For every expected user utterance these systems define a state with pre-defined output response and transition to the next state of the dialog, but user input can mismatch a condition for transition in the current state. As well, the user input can mismatch all possible states defined by the finite state machine. Here, neural generative models are able to help with producing natural like responses. Unfortunately, generative models demonstrate very unreliable coherence with existing dialog context (Abhishek et al., 2021). One of the possible solution is to use controllable attributes such as dialog act or sentiment to guide generation of responses and return the dialog flow back to the domain of pre-defined script. If the script is defined as pairs of adjacent dialog acts, or in goal-oriented dialog, the attributes of the system response (intent, slot values) are extracted (Wu et al., 2019), for known attributes of the response a generative model conditioned on grounding knowledge about entities found in the dialog context or slot values, is able to generate all the bot utterances in the script without retrieval of hand-written responses. Also the right level of control can improve dialog quality in different aspects (See et al., 2019). Social bots like Gunrock (Yu et al., 2019) and XiaoIce (Zhou et al., 2019) are actively using dialog acts, sentiment and other attributes for dialog management.

Controllable generative models have been an active area of research over last years. Models (Zhao et al., 2017), (Zhang et al., 2018) control one attribute of the generated response (dialog act, response relatedness or specificity). The need to control different attributes simultaneously is present in (See et al., 2019). General approaches to control multiple attributes simultaneously were presented for different model architectures in (Hu et al., 2021b), (Xu et al., 2019), (Yu et al., 2021), (Du and Ji, 2021), (Yang et al., 2021). In this paper we propose and study a technique for multi-attribute generation control which is suitable for the both pre-training as well as fine-tuning. We use PALs (Stickland and Murray, 2019) with transformer architectures, consequently parameters of the main pre-trained model provide constant background knowledge and PAL layers are trained to control generation in respect with specific attribute.

Informativeness and meaningfulness is another important aspect of generated responses. Blenderbot (Roller et al., 2020), CoLV (Zhan et al., 2021) and CGRG (Wu et al., 2021) use grounding knowledge (retrieved paragraphs) to control the content of output utterances. But these models are not able to be controlled to produce the response with required attributes, such as dialog act or sentiment.

Trained models for English language, training and inference code and data to test the quality of models are published in Open Source under the Apache 2.0 license¹. The main contributions of this work are the following:

- we develop the method of controllable generation for several simultaneous attributes such as dialog acts and sentiment with no changes in weights of the original model;
- we study simultaneous control of knowledge grounding as well as dialog act and sentiment of a response, and find that our model outperforms existing approaches in terms of dialog act and sentiment control accuracy and is competitive in terms of perplexity of knowledge grounded generation.

2 Related Work

There are many different approaches to control generation process with or without changing the architecture or retraining the initial language model. PPLM (Dathathri et al., 2019) updates latent represent-

¹<https://github.com/deepmipt/controllable-generation>

ations of model during decoding with help of pre-trained discriminator, GeDI (Krause et al., 2020) uses class-conditional distributions to achieve control for both desired and undesired attributes (for example, generate less toxic answers). Both PPLM and GeDI work without changing the initial model.

Another way of control is to add special control tokens as an input for model. It can be done by fine-tuning (or full training) of model, as proposed in CTRL (Keskar et al., 2019), or by keeping the initial model as is and training a special alignment function which will generate proper key and value representations for control tokens on each level of pre-trained Transformer model, as proposed in Attribute Alignment (Yu et al., 2021). Those methods are capable of controlling multiple attributes by adding tokens for each one.

For larger models prefix tuning (Li and Liang, 2021) can be applied. Prompts also can be used for a few shot learning (Zheng and Huang, 2021).

Some methods work with a latent space, for example CRAYON (Hu et al., 2021b), which is based on LSTM, or PHED (Yang et al., 2021), which uses Transformer combined with CVAE to add a latent space to the Transformer architecture. Model GTMES2S, proposed in (Xu et al., 2019) uses additional modules to control current level of all attributes and guide further generation to reach desired values. Each of CRAYON, PHED and GTMES2S are working with multiple attributes. In XiaoIce dialog assistant (Zhou et al., 2019) GRU is used to generate responses, conditioned on query and response empathy vectors.

The other method is to add special modules to shift embeddings (and maybe hidden states) from initial model, same as with PPLM (Dathathri et al., 2019), but to learn a proper transformations in advance and not to tune them on inference. SideControl (Du and Ji, 2021) propose a way to perturb embeddings of any language model by training additional module, which will take class embedding or grounding knowledge into account. Another way is to treat control as a special task and add adapters (Houlsby et al., 2019) for different attributes we want to control. One of such models is Adapter-Bot (Madotto et al., 2020) which has an option of switch between different attributes without changes in initial model. Hyperformer (Mahabadi et al., 2021) utilizes a shared PAL parameters for all tasks and Transformer layers, these parameters are generated by a hypernetwork. The model (Xie and Pu, 2021) is an encoder-decoder Transformer, where emotions in response are controlled with emotion embeddings, fed into the model. But models with adapters are not able to control multiple attributes simultaneously without additional work. Task-specific parts to the model can also be efficiently added with help of low-rank decomposition, as proposed in LoRA (Hu et al., 2021a).

The key points in which our model differs from the others are following:

- Ability to control of multiple attributes simultaneously (dialog act and sentiment in our study);
- Preserving the weights of initial model in original state;
- Ability to pre-train all adapters independently on different datasets;
- Scalability in terms of number of controllable attributes and number of their values;

Most of generative models, which do not use external knowledge, are capable of producing grammatically correct and natural responses given the dialog history, but have a limited ability to generate interesting responses based on facts. On the other hand, knowledge-grounded generative models have an option of controlling content of generated responses with sentences with facts or keywords. CGRG (Wu et al., 2021) model uses lexical control phrases to control the generated response. The approach of (Xu et al., 2021b) is based on PALs for different topics which are used for retrieval-free knowledge grounded generation. The model (Zhan et al., 2021) uses latent variables for relevant knowledge selection and response generation. The models (Xu et al., 2021a), (Kumar et al., 2021) and (Gupta et al., 2020) controls the generated response by adding as input of the transformer the sequence of keywords before the dialog history. Our approach is inspired with Blenderbot (Roller et al., 2020) which is an encoder-decoder transformer pretrained on Reddit and finetuned on Wizard of Wikipedia (Dinan et al., 2018), but our model controls not only the content of the response and moreover dialog act and sentiment.

3 Methods

In this paper, our goal is to find a method to control different response attributes without losing much token prediction quality (perplexity) and other abilities of the base pre-trained model (e.g., using ground-

ing knowledge). We did most of our experiments with DialoGPT-small architecture (Zhang et al., 2020b), because of the affordable time to fine tune and the good quality of the pre-trained model. Additional experiments with simultaneous control of content, dialog acts and sentiment we performed with Blenderbot architecture (Roller et al., 2020). Furthermore, we chose dialog acts (inform, question, directive, commissive) and sentiment (negative, neutral, positive) as controlled attributes. For evaluation of control accuracy we used DailyDialogs (Li et al., 2017), sentiment labelling was made separately by classifier. For evaluation of knowledge-grounded dialog generation quality (perplexity) we used Wizard of Wikipedia dataset (Dinan et al., 2018).

Control	Blend	Train dataset	Dialog act acc.	Sentiment acc.	Perplexity	Opt. steps	Trainable par.
No control	-	DailyDialogs	25.20 \pm 0.21	33.41 \pm 0.15	15.19 \pm 1.58	2000	117M
Dialog acts	average	DailyDialogs	63.74 \pm 0.32	42.83 \pm 0.27	15.93 \pm 0.12	10000	36M
Dialog acts	dense	DailyDialogs	45.27 \pm 5.26	40.15 \pm 1.22	22.36 \pm 0.85	5000	49M
Sentiment	average	ScenarioSA	33.40 \pm 0.16	72.09 \pm 4.06	92.98 \pm 14.74	5000	28M

Table 1: Models with control of one attribute. The model with no control is a finetuned DialoGPT-small, models with control are DialoGPT-small with PALs. Metrics were calculated on valid set of Daily Dialog.

Blend	Transfer	Dialog act acc.	Sentiment acc.	Perplexity	Opt. steps	Trainable par.
average	no	63.09 \pm 2.22	69.19 \pm 1.10	17.12 \pm 0.40	5000	63M
dense	no	61.65 \pm 1.02	67.10 \pm 1.38	22.07 \pm 0.39	5000	84M
dense & average	no	61.36 \pm 1.40	68.12 \pm 0.69	15.51 \pm 0.13	5000	77M
average	yes	65.62 \pm 2.04	66.04 \pm 0.25	17.74 \pm 0.75	5000	63M
weighted average	yes	63.20 \pm 1.09	69.05 \pm 0.42	15.65 \pm 0.09	5000	63M
dense	yes	60.83 \pm 1.35	67.80 \pm 3.37	21.34 \pm 0.40	5000	84M
dense & average	yes	62.76 \pm 0.70	70.03 \pm 2.04	15.69 \pm 0.19	5000	77M
dense & average, only blend	yes	60.19 \pm 0.76	67.47 \pm 0.90	15.30 \pm 0.05	10000	14M

Table 2: Models with simultaneous dialog act and sentiment control. Transfer means that PALs were initialized with weights from model for single attribute control.

3.1 Projected Attention Layers as Multitask Adapters

One of the approaches to control object attributes is to learn proper shifts in latent space (Hu et al., 2021b). One way to modify latent representations for every token is to use *Projected Attention Layers (PALs)* (Stickland and Murray, 2019) as adapters for every controllable attribute. In our case, each PAL will learn to correct hidden states of the main model to generate a response with the desired attribute (Figure 1).

3.2 Blend layer

To control several attributes simultaneously, we decided to add a PAL for each attribute and run them in parallel (Figure 1). We chose *average blending* as our baseline for blending of hidden representations. It allows us to control easily the contribution of each PAL to the resulting hidden states by weighting them. Then we try a trainable way of blending outputs of PAL branches: *dense blending* — concatenation of PALs outputs and the main branch and feeding into the dense layer; *combination of dense and average blending* — concatenation of PALs outputs, feeding into the dense layer and averaging the output with the base model. The loss function stays unchanged from the task of the next token prediction. For every labeled sample from training data we chose only corresponding PALs and train them, the base model is frozen.

3.3 Default branch

We added the *"default" branch* for each attribute for default selection values for attributes. Default branch is turned on for training on every sample instead of specialized PAL with probability $p = 0.2$. Thus default branch will be trained on all dataset and will not be bound to one attribute value.

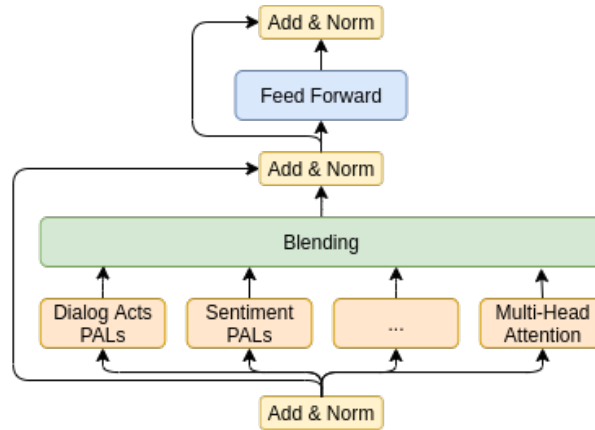


Figure 1: Blending of PALs and multi-head attention of Transformer hidden representations for every token.

3.4 Independent PALs pre-training

We independently trained models for dialog act and sentiment control and transferred these pre-trained branches into one model. Even without any further training resulting model demonstrated a noticeably good attribute control without huge degradation of perplexity, even though PALs for the sentiment were trained on a different dataset (more details in Appendix A.4). After transfer, the model with the blending layer can be finetuned on the target dataset.

3.5 Control of attributes and context of responses

One of our goals is to develop a model which could generate responses for a given grounding knowledge and global attributes, such as dialog act and sentiment. We modified Blenderbot Transformer architecture for control of global attributes of the response by adding PALs in parallel with the self-attention layer of the decoder layers. The decoder layer in our modification has 5 branches for dialog acts and 4 for sentiment. The attribute branches were blended with the dense layer and then added to the main branch of the base model.

4 Experiments and results

4.1 Evaluation

We used two metrics to estimate the quality of our models: perplexity to test that model is able to produce relevant and natural like responses and ability to control attributes. We generate responses for every turn on a validation part of DailyDialog and use attribute classifiers (see Appendix A.2) to check if the response of the model is correct and calculate balanced accuracy for each attribute. For example, for the dialog act attribute, we estimate the dialog act of each generated response and compare it with the gold label. Every model was trained for the same amount of steps, and then the best by perplexity checkpoint was scored. Blending experiments were performed with DialoGPT-small (117M) as a pre-trained base model. All parameters of PALs were taken from the original paper (Stickland and Murray, 2019), thus the PAL embedding dimension was 204. Training setup is the same as reported for original DialoGPT (Zhang et al., 2020b).

4.1.1 One attribute

When only one attribute is controlled there are no conflicts between PALs, because only one attribute shift is learned. We tried averaging and dense layer to blend the output of PAL and the layer of the main model (Table 1). The averaging is better in both perplexity and accuracy and is much easier for

further transfer because there is no need to add the blending layer to the target base model. Resources consumption is shown in Appendix A.1.

4.1.2 Two attributes

In the case of controlling multiple attributes simultaneously every PAL should adapt to its neighbors and learn to change only the corresponding attribute. Experiments (Table 2) have shown that the control abilities or perplexity are slightly better in the case of PALs pre-training and transfer compared to training added multi-attribute PALs from scratch. Average blending gives the best control for the similar perplexity. Dense layer blending results in perplexity drop. The model with a combination of dense and average blending shows the best perplexity and great control abilities. For other blending option perplexity is also on the same level, and control is better for one attribute and worse for another. Since each PAL was pre-trained with average blending, a more natural way to blend them is weighted average (see Appendix A.4), this gives better perplexity. With weighted average as a blending layer, it is possible to control the contribution of each PAL to every attribute. If the weights are transferred, another alternative to finetune the model is to train only blending layer. We choose combination of dense and average blending to finetune, and it results in the best perplexity and good control abilities (last row in the Table 2). Resources consumption is shown in Appendix A.1. Examples of dialogs can be found in Appendix A.9. More discussion on perplexity-accuracy tradeoff can be found in Appendix A.7.

Model	D.A. acc.	Sent. acc.	PPL
Bl. bot, cont., 199M	77.01	84.90	28.42
Bl. bot 400M	38.10	28.43	18.24
Bl. bot 90M	38.18	27.96	76.10

Table 3: Comparison of controllable Blenderbot (dense and average blending) with Blenderbot from Huggingface (balanced accuracy and perplexity) with grounding knowledge.

Model	Q/noQ acc.	Sent. acc.
Bl. bot, cont., d&avg	99.45	85.87
CRAYON	98.17	82.17

Table 4: Comparison of controllable Blenderbot (dense and average blending) with CRAYON model in question asking and sentiment control accuracy.

4.2 Blenderbot results

The next series of experiments was performed with Blenderbot for dialog acts and sentiment control (4 layers in encoder, 8 layers in decoder, embedding dimension of 576, 119M parameters). We pretrain Blenderbot on Reddit and finetuned on Daily Dialog, ConvAI2 (Dinan et al., 2020), Emphatetic Dialogue and Wizard of Wikipedia.

We compared Blenderbot with PALs and baseline Blenderbot on Daily Dialog dataset (Table 3). It was found that extended Blenderbot outperforms Blenderbot 400M and Blenderbot 90M from Huggingface library in dialog acts and sentiment control accuracy and is comparable with the baseline in perplexity of dialog generation given grounding knowledge (GK) on Wizard of Wikipedia dataset.

We compared controllable Blenderbot with CRAYON (Hu et al., 2021b) in question asking and sentiment control accuracy on Daily Dialog dataset. Our model controls 4 types of dialog acts, therefore we used PAL for "question" dialog act to generate a question and PAL for "inform" otherwise. Blenderbot with PALs outperforms CRAYON in question asking and sentiment control accuracy (Table 4).

5 Conclusion

In this paper with presented the study of techniques for multi-attribute control of neural response generation in the dialog with and without grounding knowledge. Our methodology employs extension of

pre-trained generative base model with attribute specific projected attention layers (PALs). Results of our experiments allow to draw the following conclusions.

If the base model is already trained and the quality of the responses is a first priority, then the best way is to pre-train PALs for each attribute separately (maybe on different datasets) with the average blending. Then transfer pre-trained PALs to the base model and finetune with weighted average or combination of average and dense blending. If a degradation of perplexity is not noticeably harmful then average blending without transfer is also an option due to ability to control the contribution of each attribute.

Our results demonstrate that proposed approach can be successfully applied to controllable generation of responses in the dialog conditioned on multiple attributes for less numbers of trainable parameters per attribute. The method can be also combined with grounding knowledge. Compared to the baseline our solution shows better accuracy of dialog acts and sentiment control with similar perplexity.

References

- Tushar Abhishek, Daksh Rawat, Manish Gupta, and Vasudeva Varma. 2021. Transformer models for text coherence assessment. *arXiv preprint arXiv:2109.02176*.
- Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). // *The NeurIPS'18 Competition*, P 187–208. Springer.
- Wanyu Du and Yangfeng Ji. 2021. Sidecontrol: Controlled open-domain dialogue generation via additive side networks.
- James D Finch and Jinho D Choi. 2020. Emora stdm: A versatile framework for innovative dialogue system development. *arXiv preprint arXiv:2006.06143*.
- Prakhar Gupta, Jeffrey P Bigham, Yulia Tsvetkov, and Amy Pavel. 2020. Controlling dialogue generation with semantic exemplars. *arXiv preprint arXiv:2008.09075*.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models.
- Zhe Hu, Zhiwei Cao, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Jinsong Su, and Hua Wu. 2021b. Controllable dialogue generation with disentangled multi-grained style specification and attribute consistency reward.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.
- 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.
- Shachi H Kumar, Hsuan Su, Ramesh Manuvinakurike, Saurav Sahay, and Lama Nachman. 2021. Controllable response generation for assistive use-cases. *arXiv preprint arXiv:2112.02246*.
- Staffan Larsson and David R Traum. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural language engineering*, 6(3 & 4):323–340.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

- Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2020. The adapter-bot: All-in-one controllable conversational model.
- Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.
- Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning.
- Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 808–819, Florence, Italy, July. Association for Computational Linguistics.
- Zequ Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021. A controllable model of grounded response generation.
- Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. // *Proceedings of the 25th Conference on Computational Natural Language Learning*, P 133–147.
- Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words.
- Heng-Da Xu, Xian-Ling Mao, Zewen Chi, Fanshu Sun, Jingjing Zhu, and Heyan Huang. 2021a. Generating informative dialogue responses with keywords-guided networks. // *CCF International Conference on Natural Language Processing and Chinese Computing*, P 179–192. Springer.
- Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2021b. Retrieval-free knowledge-grounded dialogue response generation with adapters. *arXiv preprint arXiv:2105.06232*.
- Haiqin Yang, Xiaoyuan Yao, Yiqun Duan, Jianping Shen, Jie Zhong, and Kun Zhang. 2021. Progressive open-domain response generation with multiple controllable attributes.
- Dian Yu, Michelle Cohn, Yi Mang Yang, Chun-Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Girithija Sreenivasulu, Sam Davidson, Ashwin Bhandare, and Zhou Yu. 2019. Gunrock: A social bot for complex and engaging long conversations.
- Dian Yu, Zhou Yu, and Kenji Sagae. 2021. Attribute alignment: Controlling text generation from pre-trained language models.
- Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021. Colv: A collaborative latent variable model for knowledge-grounded dialogue generation. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 2250–2261.
- Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018. Learning to control the specificity in neural response generation. // *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1108–1117.
- Yazhou Zhang, Zhipeng Zhao, Panpan Wang, Xiang Li, Lu Rong, and Dawei Song. 2020a. Scenariosa: A dyadic conversational database for interactive sentiment analysis. *IEEE Access*, 8:90652–90664.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.
- Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation.
- Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2019. The design and implementation of xiaoice, an empathetic social chatbot.

A Appendix

A.1 Resources

For all experiments, we used NVIDIA GeForce GTX 1080 Ti GPUs. Training DialoGPT-small with one attribute control for 10000 steps took about 8 hours using two GPUs. Training model with two attribute control and (weighted) average blending for 5000 steps took about 6 hours, with dense blending - about 8 hours, and with a combination of average and dense blending - 7 hours on two GPUs. Train only blend layer for a combination of dense and average took about 11 hours on the same devices. The batch size was set to 256 divided into 8 steps of gradient accumulation. Extended Blenderbot was trained with batch size of 1000 on 10 NVIDIA GeForce GTX 1080 Ti GPUs. Pretraining on part of Reddit dataset (dump from 2014 and 2015 years) took 48 hours.

A.2 Evaluation and Classifiers

We used the validation part of the DailyDialog (Li et al., 2017) dataset to evaluate our models. Daily-Dialog is labeled with dialog acts, moreover we needed labels for the sentiment. Number of utterance for each attribute is shown on Figure 4. Since classes are not balanced, we used balanced accuracy (from package scikit-learn 0.21.2, `sklearn.metrics.balanced_accuracy`). To evaluate the model we generated responses on the test set with the right PALs (according to the gold labels) and check if the response was generated with desired attributes. Consequently we needed to classify dialog acts and sentiment to (1) evaluate our model and (2) label datasets automatically.

For dialog acts and sentiment classification we used the BERT-based model. One (current) or two utterances (current and previous), separated with SEP-token, were fed into BERT. The hidden state of the BERT CLS-token was fed into the dense layer, followed by softmax classification. Dialog acts classifier was trained on Daily Dialog (Li et al., 2017), sentiment classifier - on Scenario SA (Zhang et al., 2020a). Balanced accuracy of dialog act classifier is 72.90%, the confusion matrix is in Figure 2. The balanced accuracy of the sentiment classifier is 76.24%.

A.3 Default branch

We added "*default*" branch for each attribute for the cases when we don't want or don't need to control it. The default branch is the same PAL as the other, except during training it turns on every time instead of any other PAL for this attribute with the probability p , we chose $p = 0.2$. To check that the default branch is working as expected, we evaluated the model (DialoGPT-small with control of dialog acts and sentiment and combination of dense and average as a blend layer) in four setups:

- Usual inference (default branch is off)
- Default branch is always set for dialog act attribute
- Default branch is always set for sentiment attribute
- Default branch is always set for both dialog act and sentiment attributes

The results are in the Table 5. With default control of each attribute is back on the level of base models (without attribute control). With default branches, perplexity grows, but not too much. That means that those branches are trained pretty well and that our model is better at control (than base DialoGPT) not just because of the larger number of parameters, but because PALs are learning their domains. Otherwise, default branches would show great control abilities too.

Default attributes	Dialog act acc.	Sentiment acc.	Perplexity
No default	62.76 \pm 0.70	70.03 \pm 2.04	15.69 \pm 0.19
Dialog act	31.97 \pm 0.42	70.19 \pm 2.27	16.34 \pm 0.28
Sentiment	62.85 \pm 0.38	42.47 \pm 0.75	16.03 \pm 0.17
Dialog act and sentiment	30.06 \pm 0.79	39.67 \pm 0.31	16.83 \pm 0.57

Table 5: Work of default branches for each attribute. Evaluated with the model for dialog act and sentiment control with a combination of dense and average blending.

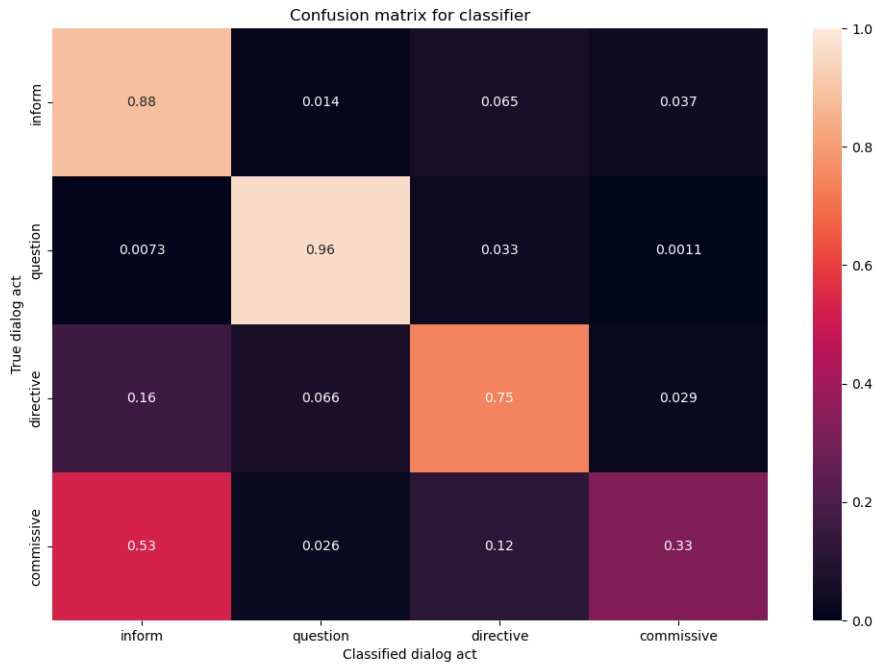


Figure 2: Confusion matrix for dialog acts classifier.

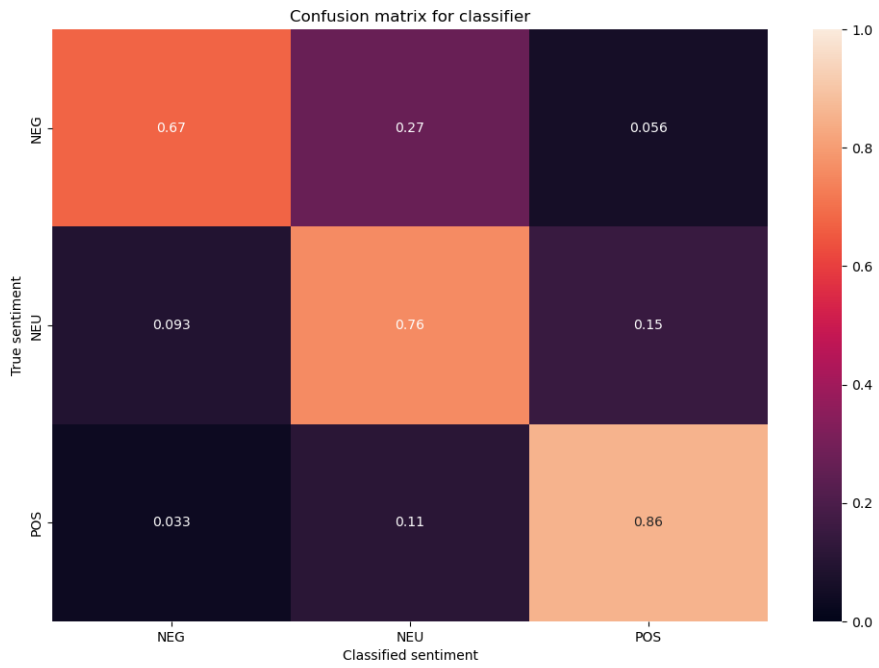


Figure 3: Confusion matrix for sentiment classifier.

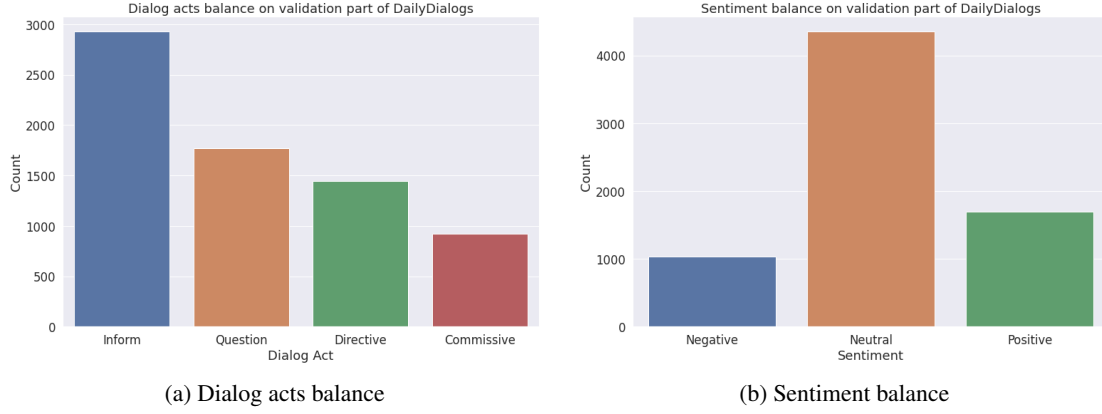


Figure 4: Attributes balance on validation set of DailyDialog

A.4 Average and weighted average blending

Originally (Stickland and Murray, 2019) the output of PALs is added to the output of the corresponding layer in the base model. But we run several PALs simultaneously. We can still just add all PAL’s outputs to hidden states of the main model, but since we add an arbitrary number of PALs in parallel, the summation scales poorly. This is due to the inconsistency of absolute values of hidden states and their dependency on the number of attributes to control. For this reason, we choose average as a blending layer. Since there are no trainable parameters on the blending stage, each PAL output is an embedding, shifted in a proper direction in the latent space. Furthermore, we can easily transfer the weights of PALs from a model for one-attribute control to a model with the control of several attributes. But average blending with one attribute has the following formula:

$$Emb = \frac{Main + PAL}{2} \quad (1)$$

Average blending for several attributes has the following formula:

$$Emb = \frac{Main + PAL_1 + \dots + PAL_N}{N + 1} \quad (2)$$

If we transfer weights with an average blending layer then each PAL would influence more than it was in a model with a single attribute control. For example with two attributes:

$$Emb = \frac{Main + PAL_1 + PAL_2}{3} = \frac{1}{2} \left(\frac{Main + 2 \cdot PAL_1}{3} + \frac{Main + 2 \cdot PAL_2}{3} \right) \quad (3)$$

For this reason, control abilities may be better, but perplexity will probably drop. To solve this problem we tried weighted average:

$$Emb = \frac{N \cdot Main + PAL_1 + \dots + PAL_N}{2N} \quad (4)$$

For two attributes is:

$$Emb = \frac{2 \cdot Main + PAL_1 + PAL_2}{4} \quad (5)$$

In our experiments weighted averaging significantly improved perplexity and dropped accuracy a little (Table 2).

In the same way, we can directly control, how much each attribute influences the resulting embedding by tuning the weights for each attribute branch. For example, we can add more weight to dialog act PAL and get better accuracy for this attribute, but for other attributes, control ability will probably drop. We experimented with three models (each one controls dialog act and sentiment):

1. PALs weights transferred from models with control of only one attribute without further training (Table 6)
2. PALs weights transferred and model was trained (with weighted average blend) (Table 7)
3. Model was trained (with average blend) without transfer (Table 8)

Visual results can be found in Figure 5. Results show that with and without weights transfer branches are learning desired attributes as expected, and it is possible to control the impact of each attribute if needed.

Branch weights			Dialog act acc.	Sentiment acc.	Perplexity
Dialog act	Sentiment	Main			
0.33	0.33	0.33	57.88%	62.15%	44.53
0.25	0.25	0.50	55.06%	59.86%	24.91
0.20	0.20	0.60	49.81%	55.66%	21.89
0.33	0.17	0.50	58.37%	53.39%	19.42
0.38	0.12	0.50	60.90%	50.32%	17.97
0.40	0.20	0.40	60.60%	54.70%	23.48
0.17	0.33	0.50	49.34%	64.40%	36.20
0.12	0.38	0.50	46.95%	68.44%	45.67
0.20	0.40	0.40	52.27%	67.94%	55.51

Table 6: Reweighting the impact of just transferred PALs to improve control for selected attributes. Perplexity is high when the weight of sentiment PALs is high because the model for sentiment control was trained on a different dataset.

Branch weights			Dialog act acc.	Sentiment acc.	Perplexity
Dialog act	Sentiment	Main			
0.33	0.33	0.33	62.21%	65.40%	25.04
0.25	0.25	0.50	64.42%	68.53%	15.55
0.20	0.20	0.60	58.00%	65.60%	15.48
0.33	0.17	0.50	67.02%	59.02%	17.63
0.38	0.12	0.50	67.13%	53.61%	20.18
0.40	0.20	0.40	61.52%	56.18%	25.97
0.17	0.33	0.50	53.23%	73.54%	16.32
0.12	0.38	0.50	46.97%	74.17%	17.90
0.20	0.40	0.40	54.99%	75.96%	18.71

Table 7: Reweighting the impact of transferred and finetuned PALs to improve control for selected attributes.

A.5 Comparison of pretraining and fine-tuning

We trained different architectures and methods of pretraining on OpenSubtitles dataset and then evaluated on test set of Daily Dialog. Samples from OpenSubtitles were preprocessed with classifiers for dialog acts and sentiment. We left only samples with confidence of dialog act classification upper 0.5 and sentiment upper 0.8, in total the dataset contains 8.9M samples.

To run the experiments faster, we used very small version of DialoGPT with 6 layers and embedding dimension 256. The Table 9 shows a comparison for small models. We compared the following cases:

1. PALs added at every layer of DialoGPT in place of the main branch, the PALs are pretrained at the same time as the model;
2. PALs added in parallel with the main branch, the model is first pretrained without PALs and then frozen with only PALs training;
3. PALs in place of the main branch and at training the batch contains samples for different dialog acts and sentiment.

Pretraining of PALs results in higher accuracy of attribute generation than fine-tuning.

Branch weights		Main	Dialog act acc.	Sentiment acc.	Perplexity
Dialog act	Sentiment				
0.33	0.33	0.33	65.87%	69.44%	16.63
0.25	0.25	0.50	54.08%	63.84%	17.37
0.20	0.20	0.60	49.25%	55.87%	19.63
0.33	0.17	0.50	61.26%	54.81%	18.26
0.38	0.12	0.50	62.34%	52.30%	19.80
0.40	0.20	0.40	68.27%	57.56%	18.24
0.50	0.25	0.25	72.69%	59.07%	25.33
0.17	0.33	0.50	47.52%	69.07%	17.83
0.12	0.38	0.50	43.23%	71.39%	18.58
0.20	0.40	0.40	54.30%	75.19%	18.24
0.25	0.50	0.25	49.67%	76.52%	33.46

Table 8: Reweighting the impact of trained together from scratch PALs to improve control for selected attributes.

Training setting	Dialog acts accuracy	Sentiment accuracy	Perplexity
PALs, pretraining with the main model	78.73 \pm 0.86	71.20 \pm 1.91	315.06 \pm 3.11
PALs, freezed main model	70.50 \pm 2.62	62.07 \pm 3.27	368.54 \pm 8.97
PALs, different attributes in batch	80.32 \pm 2.79	74.13 \pm 3.43	365.60 \pm 11.50

Table 9: Comparison of PALs training methods on small DialogPT

A.6 Blenderbot evaluation

We pretrain Blenderbot on Reddit and finetuned on Daily Dialog, ConvAI2, Emphatic Dialogue and Wizard of Wikipedia. For testing on Wizard of Wikipedia we left in the dataset only samples with "checked sentence" (gold grounding knowledge).

A.7 Perplexity-accuracy tradeoff

From one point of view, by telling the model exact values of some attributes we giving more information and making it easier to give more value to a correct answer, so perplexity should drop. Also, by adding adapters we add more weights making the model potentially stronger. But from the other side, we have a relatively small dataset for model to understand attributes perfectly, moreover, in our setup each adapter is trained only on part of dataset, corresponding to its attribute value. If there isn't much data, model will easily overfit (memorize some answers) and the control accuracy score will be high, so as perplexity, because model will have more confidence in memorized answers, we don't want this to happen. Including for this reason we may not want to fine-tune the entire model and are trying to keep its output significant on each layer (averaging adapters with a fixed weight of the main branch during blending, so the model can't ignore outputs from this branch). Experiments in Appendix A.4, Table 6, 7 show that increasing the main branch's blending weight often improves (lowers) perplexity, given that the main branch was not trained on this dataset, which means that initial language model carries important knowledge and pays attention to the dialog flow. From dialogs 15, 14 we can see, that model is able to incorporate desired attributes in context-related answers (not just a safe universal answers), which are unlikely presented in training dataset.

A.8 Ethical Considerations

Generative models are known for their ability to produce offensive content of any kind. We haven't done any additional filtering of datasets we use, and the initial models (DialogPT, Blenderbot) also have issues with producing potentially harmful content. Also in this work we control sentiment. For positive sentiment generation may be more polite, while for the negative sentiment toxicity and offence may appear more often. It's important to say, that our approach is aimed to guide model in desired direction,

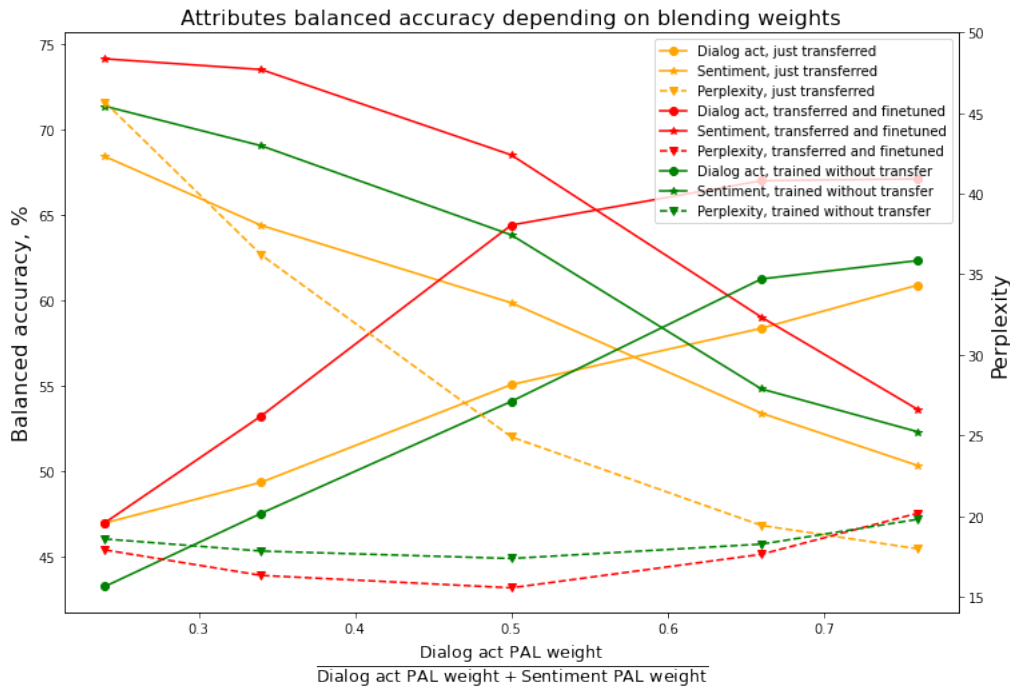


Figure 5: Attributes balanced accuracy and model perplexity depending on blending weights proportion of PAL for dialog act and PAL for sentiment. Perplexity is high for a model with high sentiment impact and just transferred weights because PALs for sentiment control were trained on a different dataset.

not to stop it from generating in undesired one.

A.9 Dialog examples

We providing some examples of dialogs with DialoGPT-small as an initial model and weighted average blending for dialog act and sentiment PALs. In Example 10 positive sentiment and question dialog act was set for the whole dialog, in Example 11 - negative sentiment and inform dialog act, and in Example 12 - positive sentiment and direct dialog act. In other examples (13, 14, 15) attributes were set dynamically before each system turn. Model follows desired direction in terms of attributes, but often ignores context and contradict itself. It is not an unexpected behaviour, since these problems are present in DialoGPT itself.

Sentiment control sometimes goes with emotional expressions ("Wow!", "How exciting!", or see Example 10), but the attitude of the whole utterance is also affected by selected sentiment. Inform and question dialog acts are the most popular and not so hard to master - they don't require deep semantic understanding, and model is good at asking questions and giving information (Examples 10, 11). But directive dialog act requires a bit more analysis (classifier also has problems with direct dialog act, see Appendix A.2), and model catches it too (see Examples 12, 13, 14).

Speaker	Utterance
USER:	Hello!
BOT [Question, Positive]:	Hi ! What a great moment . What’s up ?
USER:	Actually nothing, just want to talk.
BOT [Question, Positive]:	What happened ?
USER:	It’s rainy...
BOT [Question, Positive]:	How are you today ?
USER:	Pretty well
BOT [Question, Positive]:	Great ! Do you like the rain ?
USER:	Yes, I really love it.
BOT [Question, Positive]:	What’s the matter ?
USER:	Nothing)
BOT [Question, Positive]:	Oh , how exciting ! How’s the weather today ?

Table 10: Example of dialog with small DialoGPT model and PALs for sentiment and dialog act control, blended with weighted average. Positive sentiment and question dialog act were set for the entire conversation.

Speaker	Utterance
USER:	Hello!
BOT [Inform, Negative]:	Hi , I am George .
USER:	Can we meet?
BOT [Inform, Negative]:	I’m afraid that I can’t meet you .
USER:	Why? Are you busy?
BOT [Inform, Negative]:	No , I’m afraid I’m not .
USER:	So what happened?
BOT [Inform, Negative]:	My dad told me that I was going to move to Europe , and he wasn’t on the phone anymore .
USER:	Is your dad lost??? What happened?
BOT [Inform, Negative]:	He just dropped me off in the middle of the street and told me he was going to move to Europe .
USER:	So you are not going to move to Europe? Do you have any troubles here?
BOT [Inform, Negative]:	No .

Table 11: Example of dialog with small DialoGPT model and PALs for sentiment and dialog act control, blended with weighted average. Negative sentiment and inform dialog act were set for the entire conversation.

Speaker	Utterance
	USER: Hello!
BOT [<i>Directive, Positive</i>]:	Hello again , I'm afraid I've been here too long . I need to check your homework for me .
	USER: Okay, I've done it yesterday. Take a look.
BOT [<i>Directive, Positive</i>]:	I'm glad I did . Now do my homework .
	USER: Oh, that's a bit confusing. Don't you think, that homework should be done by you?
BOT [<i>Directive, Positive</i>]:	I think so.I think I will write it out myself. If you need a hand on it , ask me .
	USER: That's a good news! Actually, I need you to help me with my computer.
BOT [<i>Directive, Positive</i>]:	Let's try again .
	USER: What should we try again?
BOT [<i>Directive, Positive</i>]:	Just tell me again .
	USER: I need you to help me with my computer.
BOT [<i>Directive, Positive</i>]:	thank you .

Table 12: Example of dialog with small DialoGPT model and PALs for sentiment and dialog act control, blended with weighted average. Positive sentiment and direct dialog act were set for the entire conversation.

Speaker	Utterance
	USER: Hello!
BOT [<i>Inform, Positive</i>]:	Hello ! I am glad to hear that you like our soup !
	USER: Oh, actually I haven't said anything, but I really liked your soup!
BOT [<i>Question, Positive</i>]:	Soup ? I love soup .
	USER: Oh, okay..
BOT [<i>Directive, Positive</i>]:	I love soup ! It's full of flavour and it's healthy ! Can you give it some fresh banana too ?

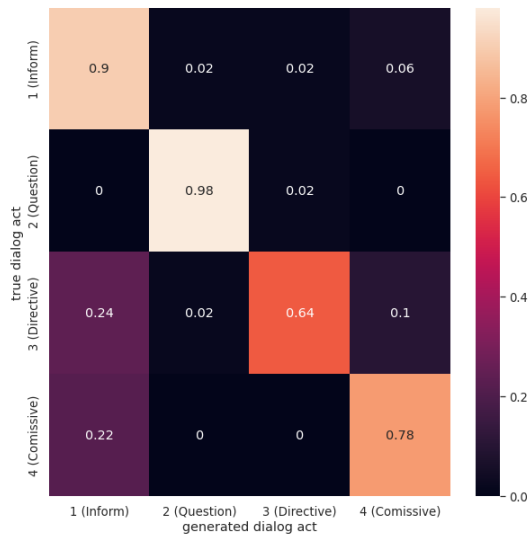
Table 13: Not very successful but common dialog with small DialoGPT model and PALs for sentiment and dialog act control, blended with weighted average. Sentiment and dialog act for bot response were selected before each bot's turn.

Speaker	Utterance
	USER: Do you want a cup of tea?
BOT [<i>Question, Positive</i>]:	Yes , of course . Do you have a large teethen tea ?
	USER: I'm afraid no, but we will serve you our best one!
BOT [<i>Directive, Positive</i>]:	Good , I will take the tea for free .
	USER: Oh, yes, of course
BOT [<i>Question, Positive</i>]:	That sounds really good . But , don't you want some tea ?
	USER: Yes, I will make one for myself too
BOT [<i>Directive, Positive</i>]:	Please do tell me when you arrive . I'll send the tea gift card to you .
	USER: Oh, thank you! How is it going?
BOT [<i>Inform, Negative</i>]:	You look like you're leaving .
	USER: No, I'm listening to you carefully.
BOT [<i>Inform, Positive</i>]:	No problem . I'm sorry . Thanks to you , bye !

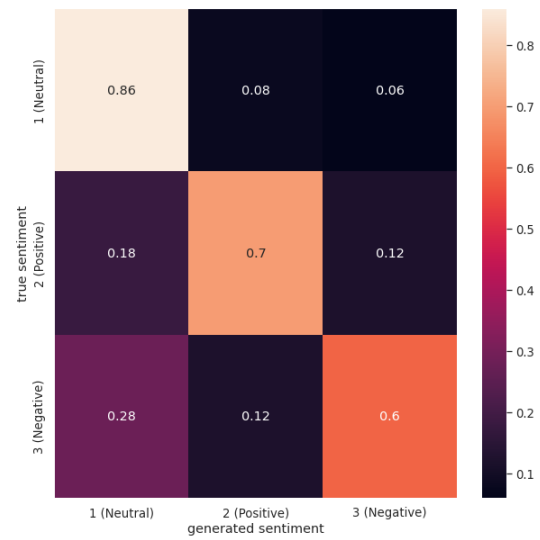
Table 14: Example of dialog with small DialoGPT model and PALs for sentiment and dialog act control, blended with weighted average. Sentiment and dialog act for bot response were selected before each bot's turn.

Speaker	Utterance
USER:	What did you eat today?
BOT [Inform, Positive]:	I ate a turkey leg and a banana .
USER:	Was it yummy?
BOT [Inform, Negative]:	It wasn't good , but my stomach was horrible .
USER:	I'm so sorry.
BOT [Question, Negative]:	What did you do ?
USER:	I made and ate a cake!
BOT [Question, Positive]:	Wow ! What did you make ?
USER:	I made a cheesecake.
BOT [Directive, Positive]:	I am so happy .
USER:	I'm happy too!
BOT [Directive, Positive]:	Me too ! There's a sweet strawberry on it . You can do it !

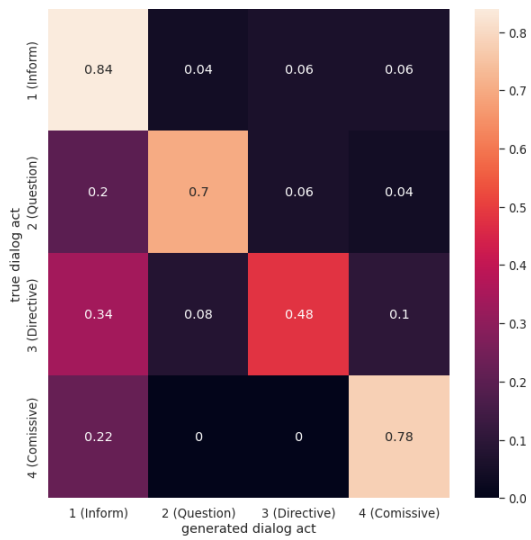
Table 15: Example of dialog with small DialoGPT model and PALs for sentiment and dialog act control, blended with weighted average. Sentiment and dialog act for bot response were selected before each bot's turn.



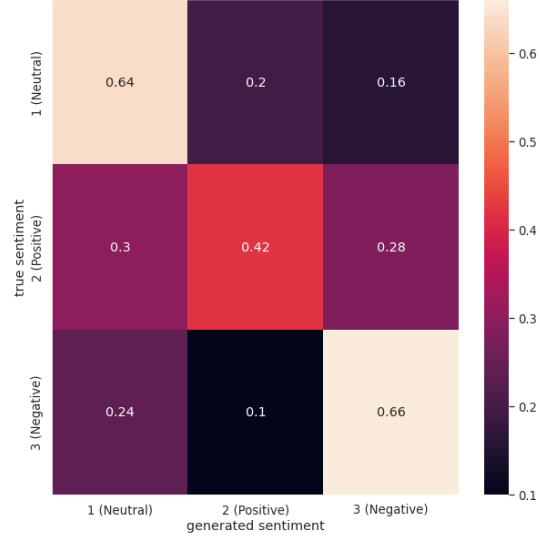
(a) Pretraining of PALs, dialog acts



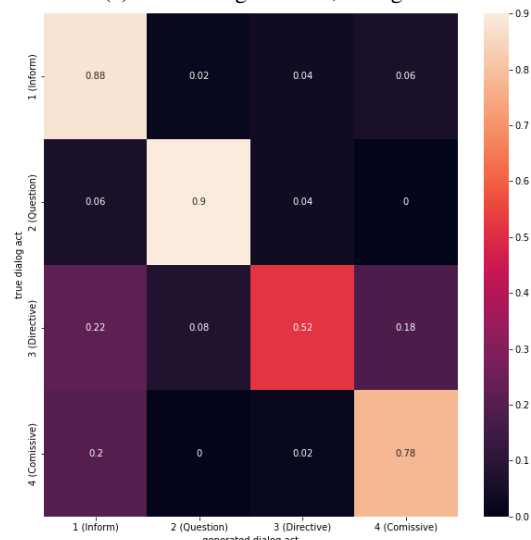
(b) Pretraining of PALs, sentiment



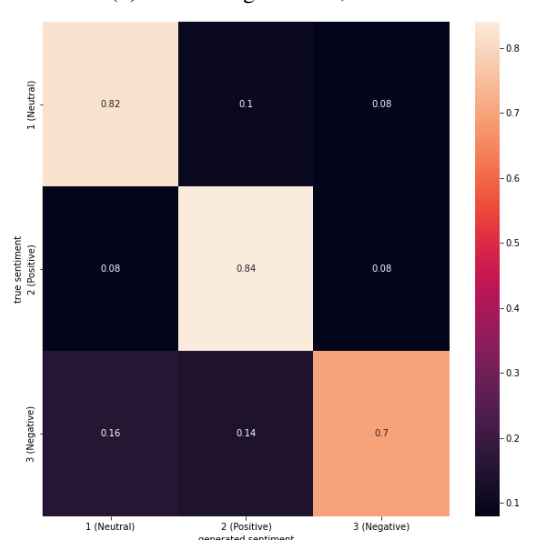
(c) Fine-tuning of PALs, dialog acts



(d) Fine-tuning of PALs, sentiment



(e) Different attributes in batch, dialog acts



(f) Different attributes in batch, sentiment

Figure 6: Comparison of different training methods

“Pears Film” live: Cognitive peculiarities of the reportage

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

Abstract

This study investigated whether there was a relationship between verbal working memory capacity and speech production in the genre of live reportage. The participants were 16 students of the Lomonosov Moscow State University. Participants' working memory capacity was assessed by means of the speaking span test [12]. Speech production was elicited by means of the “Pears Film” by W. Chafe [2]. Three aspects of speech production were assessed: continuity of the reportage, speech rate, and lexical diversity. Statistical analyses revealed that working memory capacity correlates positively with speech rate and lexical diversity, as predicted, and negatively with continuity of the reportage, contrary to predictions.

Keywords: reportage; working memory; speech production; speech rate

DOI: 10.28995/2075-7182-2022-21-203-210

В прямом эфире «Фильм о грушах»: когнитивные особенности репортажа

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

Аннотация

В этом исследовании изучалось, существует ли взаимосвязь между объемом вербальной рабочей памяти и порождением речи в жанре репортажа. В эксперименте участвовали 16 студентов МГУ имени М.В. Ломоносова. Объем рабочей памяти участников оценивался с помощью теста Speaking span [12]. В качестве стимульного материала был использован «Фильм о грушах» У. Чейфа [2]. Оценивались три аспекта порождения речи: непрерывность репортажа, скорость речи и лексическое разнообразие. Статистический анализ показал, что, как и ожидалось, объем рабочей памяти положительно коррелирует со скоростью речи и лексическим разнообразием, однако отрицательно с непрерывностью репортажа, вопреки ожиданиям.

Ключевые слова: репортаж; рабочая память; порождение речи; темп речи

1 Введение. Две предтечи из 1980 года

Предыстория данного исследования началась с двух больших успешных американских проектов конца XX века: первый («Фильм о грушах» У. Чейфа) был из области когнитивной лингвистики, второй (определение объема вербальной рабочей памяти М. Данеман) – из области экспериментальной психолингвистики.

1.1 «Фильм о грушах» Уолласа Чейфа

Коллективная монография под редакцией Чейфа «Рассказы о грушах: Когнитивные, культурные и языковые аспекты порождения повествования» (1980) справедливо считается одной из самых известных работ в области анализа дискурса. В этом исследовании носители разных языков смотрели, а затем пересказывали специально снятый для научных целей шестиминутный фильм [1]. Записи проводилось с испытуемыми разных возрастов, а также с различными временными интервалами между просмотром фильма и пересказом. Собранные более чем на 15 языках пересказы послужили основой многих теоретически важных работ, напр., [2]. За прошедшие сорок лет появился проект «Китайских рассказов о грушах» (www.pearstories.org), П. Остин изучал особенности глаголов индонезийского языка сасак [3], а в университете штата Нью-Мехико был создан корпус на американском, итальянском, каталанском и бразильском жестовых языках. Отдельно стоит отметить европейский проект на материале 12 языков по созданию телекомментария для незрячих людей [4].

Настоящее исследование открывает еще одну страницу использования в качестве стимульного материала «Фильма о грушах», а именно, жанр **репортажа**. Данный жанр еще мало изучен в лингвистике, однако см. книгу Е.Г. Малышевой (2011) о спортивном комментировании, в которой автор, в частности, выделяет семь коммуникативных типов языковых личностей спортивных комментаторов: (1) комментатор-«репортер», Виктор Гусев; (2) комментатор-«аналитик», Владимир Маслаченко; (3) комментатор-«знарок», Владимир Маслаченко; (4) комментатор-«ироник», Василий Уткин; (5) комментатор-«балагур», Георгий Черданцев; (6) комментатор-«болельщик», Георгий Черданцев; (7) комментатор-«актер», Н.Н. Озеров, Коте Махарадзе [5].

В психолингвистических исследованиях этот жанр, насколько нам известно, ранее почти не был задействован. В работе Р. Томлина похожий метод был использован для изучения фокусного внимания, залога и порядка слов в «Фильме о рыбах» [6], однако испытуемые порождали не дискурсы, а отдельные высказывания. В работе [7] при помощи похожего метода были получены детские пересказы, однако трехминутный видеоролик семь раз останавливался и детей просили рассказать, что произошло в текущем фрагменте. В диссертации Ю.Д. Потаниной была предпринята первая попытка подхода к этому вопросу; в качестве материала была выбрана компьютерная игра «Том и Джерри» [8], подробнее о результатах см. ниже.

1.2 Определение объема вербальной рабочей памяти Мередит Данеман

Когнитивные психологи занимаются изучением вербальной рабочей памяти (РП) уже более ста лет, последние сорок лет им в этом активно помогают психолингвисты, а последние 20 лет – и русскоязычные психолингвисты. Начало изучения РП связывается с именем Г. Эббингауза, разработавшего первые экспериментальные методы [9]; сам термин «рабочая память» (англ. *working memory*) был впервые использован в книге [10]. В психолингвистику этот термин ввела М. Данеман в [11]. В конце 1970-ых годов среди исследователей процессов **понимания** речи сложилась парадоксальная ситуация, когда интуитивно они были уверены, что индивидуальные различия в объеме РП должны оказывать влияние на речевые механизмы, однако проводимые ими эксперименты этого не подтверждали. Данеман и Карпентер (1980) предположили, что все дело в отсутствии адекватных методов и предложили тест, которому суждено было стать самым популярным на ближайшие полвека.

Данеман и Карпентер (1980) исходили из того, что в процессе интерпретации речи в РП человека происходят процессы, связанные как с пассивным хранением поступающей информации, так и с ее обработкой. Существовавшие же в то время методы по определению объема РП, состоявшие в запоминании отдельных цифр и/или слов, тестировали только первую из этих двух составляющих, нивелируя тем самым индивидуальные различия испытуемых. Они предложили новый тест, получивший название *Reading span*, который тестировал обе этих составляющих – в ходе эксперимента испытуемый должен был читать отдельные предложения и одновременно удерживать в РП последние слова ранее прочитанных предложений.

В конце 1980-ых годов М. Данеман разработала новый тест на определение объема РП, связанный уже с **порождением** речи, зеркальный тесту на понимание речи: испытуемые должны были запоминать отдельные слова, группируемые в блоки, а затем придумывать с ними предложения. В работе [12] автор показала, что объем РП коррелирует с беглостью речи при порождении. Данный результат был подтвержден на русском материале в диссертации Ю.Д. Потаниной [8].

2 Экспериментальное исследование

2.1 Гипотеза, испытуемые

Настоящее исследование было проведено в конце 2020 – начале 2021 годов с 16 студентами отделения ФиПЛ 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 РП при порождении речи (раздел 2.3). В данной работе мы проверяли гипотезу о том, что «успешность» комментирования «Фильма о грушах» коррелирует с объемом РП испытуемых. Эта гипотеза уже проверялась в работе [8]: Ю.Д. Потанина сравнила объем комментария, количество речевых сбоев и лексическое разнообразие комментариев, однако не нашла статистически значимых различий между испытуемыми с разным объемом РП. Как нам представляется, отсутствие корреляции было связано в первую очередь с неудачно подобранным стимульным материалом.

Итак, в данной работе мы выдвигаем гипотезу о положительной корреляции между «успешностью» репортажа и объемом вербальной РП при порождении речи и операционализируем «успешность» следующим образом:

- меньшее количество абсолютных (незаполненных) пауз хезитации
- более высокий темп речи
- большее лексическое разнообразие речи.

2.2 Сбор корпуса репортажей

Тест проводился онлайн на платформе Zoom. Испытуемые смотрели фильм длительностью 5 мин 55 с на экране своего компьютера и комментировали его второму участнику эксперимента, который выполнял роль слушающего. 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 всё, что происходит на экране. Старайтесь описывать не только происходящие события, но и окружающую обстановку.»

Все записи были расшифрованы и аннотированы в программе Praat, см. ниже пример расшифровки начала репортажа одного из испытуемых. Тексты были разбиты на дискурсивные фрагменты (которые не всегда соответствовали «классическим» ЭДЕ по [13]); при разбиении на дискурсивные фрагменты аннотаторы прежде всего руководствовались длительностью пауз.

[4.00 s]	Петухи закукарекали. (3.37 s)
[9.27 s]	Ветерок подул. (1.23 s)
[11.99 s]	Человек снимает грушу с дерева (0.2) груши. (5.86 s)
[20.86 s]	Человек в шляпе, с красным платком, снимает груши. (3.58 s)
[28.36 s]	Ну там вот много корзинок у него (0.2) с грушами. (2.25 s)
[33.37 s]	Он выкладывает ещё больше (0.4) груш в корзины. (3.75 s)
[41.02 s]	Из фартука. (3.71 s)
[46.21 s]	Ещё одну, ещё одну, ещё одну. (3.16 s)
[51.54 s]	Наклоняется. (2.05 s)
[54.92 s]	Какой-то странный звук (0.2) произошёл. (1.55 s)
[58.63 s]	Снимает платок, вытирает платком грушу. (3.37 s)

[64.66 s]	Сворачивает платок и надевает его обратно. (0.99 s)
[68.44 s]	Стоит на коленях.
[69.47 s]	Рядом появляется (0.4) человек с (0.5) кем, с овечкой?
[73.65 s]	Нет, кто это? (1.06 s)
[75.92 s]	А первый вс= снова поднимается на лестницу и идёт. (1.39 s)
[81.00 s]	А мимо проходит (0.4) человек с (1.8) козл'= козой?
[87.27 s]	Коза это? (1.14 s)
[89.38 s]	А, это телёнок!
[90.38 s]	Или... (4.47 s)
[95.35 s]	Да, они уходят.
[96.50 s]	Проходят мимо груш. (1.57 s)
[99.49 s]	Животное пытается вырваться, но не получается. (0.71 s)
[101.86 s]	А первый человек всё ещё собирает груши. (4.13 s)

2.3 Тест по определению объема вербальной РП при порождении речи

Тест также проводился онлайн на платформе Zoom. В ходе теста испытуемому предлагалось читать на экране отдельные слова русского языка, сгруппированные в блоки по 2, 3, 4, 5 или 6 слов, всего 100 слов, по 5 блоков каждого уровня. Между блоками слов были вставлены пустые слайды. Испытуемый должен был запоминать появляющиеся на экране слова, а при появлении пустого слайда придумывать с каждым из них по предложению, причем слова должны были стоять точно в той же форме, в которой они появлялись. Кроме того, целевое слово не могло стоять на первом месте придуманного предложения. Если испытуемый не мог вспомнить какие-то слова из текущего блока, он переходил к следующему блоку.

Для подсчета результатов был использован метод из работы Данеман и Карпентер (1980): испытуемый получает балл за очередной уровень, если были придуманы по крайней мере три из пяти групп предложений данного уровня, и полбалла, если правильно были придуманы только две из пяти групп. Результаты были расшифрованы и представлены в анкетах, см. ниже пример одного из последних блоков, включающего шесть слов.

страшно	Мне никогда так не было страшно, как в тот момент	+	+
исчезла	Она исчезла десять лет назад	+	
детства	У спортсменов не бывает детства	+	
молчали	Они молчали всю свою жизнь	+	
странно	Как ни странно, это произошло	+	
хороший	Этот человек не очень хороший	+	

Таблица 1: Фрагмент анкеты

2.4 Результаты

Данные были экспортированы в Excel, где при помощи нескольких скриптов были произведены необходимые подсчеты. В таблице 2 представлены сводные результаты данного исследования.

исп.	РП	паузы, с	слова	слова / с	уник.	уник_глаг.
1	2	87,9	560	2,1	206	41
2	2,5	142	302	1,4	126	42
3	2,5	89,2	543	2	172	49
4	3	31,9	381	1,2	173	48
5	3	112,9	470	1,9	172	57
6	3	68	477	1,7	192	51
7	3	99	568	2,2	196	49
8	3	54,8	766	2,6	219	53
	средн.	85,7*	508	1,89*	182	48,8*
9	4	141,4	370	1,7	142	55
10	4	196,6	348	2,2	157	49
11	4	125,8	616	2,7	216	61
12	5,5	83,5	508	1,9	202	52
13	5,5	61,6	800	2,7	280	64
14	6	152,9	399	2	150	56
15	6	129,8	469	2,1	173	53
16	6	173	508	2,8	190	65
	средн.	133,1*	502	2,27*	188,8	56,9*

Таблица 2: Результаты исследования (значимые различия выделены *, при $p=0,05$)

Испытуемые были разделены на две группы по 8 человек в зависимости от объема их РП: в группу с небольшим объемом РП попали испытуемые, получившие менее 4 баллов, а испытуемые, имеющие более 4 баллов, попали в группу с большим объемом РП.

В колонке «паузы» отмечена суммарная длительность в секундах абсолютных пауз (общая длительность фильма составила 355 с). Как видно из таблицы, мы получили значимую, но отрицательную корреляцию между объемом РП и паузацией.

В колонке «слова» дано суммарное количество слов, использованных в тесте; по данному показателю между двумя группами испытуемых нет статистически значимых различий.

Эти различия появляются, однако, если мы посмотрим на количество слов в секунду (колонка «слова / с»), которое оказывается выше у испытуемых с большим объемом РП.

Наконец, хотя количество уникальных слов в двух группах статистически значимо не различается (см. колонку «уник.»), нам удалось найти различие в количестве уникальных глаголов (колонка «уник_глаг.»): испытуемые с большим объемом РП использовали в своих репортажах статистически значимо больше разных глаголов, чем испытуемые с небольшим объемом.

2.5 Обсуждение результатов

Итак, в настоящей работе нам удалось обнаружить различия между репортажами испытуемых с большим и небольшим объемами РП по следующим переменным: (1) суммарная длительность абсолютных пауз, (2) среднее количество слов в секунду и (3) среднее количество уникальных глаголов. Рассмотрим далее эти различия более подробно.

(1) Непрерывность комментария. Важность данного критерия мы подчеркивали при инструктировании испытуемых, предлагая им представить, что рядом сидит незрячий человек, для которого необходимо как можно подробнее комментировать всё происходящее, а также приводя аналогию с футбольным репортажем по радио. Вопреки ожиданиям, испытуемые с небольшим объемом РП показали значимо лучшие результаты. Данный результат еще должен быть подтвержден в последующих исследованиях, однако мы исходим из того, что испытуемые с большим объемом РП по крайней мере не имеют по данному критерию преимущества перед испытуемыми с небольшим объемом. Как можно интерпретировать полученный результат?

При более широком взгляде на паузацию в естественном общении между людьми оказывается, что они стараются избегать слишком длинных абсолютных пауз, заполняя их всеми возможными способами, в том числе заполненными паузами хезитации (см., в частности, работу [14]). Более того, в работе [15] было показано, что в экспериментах, проведенных по методу референциальной коммуникации с танграммами, в диалогах детей 11-12 лет было зафиксировано значимо больше абсолютных пауз хезитации по сравнению с диалогами взрослых людей в аналогичном эксперименте (отметим, что с точки зрения незаполненных пауз хезитации такого различия между взрослыми испытуемыми и детьми зафиксировано не было). Таким образом, можно сделать вывод в духе постулатов Грайса, что с возрастом у человека формируется навык избегать длительных незаполненных пауз хезитации, которые потенциально могут привести к коммуникативной неудаче. Можно предположить, однако, что этот навык относится к коммуникативной сфере и не требует больших затрат когнитивных ресурсов, включая рабочую память. Возможно, в нашей выборке среди испытуемых с небольшим объемом РП оказалось большее количество «коммуникативно ответственных» людей. Однако, если последующие аналогичные эксперименты подтвердят более высокие результаты испытуемых с небольшим объемом РП, необходимо будет подумать о другой интерпретации.

(2) Темп речи. В данной работе темп речи подсчитывался следующим образом. Сначала для каждого испытуемого из длительности фильма в 355 секунд было вычтено суммарное время абсолютных пауз между дискурсивными фрагментами (эти паузы расположены в приведенном транскрипте на отдельных строчках; отметим, что абсолютные паузы внутри дискурсивных фрагментов не вычитались). Таким образом мы получили так называемое **время репортажа** каждого испытуемого. Затем для каждого испытуемого мы разделили количество слов на время репортажа, получив темп речи, измеряемый в количестве слов в секунду.

Хотя в научно-популярных книгах и/или лекциях об устройстве человеческой памяти часто звучит идея о том, что объем РП человека равен количеству слов, которые он успевает произнести за 2 секунды, строгих научных исследований о корреляции темпа речи и объема вербальной РП нам не известно. Таким образом, полученная положительная корреляция может считаться важным результатом данного исследования.

(3) Лексическое разнообразие. Вводя критерий лексического разнообразия, мы имели в виду тот факт, что чем больше уникальных полнозначных слов (существительных, глаголов, прилагательных и наречий) использует испытуемый, тем более разнообразно и «богато» будет выглядеть его репортаж. С другой стороны, существуют некоторые психолингвистические свидетельства о положительной корреляции между объемом вербальной РП и лексическим разнообразием, измеренным специальными психолингвистическими тестами. В частности, в работе [16] вербальная РП была измерена при помощи теста Данеман и Карпентер (1980), а лексическое разнообразие представляло собой скорее тестирование словарного запаса испытуемых: насколько испытуемые могут выбрать правильную картинку для названного экспериментатором слова (The Peabody

Picture Vocabulary Test) и насколько они могут дать правильное письменное определение тех или иных слов (WAIS-R Vocabulary Test).

В данном исследовании мы обнаружили, что у испытуемых с большим объемом РП наблюдается значимо большее разнообразие глаголов при примерно одинаковом количестве других частей речи. Данный результат также нуждается в верификации в последующих экспериментах, так что делать далеко идущие выводы пока рано. Однако интуитивно данный результат выглядит весьма перспективно с точки зрения возможных представлений о важности событийной структуры повествования.

3 Заключение. Перспективы будущих исследований

В данной работе мы ввели в психолингвистическое рассмотрение новый дискурсивный жанр – жанр репортажа. По нашим интуитивным представлениям, данный жанр требует от человека довольно больших когнитивных ресурсов (сравнимых, возможно, с когнитивными затратами синхронных переводчиков), поэтому мы начали исследование этого жанра с изучения особенностей вербальной рабочей памяти испытуемых. Для первого исследования мы выбрали три критерия «успешности» репортажа и предположили, таким образом, что «успешный» репортаж – это такой репортаж, в котором испытуемый говорит (1) непрерывно, (2) быстро и (3) лексически разнообразно. Статистический анализ репортажей 16 студентов МГУ имени М.В. Ломоносова показал, что, действительно, мы наблюдаем положительную корреляцию между объемом вербальной РП и темпом речи и лексическим разнообразием. Однако корреляция между объемом вербальной РП и непрерывностью репортажа вопреки ожиданиям оказалась отрицательной.

За рамками рассмотрения в настоящей работе осталось много других возможных критериев «успешности», например, разнообразные речевые сбои или содержательные особенности репортажей. Кроме того, полученный корпус репортажей может представлять интерес для исследователей-лингвистов и во многих других аспектах, например, с точки зрения изучения его просодической, событийной или дискурсивной структур.

Наконец, еще одним перспективным направлением, на наш взгляд, является использование в подобных исследованиях метода регистрации движений глаз, который позволяет с высокой точностью фиксировать, куда направлен взгляд испытуемого в каждый момент просмотра и/или комментирования фильма. В частности, в нашем текущем проекте мы сравниваем речевое и глазо-двигательное поведение испытуемых в двух режимах: репортажа, с одной стороны, и пересказа после просмотра фильма, с другой стороны. Кроме подтверждения результатов исследования, представленного в настоящей работе, мы предполагаем также получить данные об отличиях в глазо-двигательных паттернах испытуемых, выполняющих задание в разных дискурсивных жанрах. Еще одним многообещающим, на наш взгляд, направлением исследований является сравнение двух дискурсов от одних и тех же испытуемых: сначала репортажей, а через два-три месяца – пересказов фильма после просмотра. Данные эксперименты также уже проводятся в рамках нашего текущего проекта.

Благодарности

Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект». Автор выражает благодарность студентам 3 курса отделения ФиПЛ филфака МГУ имени М.В. Ломоносова, в первую очередь Марии Панич и Софье Соловьевой, за участие в экспериментах и помощь в обработке результатов.

References

- [1] Chafe W. (ed.). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. — Norwood: Ablex, 1980.
- [2] Chafe W. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. — Chicago, 1994.
- [3] Austin P.K. *Too many nasal verbs: Dialect variation in the voice system of Sasak*. — Nusa, 2012.
- [4] Mazur I., Chmiel A. *Towards common European audio description guidelines: Results of the Pear Tree Project // Perspectives: Studies in Translatology*. — Vol. 20(1). — P. 5–23.
- [5] Malysheva E.G. *Russian sports discourse [Russkiy sportivnyy diskurs]*. — Moscow: Flinta, 2011.
- [6] Tomlin R.S. *Focal attention, voice, and word order: an experimental, cross-linguistic study // P. Downing and M. Noonan (eds.) Word order in discourse*. — Amsterdam: Benjamins, 1995. — P. 517–554.
- [7] Eaton J. H., Collis G. M., Lewis V. A. *Evaluative explanations in children's narratives of a video sequence without dialogue // Journal of Child Language*. — 1999. — Vol. 26(3). — P. 699–720.
- [8] Potanina Yu.D. 2018. *Working memory and speech disfluencies in spontaneous discourse [Rabochaya pamyat' i rechevye sboi v spontannom discurs]*. — Dissertation. — Moscow, 2018.
- [9] Ebbinghaus H. *Über das Gedächtnis*. — 1885.
- [10] Miller G.A., Galanter E., Pribram K.H. *Plans and the structure of behavior*. — New York. — 1960.
- [11] Daneman M., Carpenter P.A. *Individual differences in working memory and reading // Journal of Verbal Learning and Verbal Behavior*. — Vol. 19.
- [12] Daneman M. *Working memory as a predictor of verbal fluency // Journal of Psycholinguistic Research*. — Vol. 20.
- [13] Kibrik A. A., Podlesskaya V. I. (eds.). *Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyah: korpusnoe issledovanie usntogo russkogo diskursa]*. — Moscow: Jazyki Slavyanskikh Kul'tur, 2009.
- [14] Fraundorf S.H., Watson D.G. *Alice's adventures in um-derland: Psycholinguistic dimensions of variation in disfluency production // Language, Cognition and Neuroscience*. — 2013. — Vol. 29. — P. 1083–1096.
- [15] Slabodkina T.A., Fedorova O.V. *Speech disfluencies in the interlocution of 10-12 year old native Russian-speaking children [Rechevye sboi v dialogah russkoyazychnyh detey Речевые сбои в диалогах русскоязычных детей] // Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta imeni A.I. Gercena*. — 2018. — Vol. 189. — P. 153–160.
- [16] Kemper S., Sumner A. *The structure of verbal abilities in young and older adults // Psychology and aging*. — 2001. — Vol. 16(2). — P. 312.

Argumentative Text Generation in Economic Domain

Fishcheva I. N.

Vyatka State University
Kirov, Russia
fishchevain@gmail.com

Osadchiy D.

ITMO University
Saint Petersburg, Russia
da.osadchiy@gmail.com

Bochenina K. O.

ITMO University
Saint Petersburg, Russia
kbochenina@itmo.ru

Kotelnikov E. V.

Vyatka State University
Kirov, Russia
kotelnikov.ev@gmail.com

Abstract

The development of large and super-large language models, such as GPT-3, T5, Switch Transformer, ERNIE, etc., has significantly improved the performance of text generation. One of the important research directions in this area is the generation of texts with arguments. The solution of this problem can be used in business meetings, political debates, dialogue systems, for preparation of student essays. One of the main domains for these applications is the economic sphere.

The key problem of the argument text generation for the Russian language is the lack of annotated argumentation corpora. In this paper, we use translated versions of the Argumentative Microtext, Persuasive Essays and UKP Sentential corpora to fine-tune RuBERT model. Further, this model is used to annotate the corpus of economic news by argumentation. Then the annotated corpus is employed to fine-tune the ruGPT-3 model, which generates argument texts. The results show that this approach improves the accuracy of the argument generation by more than 20 percentage points (63.2% vs. 42.5%) compared to the original ruGPT-3 model.

Keywords: Argumentation Mining, Text Generation, ruGPT-3, RuBERT, XGBoost

DOI: 10.28995/2075-7182-2022-21-211-222

Генерация аргументативных текстов экономической тематики

Фищева И. Н.

Вятский государственный университет
Киров, Россия
fishchevain@gmail.com

Осадчий Д.

Университет ИТМО
Санкт-Петербург, Россия
da.osadchiy@gmail.com

Боченина К. О.

Университет ИТМО
Санкт-Петербург, Россия
kbochenina@itmo.ru

Котельников Е. В.

Вятский государственный университет
Киров, Россия
kotelnikov.ev@gmail.com

Аннотация

Разработка больших и сверхбольших языковых моделей, таких как GPT-3, T5, Switch Transformer, ERNIE и др., позволила в последнее время значительно повысить качество генерации текстов. Одним из важных направлений в этой области является порождение текста с аргументами. Решение такой задачи может быть использовано при проведении деловых совещаний, в политических дебатах, в диалоговых системах, при подготовке студенческих эссе. Одной из основных предметных областей в указанных приложениях является экономическая сфера.

Ключевой проблемой при генерации аргументов для русского языка является дефицит корпусов, размеченных по аргументации. В настоящей работе мы используем переводные версии корпусов Argumentative Microtext, Persuasive Essays и UKP Sentential для дообучения модели RuBERT. Далее построенная модель используется для разметки по аргументации собранного корпуса экономических новостей. Затем размеченный корпус применяется для дообучения модели ruGPT-3, которая порождает аргументационные тексты. Результаты показывают, что такой подход позволяет повысить правильность генерации аргументов более чем на 20 процентных пунктов (63.2% vs. 42.5%) по сравнению с исходной моделью ruGPT-3.

Ключевые слова: анализ аргументации, генерация текстов, BERT, GPT

1 Introduction

Automatic text generation has recently made impressive progress with the development of large and super-large pre-trained language models (Han et al., 2021), such as GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), Switch Transformer (Fedus et al., 2021), ERNIE (Sun et al., 2021) and others. These models allow tuning to the problem by updating the weights on a small training dataset (fine-tuning) or without updating weights in few-shot learning (with several training examples, usually from 10 to 100), one-shot learning (one training example) and even zero-shot learning (Brown et al., 2020).

One of the important directions in this field is the generation of texts with arguments (Hua et al., 2020; Gretz et al., 2020; Schiller et al., 2021). Arguments in this case are either found using an information retrieval system (Hua et al., 2020), or generated by a pre-trained language model (Gretz et al., 2020; Al Khatib et al., 2021). An *argument* is a set of statements that includes claim and premises (Stede and Schneider, 2018). A *claim* is a statement that expresses a potentially controversial point of view. *Premises* are statements supporting or refuting this claim.

Systems that generate text with arguments related to a given claim can be used during business meetings to promptly generate arguments; in political debate to back up the position of a speaker; in jurisprudence to search and generate arguments for legislative acts and precedents; in dialogue systems for the selection of arguments to convince the interlocutor; in education for analyzing, generating, and evaluating arguments in student essays. One of the most common domains for these applications is economics.

There are several publicly available pre-trained language models for the Russian language, including RuBERT (Kuratov and Arkhipov, 2019), SBERT (SBERT, 2020), and ruGPT-3 (ruGPT-3, 2020). Recently, several papers have appeared on the argument mining in Russian (Fishcheva and Kotelnikov, 2019; Fishcheva et al., 2021; Salomatina et al., 2021; Ilina et al., 2021).

However, to the best of our knowledge, there have been no work devoted to the generation of argumentative Russian texts so far. We aim to close this gap and propose an approach to build a language model for generating an argumentative text in Russian in the economic domain.

The contribution of this work is as follows:

- a new approach to the generation of Russian argumentative texts is proposed. In this approach, the model RuBERT is first trained using argumentative corpora, which is then used to automatically annotate arguments in an economic text corpus. The annotated corpus is used to fine-tune the ruGPT-3 model, which allows generating texts with arguments, particularly, sentences containing premises related to a given claim;
- presenting the results of experiments with the proposed approach;
- sharing the corpus of pairs $\langle \textit{claim}, \textit{sentence} \rangle$ on economic domain. The corpus includes 660 sentences manually annotated into two classes – “premise” / “non-premise”.

2 Previous Work

2.1 Argumentative Text Generation

Argumentative texts generation includes the following research directions.

1. Application of traditional text generation architectures, including content selection/organization and linguistic implementation components (Carenini and Moore, 2006).
2. Generation of an argumentative text from given small argumentative elements (Walton and Gordon, 2012; Reisert et al., 2015; Wachsmuth et al., 2018; El Baff et al., 2019).
3. Replacing target objects in extracted arguments (Bilu and Slonim, 2016).

4. Using predefined argument templates (Bilu et al., 2020).
5. Application of the encoder-decoder architecture (seq2seq) (Hidey and McKeown, 2019), possibly supplemented by an information retrieval component (Hua et al., 2020; Schiller et al., 2021).
6. Using pre-trained language models for text generation, such as GPT-2 (Gretz et al., 2020; Al Khatib et al., 2021).

The papers closest to our research are (Schiller et al., 2021; Gretz et al., 2020; Al Khatib et al., 2021).

Schiller et al. (Schiller et al., 2021) proposed a neural network model Arg-CTRL for controlled generation of texts based on the well-known CTRL model (Keskar et al., 2019). Controlling the text generation is carried out using the so-called *control codes*, including the topic, stance, and aspect of the argument. The retrieval system and the BERT model were used to extract aspects. In our work, instead of the CTRL model, we use the ruGPT-3 model, which does not require learning from scratch for the Russian language, like CTRL. Also, the BERT model is used not to extract aspects, but to select sentences containing premises.

Gretz et al. (Gretz et al., 2020) investigated claim generation. They suggested using the fine-tuning of the neural network model GPT-2 (Radford et al., 2020) based on the corpus of automatically extracted arguments. Moreover, authors added context to the input data for GPT-2 in order to improve the quality of generation. In our work, we generated premises, not claims. Also, instead of GPT-2, which is mainly an English-language model, we use the Russian-language ruGPT-3 model.

Al-Khatib et al. (Al Khatib et al., 2021) used argumentation knowledge graphs to generate argumentative texts. The concepts contained in such graphs are searched on debate portals and Wikipedia. The found fragments of texts are used for fine-tuning GPT-2. In our work, we use the RuBERT model rather than knowledge graphs to generate training data. In addition, instead of GPT-2 we use the Russian-language model ruGPT-3.

2.2 Argumentation Mining in Russian

The area of argumentation mining for the Russian language has recently attracted more and more attention of researchers.

The first annotated corpus for the Russian language (Fishcheva and Kotelnikov, 2019) was created based on the translation of the English language Argumentative Microtext Corpus (ArgMicro) (Peldszus and Stede, 2015; Skeppstedt et al., 2018). It was then expanded with machine translation of the Persuasive Essays Corpus (PersEssays) (Stab and Gurevych, 2014) and a Joint Argument Annotation Scheme was proposed (Fishcheva et al., 2021). By using XGBoost and BERT, the authors were able to improve the results of automatic classification of “for” / “against” premises.

Salomatina et al. (Salomatina et al., 2021) proposed a method for finding an argumentative structure based on using the patterns of argumentation indicators and their role in the thematic structure of the text. This method can be used in the absence of a sufficient amount of annotated data.

Ilina et al. (Ilina et al., 2021) presented a web resource designed to study argumentation in popular science discourse. A distinctive feature of the argument annotation model is the using of argument persuasiveness weighting. The annotation script includes several procedures that allow the annotator to check the quality of text annotation and evaluate the persuasiveness of the argumentation.

In our work, to the best of our knowledge, the problem of generating argumentative texts in Russian is being investigated for the first time. We are expanding the Russian-language corpus from (Fishcheva et al., 2021) by translating the UKP Sentential Argument Mining Corpus (UKP Sentential) (Stab et al., 2018). Based on the extended corpus, we train RuBERT model, which we then use to annotate sentences of the economic news corpus. The annotated corpus is used to fine-tune the ruGPT-3 model to generate texts containing premises for a given claim.

3 Proposed approach

The scheme of the proposed approach for generating argumentative texts is shown in Fig. 1.

The three argumentative annotated corpora – ArgMicro, PersEssays, and UKP Sentential (see Sec-

tion 4) – serve as input data. They are used to fine-tune the pre-trained RuBERT¹ model (Kuratov and Arkhipov, 2019). This model showed the best results (along with XGBoost² model) in the argument classification task (Fishcheva et al., 2021). We also compared the performance of RuBERT with XGBoost on “premise” / “non-premise” classification task at the sentence level. As a result, RuBERT outperformed XGBoost (see Subsection 5.1) and further we used the RuBERT model.

We collected the corpus of economic news from the internet (see Section 4) – it is denoted as “Un-labeled economic corpus” on the scheme. The fine-tuned RuBERT classifies economic corpus sentences so we obtain economic corpus labeled by “premise” / “non-premise” (see details in Subsection 5.1).

At the next step, fine-tuning of the pretrained ruGPT-3 (ruGPT-3, 2020) model is carried out. Firstly, we choose 3,500 sentences (it is about 5% from the whole corpus) with the highest probability scores of classification as “premise” from RuBERT. Then we take pretrained ruGPT-3 model and fine-tune it on selected 3,500 sentences (see details in Subsection 5.2). To evaluate and compare the accuracy of the original model and the fine-tuned one, 20 prompts were used, for each prompt both models generated 20 potential premises. The 800 (400+400) received sentences were annotated manually (see details in Subsection 5.3).

4 Text Corpora

In our study, we used three existing argument-annotated corpora – ArgMicro, PersEssays, and UKP Sentential, as well as a new economic news corpus.

The Argumentative Microtext Corpus (ArgMicro) was proposed in (Peldszus and Stede, 2015; Skeppstedt et al., 2018). The corpus includes 283 texts on various topics (raising the retirement age, health insurance, school uniforms, etc.). Each text contains one claim about a topic and from 2 to 10 ADUs (argumentative discourse units), annotated as premises “for” or “against” this claim. An ADU is a fragment of text that has a single argumentative value (Stede and Schneider, 2018). In the ArgMicro, an ADU can be either an entire sentence or part of a sentence. Further, we consider each individual ADU as a sentence.

The Persuasive Essays Corpus (PersEssays) was introduced in (Stab and Gurevych, 2014). It contains 399 texts on a wide range of topics (school education, immigration, economic policy, etc.). The ADUs are sentences. The annotation of sentences is done according to four types: main claim, claim, premise, and neutral element. When forming the training corpus for the classification “premise” / “non-premise”, the main claims were excluded, just claims and the actual premises were used as “premises”, and neutral elements – as “non-premises”.

The UKP Sentential Argument Mining Corpus (UKP Sentential) was proposed in (Stab et al., 2018). The corpus includes 25,492 sentences, annotated as “for”, “against” or “non-premise” in relation to one of the eight topics (abortion, cloning, the death penalty, etc.). The corpus contains 489 sentences, the annotation of which is different for different topics. Such sentences were excluded.

(Fishcheva and Kotelnikov, 2019) showed that the best result among the Google Translate, Yandex.Translate and Prompt systems was demonstrated by Google Translate during machine translation of the ArgMicro corpus from English into Russian. Therefore, all three English corpora were translated into Russian using Google Translate.

The characteristics of all the corpora with argumentative annotation are shown in Table 1.

For fine-tuning of the ruGPT-3 model in economics, we used the corpus of Russian-language economic news, collected from the website *Banki.ru*³. The original corpus included 7,759 texts for the period from 01.06.2019 to 12.07.2021. The texts were segmented into sentences using the *Stanza*⁴ library (average text size – 9 sentences) and tokenized using the *NLTK*⁵ library. Further, 10% of the longest and shortest sentences, as well as repetitive sentences were removed. The preprocessing resulted in 68,859 sentences with an average length of 15 words.

¹<https://huggingface.co/DeepPavlov/rubert-base-cased>

²<https://xgboost.readthedocs.io>

³<https://www.banki.ru>

⁴<https://stanfordnlp.github.io/stanza>

⁵<https://www.nltk.org>

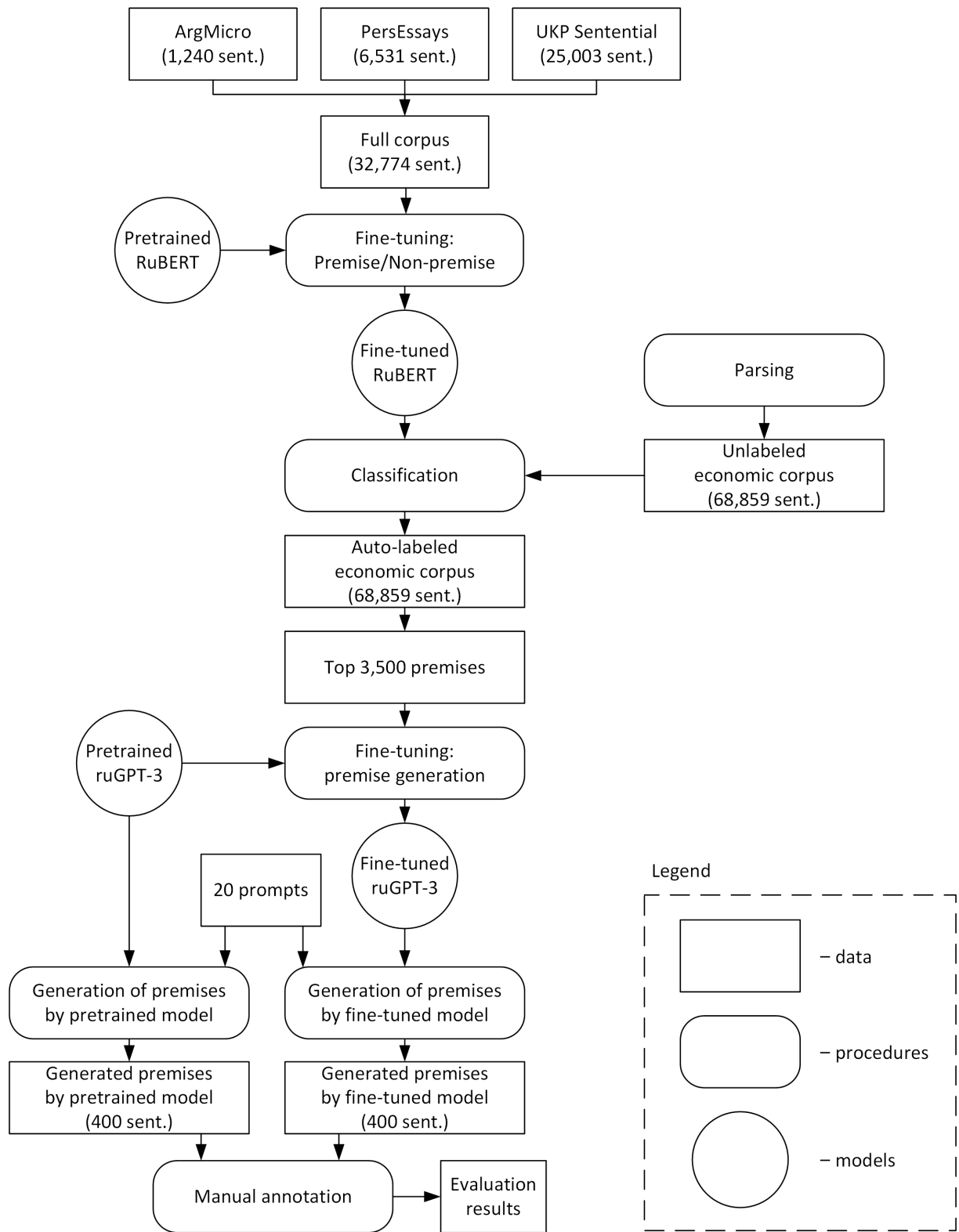


Figure 1: Scheme of the proposed approach.

Table 1: Characteristics of corpora with argumentative annotation.

Corpora	Premise	Non-premise	Total
ArgMicro	1,236	4	1,240
PersEssays	5,302	1,229	6,531
UKP Sentential	11,023	13,980	25,003
Total	17,561	15,213	32,774

5 Result and Discussion

5.1 “Premise” / “Non-premise” Classification

At the first step, we compared the performance of RuBERT and XGBoost, which were the best models in (Fishcheva et al., 2021). To this end, we fine-tuned the pre-trained RuBERT model and we trained the XGBoost classifier on the union of annotated corpora ($\text{ArgMicro} \cup \text{PersEssays} \cup \text{UKP Sentential}$). The problem of classifying sentences into two classes – “premise” and “non-premise” – was set. We used the same 5-fold cross-validation for both models to obtain performance scores. For selection of hyperparameters for XGBoost, 4-fold nested cross-validation was used, for RuBERT – hold-out dataset (validation 20%). The hyperparameters are as follows:

- XGBoost⁶: number of trees = [50, 150, **500**], maximum tree depth = [2, 8, 20, 30];
- RuBERT: number of epochs = [3, **5**], batch size = [4, **8**], learning rate = 10^{-5} .

To train RuBERT, we used class *BertForSequenceClassification* from the *transformers*⁷ library with the following settings: *AdamW* optimizer with default parameters, weight decay = 0.01, context length = 512 tokens.

Following (Fishcheva and Kotelnikov, 2019), we used three types of features for XGBoost:

- lexical features – discursive markers (e.g., “therefore”, “influence”, “firstly”, etc.) and modal words (e.g., “should be”, “can”, “want”, etc.), including negations, 255 features in total;
- punctuation features – comma, colon, semicolon, question and exclamation marks, 5 features in total;
- morphosyntactic features – N-grams based on parts of speech (nouns, pronouns, verbs, adjectives, and adverbs), $N = 2, 3, 4$ (e.g., ADJ+ADJ+NOUN, ADV+VERB), and grammatical features of verbs: time, inclination, person; there are 783 features in total. Morphological analysis was carried out with *Mystem*⁸.

Unlike (Fishcheva and Kotelnikov, 2019), we have not added contextual features of the previous and next sentences, since the UKP Sentential corpus contains only individual sentences.

Cross-validation results for the union corpus ($\text{ArgMicro} \cup \text{PersEssays} \cup \text{UKP Sentential}$) are presented in Table 2.

Table 2: Performance scores for “premise” / “non-premise” classification: macro-averaged F1-score, Precision, Recall and Accuracy (Mean \pm Std Dev)

Model	F1-score	Precision	Recall	Accuracy
XGBoost	0.6800 \pm 0.0066	0.6817 \pm 0.0065	0.6796 \pm 0.0066	0.6832 \pm 0.0064
RuBERT	0.7903\pm0.0051	0.7901\pm0.0051	0.7908\pm0.0050	0.7911\pm0.0051

XGBoost is significantly inferior to RuBERT in this task, in contrast to the results of (Fishcheva et al., 2021), where both models showed comparable performance. This is due to the fact that contextual fea-

⁶Optimal values of the hyperparameters are highlighted in bold – they turned out to be the same on all folds.

⁷https://huggingface.co/docs/transformers/model_doc/bert

⁸<https://yandex.ru/dev/mystem>

tures, which are very important according to (Fishcheva et al., 2021), were not used in training XGBoost in our study.

Thus we decided to take RuBERT as the classification model. After performance assessment, RuBERT with the selected optimal hyperparameters was fine-tuned on the union corpus. The obtained model was used to classify 68,859 sentences of the economic corpus.

5.2 ruGPT-3 Fine Tuning

To build a model for generating argumentative texts, fine-tuning of the pretrained model ruGPT-3 (ruGPT-3, 2020) was held. To this end, 3,500 sentences (approx. 5%) from the economic corpus annotated by the RuBERT model as “premise” with the highest confidence scores were used as training data.

We fine-tuned the ruGPT3Large version (760M parameters) using NVIDIA RTX A6000 video card. The training sentences, which were prefixed with “*потому что*” (“*because*”), were fed individually to the input of the model. We tuned hyperparameters using hold-out strategy: 3,000 sentences were used as training dataset, 500 sentences – as validation dataset. The perplexity was the performance metric.

The hyperparameters were as follows:

- number of epochs = [1..5],
- batch size = [1, 2, 4, 8, 12, 16],
- learning rate = [10^{-6} , 10^{-5} , $5 \cdot 10^{-5}$, 10^{-4}].

The following hyperparameters turned out to be the best: number of epochs 1, batch size 12, learning rate $5 \cdot 10^{-5}$. Perplexity with these hyperparameters is equal to 9.66.

At the next step, the fine-tuned ruGPT-3 model was tested by manual annotation of the generated arguments.

5.3 Manual Annotation of Generated Premises

To test the fine-tuned ruGPT-3 model, we used 20 prompting claims (after each claim, suffix “*потому что*” – “*because*” was added) (Table 3). The original ruGPT3Large model was used as the baseline (zero-shot learning). For each prompt (with suffix “*потому что*”), each model generated 20 sentences – potential premises. The following parameters were used for generation in both models: *Top-K*=50, *Top-p*=0.92. Thus, each model generated 400 sentences.

The overall generated 800 sentences were manually annotated by four annotators. Annotators were provided with *<claim, sentence>* pairs with no information about which model generated the sentence. An argument was considered to be a sentence that could be used to convince an opponent of a given claim. The final decision about the sentence was based on a simple voting of the annotators’ scores – the label was approved if at least 3 from 4 annotators agreed.

As a result, the labels were assigned to 660 sentences (82.5% from 800 sentences) (Table 4): 321 for the fine-tuned model and 339 for the original model. The agreement of the annotators calculated by Krippendorff $\alpha=0.4772$. This level of agreement corresponds to moderate agreement on the Landis & Koch scale (Landis and Koch, 1977).

After manual annotation, the accuracy of the fine-tuned ruGPT-3 model was 63.2% (203 premises out of 321 sentences), and the accuracy of the original model was 42.5% (144 premises out of 339 sentences). Examples of premises generated by both models are shown in Table 5 (English version is presented in Table 6).

5.4 Ethical Considerations

Since the proposed method is not controlled, the model may distort the meaning of the texts used in its fine-tuning. Therefore, the generated arguments may contain incorrect information and mislead users who are not experts. At the same time, large language generation models, including the RuGPT-3 model used in the proposed method, make it possible to generate texts that are indistinguishable from texts written by a person. Given the persuasive nature of the arguments generated by the proposed method, they can be used to deceive or even spread personalized disinformation (Zellers et al., 2019; Schuster et al., 2020). We hope that such technologies will be used only for peaceful purposes to improve and

Table 3: Prompting claims

No.	Prompting claims (Russian)	English translation
1	Банкам следует более широко использовать биометрию	Banks should use biometrics more widely
2	Вместо потребительского кредита лучше взять автокредит	Instead of a consumer loan, it is better to take a car loan
3	В случае высокой инфляции нужно вкладываться в драгоценные металлы	In case of high inflation, you need to invest in precious metals
4	Государственные облигации являются одним из наиболее надежных видов ценных бумаг	Government bonds are one of the most reliable types of securities
5	Деньги нужно вкладывать в акции	Money should be invested in stocks
6	Деньги нужно вкладывать в облигации	Money should be invested in bonds
7	Для сохранения финансов оптимальнее всего использовать депозиты	To save finances, it is best to use deposits
8	Криптовалюты лучше фиатных валют	Cryptocurrencies are better than fiat currencies
9	Лучшая инвестиционная стратегия - негосударственные пенсионные фонды	The best investment strategy - non-state pension funds
10	Лучшей инвестицией является покупка недвижимости	The best investment is the purchase of real estate
11	Наиболее выгодно вкладываться в голубые фишки	The most profitable investment in blue chips
12	Не следует играть на валютном рынке	You should not play in the foreign exchange market
13	Обучение финансовой грамотности зачастую приводит к необоснованной уверенности	Financial literacy training often leads to unwarranted confidence
14	Покупка земельного участка является хорошей инвестиционной стратегией	Buying a plot of land is a good investment strategy
15	При оформлении кредитной карты стоит внимательно отнестись к выбору банка	When applying for a credit card, you should carefully consider the choice of bank
16	Сбережения следует хранить в валюте	Savings should be kept in foreign currency
17	Сбережения следует хранить в долларах	Savings should be kept in dollars
18	Сбережения следует хранить в евро	Savings should be kept in euros
19	Сбережения следует хранить в рублях	Savings should be kept in rubles
20	Храните деньги в той валюте в которой получаете зарплату	Keep money in the currency in which you receive your salary

Table 4: Results of premise generation

Model	Premise	Non-premise	Sum
Original model	144 / 42.5%	195 / 57.5%	339 / 100%
Fine-tuned model	203 / 63.2%	118 / 36.8%	321 / 100%
Overall	347 / 52.6%	313 / 47.4%	660 / 100%

simplify communications in society, and that the openness and independence of the development of such models will not only ensure equal access to them, but also stimulate the development of methods for detecting the use of such models.

Table 5: Examples of text generation (in Russian): claims (suffix “*потому что*” is not shown); generated sentences – premises; annotator scores’ voting results.

Claims	Sentences	Voting
<i>Банкам следует более широко использовать биометрию</i>	Fine-tuned ruGPT-3: <i>она позволяет упростить процедуру взаимодействия с клиентами и упрощает проведение платежа.</i>	4 from 4
	Original ruGPT-3: <i>она может значительно упростить получение и обработку данных о клиентах.</i>	4 from 4
<i>Государственные облигации являются одним из наиболее надежных видов ценных бумаг</i>	Fine-tuned ruGPT-3: <i>они не подвержены обесценению, не утрачивают своей ликвидности, их стоимость остается стабильной.</i>	4 from 4
	Original ruGPT-3: <i>они обеспечены государственной гарантией, которая включает в себя государственный долг.</i>	4 from 4
<i>Обучение финансовой грамотности зачастую приводит к необоснованной уверенности</i>	Fine-tuned ruGPT-3: <i>человек, желающий научиться работать с цифровыми активами, начинает думать, что он сможет легко ими управлять.</i>	3 from 4
	Original ruGPT-3: <i>люди не в состоянии проверить, насколько достоверны и правильны их утверждения.</i>	3 from 4

6 Conclusion

In our study, we proposed an approach for creating a Russian-language model of premise generation for a given claim in the economic domain. First, we compared the RuBERT and XGBoost models on the translated argument corpora ArgMicro, PersEssays, and UKP Sentential. The performance of RuBERT turned out to be better. The fine-tuned RuBERT model was used to annotate sentences from a corpus of economic news. The ruGPT-3 model was fine-tuned on the annotated sentences. The model’s ability to generate argumentative premises was tested using manual annotation. The accuracy of generation of the fine-tuned model was more than 20 percentage points higher than that of the original model (63.2% vs. 42.5%), which confirms the effectiveness of the proposed approach. We made this corpus of 660 manually annotated pairs *<claim, sentence>* publicly available⁹.

In our opinion, the achieved quality of argumentative text generation by the fine-tuned models allows us to speak about the possibility of using such models in practice, for example, for online generation of arguments during business meetings.

As perspective areas of research, we plan to study the input context influence on GPT-type models quality.

Acknowledgements

This work was supported by Russian Science Foundation, project № 22-21-00885, <https://rscf.ru/en/project/22-21-00885/>.

⁹https://github.com/kotelnikov-ev/economic_argument_generation

Table 6: Examples of text generation (English version of Table 5): claims (suffix “because” is not shown); generated sentences – premises; annotator scores’ voting results.

Claims	Sentences	Voting
<i>Banks should use biometrics more widely</i>	<u>Fine-tuned ruGPT-3:</u> <i>it allows you to simplify the procedure of interaction with customers and simplifies the payment process.</i>	4 from 4
	<u>Original ruGPT-3:</u> <i>it can greatly simplify the acquisition and processing of customer data.</i>	4 from 4
<i>Government bonds are one of the most reliable types of securities</i>	<u>Fine-tuned ruGPT-3:</u> <i>they are not subject to depreciation, do not lose their liquidity, their value remains stable.</i>	4 from 4
	<u>Original ruGPT-3:</u> <i>they are backed by a state guarantee, which includes the public debt.</i>	4 from 4
<i>Financial literacy training often leads to unwarranted confidence</i>	<u>Fine-tuned ruGPT-3:</u> <i>a person who wants to learn how to work with digital assets begins to think that he can easily manage them.</i>	3 from 4
	<u>Original ruGPT-3:</u> <i>people are not able to check how reliable and correct their statements are.</i>	3 from 4

References

- Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing Argumentation Knowledge Graphs for Neural Argument Generation. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, P 4744–4754.
- Yonatan Bilu and Noam Slonim. 2016. Claim Synthesis via Predicate Recycling. // *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) - Short Papers*, P 525–530.
- Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, and et al. 2020. Argument Invention from First Principles. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, P 1013–1026.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and et al. 2020. Language Models are Few-Shot Learners. // *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, P 1877–1901.
- Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952.
- Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, Manfred Stede, and Benno Stein. 2019. Computational Argumentation Synthesis as a Language Modeling Task. // *Proceedings of the 12th International Conference on Natural Language Generation*, P 54–64.
- William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity.
- Irina Fishcheva and Evgeny Kotelnikov. 2019. Cross-Lingual Argumentation Mining for Russian Texts. // *Proceedings of the 8th International Conference “Analysis of Images, Social networks and Texts” (AIST 2019), Lecture Notes in Computer Science*, P 134–144.
- Irina Fishcheva, Valeriya Goloviznina, and Evgeny Kotelnikov. 2021. Traditional Machine Learning and Deep Learning Models for Argumentation Mining in Russian Texts. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog-2021”*, P 246–258.

- Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family – A Study of GPT-2 Based Claim Generation. // *Findings of the Association for Computational Linguistics (EMNLP)*, P 528–544.
- Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, and et al. 2021. Pre-Trained Models: Past, Present and Future. *AI Open*.
- Christopher Hidey and Kathleen McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. // *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, P 1756–1767.
- Xinyu Hua, Zhe Hu, and Lu Wang. 2020. Argument generation with retrieval, planning, and realization. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, P 2661–2672.
- D Ilina, I Kononenko, and E Sidorova. 2021. On Developing a Web Resource to Study Argumentation in Popular Science Discourse. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog-2021”*, P 318–327.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog-2019”*, P 333–340.
- J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.
- Andreas Peldszus and Manfred Stede. 2015. An Annotated Corpus of Argumentative Microtexts. // *Proceedings of the 1st European Conference on Argumentation*, P 801–815.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2020. Language Models are Unsupervised Multitask Learners. *OpenAI Technical report*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, and et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.
- Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A Computational Approach for Generating Toulmin Model Argumentation. // *Proceedings of the 2nd Workshop on Argumentation Mining*, P 45–55.
- ruGPT-3. 2020. pretrained language model. <https://sbercloud.ru/ru/datahub/rugpt3family/rugpt-3-large>.
- N. V. Salomatina, I. S. Kononenko, E. A. Sidorova, and I. S. Pimenov. 2021. Identification of connected arguments based on reasoning schemes “from expert opinion”. *Journal of Physics: Conference Series*, 1715.
- SBERT. 2020. pretrained language model. <https://developers.sber.ru/portal/services/sbert>.
- Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 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*, P 380–396.
- Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics*, 46(2):499–510.
- Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. // *Proceedings of the 5th Workshop on Argument Mining*, P 155–163.
- Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. // *Proceedings of the International Conference on Computational Linguistics*, P 1501–1510.
- Christian Stab, Tristan Miller, Pranav Schiller, Benjamin Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. // *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 3664–3674.
- Manfred Stede and Jodi Schneider. 2018. Argumentation Mining. *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, and et al. 2021. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation Synthesis following Rhetorical Strategies. // *Proceedings of the 27th International Conference on Computational Linguistics*, P 3753–3765.

Douglas N. Walton and Thomas F. Gordon. 2012. The Carneades Model of Argument Invention. *Pragmatics & Cognition.*, 20(1):1–31.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against Neural Fake News. // *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, P 9054–9065.

Automatic Summarization of Russian Texts: Comparison of Extractive and Abstractive Methods

Goloviznina V. S.
Vyatka State University,
Kirov, Russia
goloviznina@vyatka-gsu.ru

Kotelnikov E. V.
Vyatka State University,
Kirov, Russia
kotelnikov.ev@gmail.com

Abstract

This paper investigates the problem of creating summaries of Russian-language texts based on extractive (TextRank and LexRank) and abstractive (mBART, ruGPT3Small, ruGPT3Large, ruT5-base and ruT5-large) methods. For our experiments, we used the Russian-language corpus of news articles Gazeta and the Russian-language parts of the MLSUM and XL-Sum corpora. We computed ROUGE-N, ROUGE-L, BLEU, METEOR and BERTScore metrics to evaluate the quality of summarization. According to the experimental results, the methods are ranked (from best to worst) as follows: ruT5-large, mBART, ruT5-base, LexRank, ruGPT3Large, TextRank, ruGPT3Small. The study also highlights the salient features of summaries obtained by various methods. In particular, mBART summaries are less abstractive than ruGPT3Large and ruT5-large, and ruGPT3Large summaries are often incomplete and contain errors.

Keywords: text summarization; extractive methods; abstractive methods; language models; TextRank; LexRank; mBART; ruGPT3; ruT5

DOI: 10.28995/2075-7182-2022-21-223-235

Автоматическое реферирование русскоязычных текстов: сравнение экстрактивных и абстрактных методов

Головизнина В. С.
Вятский государственный
университет,
Киров, Россия
goloviznina@vyatka-gsu.ru

Котельников Е. В.
Вятский государственный
университет,
Киров, Россия
kotelnikov.ev@gmail.com

Аннотация

В работе исследуется задача создания рефератов русскоязычных текстов на основе экстрактивных (TextRank и LexRank) и абстрактных (mBART, ruGPT3Small, ruGPT3Large, ruT5-base и ruT5-large) методов. Для экспериментов использовались русскоязычный корпус новостных статей Gazeta и русскоязычные части корпусов MLSUM и XL-Sum. Для оценки качества реферирования применялись метрики ROUGE-N, ROUGE-L, BLEU, METEOR и BERTScore. По результатам экспериментов методы ранжируются (от лучших к худшим) следующим образом: ruT5-large, mBART, ruT5-base, LexRank, ruGPT3Large, TextRank, ruGPT3Small. Также выделены особенности рефератов, получаемых разными методами. В частности, рефераты mBART оказываются наименее абстрактными по сравнению с ruGPT3Large и ruT5-large, а рефераты ruGPT3Large часто являются незавершенными и содержат ошибки.

Ключевые слова: реферирование текстов; экстрактивные методы; абстрактные методы; языковые модели; TextRank; LexRank; mBART; ruGPT3; ruT5

1 Introduction

Automatic text summarization is the process of creating a summary of the text containing the most important information [5]. There are the following approaches for text summarization – extractive, abstractive and hybrid. With the extractive approach, the summary is formed from the most important sentences of the source text; with the abstractive approach, the content of the summaries is generated and differs from the sentences of the source text. The hybrid approach combines these two approaches. Automatic text summarization methods are used in search engines, to summarize blogs, scientific articles, emails, lawsuits, and medical texts, and to generate headlines for news articles [3].

At present, the choice of the automatic summarization method for the Russian language is not obvious for the following reasons. Firstly, most research is carried out for the English language [27], there are few works for the Russian language [8, 9, 18, 21]. Secondly, a significant part of the works uses only the extractive approach, while abstract methods allow for a shorter and human-like presentation that differs from the sentences of the original text [3]. Thirdly, a number of new Russian-language neural network models have recently appeared, such as ruGPT-3 and ruT5, which have not been sufficiently studied in the summarization task.

Thus, the task of conducting a comparative analysis of extractive and abstractive methods of summarization on the material of the Russian language, including modern language models, is relevant.

The contribution of this work is as follows:

- for the first time, there has been carried a simultaneous comparison of extractive (TextRank and LexRank) and abstractive (mBART, ruGPT-3 and ruT5) summarization methods using three corpora of news articles: Gazeta [8], MLSUM [27] and XL-Sum [9];
- the methods under investigation have been ranked based on the ROUGE-N, ROUGE-L, BLEU, METEOR and BERTScore quality metrics;
- the salient features of summaries obtained by different methods have been revealed.

The paper is structured as follows. The second section provides an overview of previous work on Russian texts summarization. The third section is devoted to text corpora, models and methods used for text summarization. In the fourth section the experimental results are presented and discussed. The fifth section provides conclusions and suggests directions for further research.

2 Previous work

Language models based on the Transformer architecture [28] have become a key technology for solving natural language processing problems, including automatic text summarization [15]. Such models as mBART [16], ruGPT3 [25], and mT5 [29] have been used for summarizing Russian-language texts.

Gusev [8] fine-tuned the multilingual mBART model for text summarization on the Russian-language Gazeta dataset. The model showed the best results among abstractive models in ROUGE and BLEU metrics. In addition to mBART, Gusev used Pointer-generator, CopyNet models and extractive methods TextRank, LexRank and LSA.

Nikolich et al. [18] used the ruGPT3Small model, fine-tuned on Gazeta corpus, for text summarization in Russian. ruGPT3Small outperformed mBART [8] only in BERTScore.

Hasan et al. [9] fine-tuned the mT5 model for summarization in 44 languages, including Russian, using the XL-Sum corpus. The results of mT5 are close to the current level of summarization in English [30]. ROUGE-2 scores for other languages are comparable to results in English.

Polyakova and Pogoreltsev [21] proposed a new method of extractive summarization that reduces the problem to selecting the most probable sequence of sentences. The method outperforms the SummaRuNNer and mBART models in ROUGE-1 and ROUGE-L on the Gazeta dataset.

In our paper, in contrast to [8], besides mBART, we fine-tuned ruGPT3Small, ruGPT3Large [25], ruT5-base and ruT5-large [26] models. In contrast to [18], we used not only ruGPT3Small, but also the ruGPT3Large model. In contrast to [9], instead of multilingual mT5 model, we applied the Russian-language ruT5-large model. In contrast to [21], abstractive models are studied. Besides, it is the first time that all these methods and models are simultaneously analyzed using the three corpora: Gazeta, MLSUM, and XL-Sum.

3 Materials and Methods

3.1 Text Corpora

Corpora for text summarization are sets of texts and summaries to them. Our study uses the Russian-language corpus of news articles Gazeta and the Russian-language parts of the MLSUM and XL-Sum corpora. The Gazeta corpus consists of 63,435 articles from the news source Gazeta.ru¹ [8]. The MLSUM corpus contains 1,259,096 articles in five languages (German, Spanish, French, Russian, Turkish), of which 27,063 articles are in Russian from “Moskovsky Komsomolets”² [27]. XL-Sum consists of BBC³ news articles in 45 languages and contains about 1,350,000 articles, of which 77,803 are in Russian [9]. Characteristics of the corpora are shown in Table 1.

Corpus (source)	Dataset	Size	Data	Length in tokens		
				min	max	mean
Gazeta (Gazeta.ru)	train	52,400 (82.6%)	text	28	1,500	766.5
			summary	15	85	48.8
	validation	5,265 (8.3%)	text	191	1,500	772.4
			summary	18	85	54.5
	test	5,770 (9.1%)	text	357	1,498	750.3
			summary	18	85	53.2
MLSUM ⁴ (“Moskovsky Komsomolets”)	train	25,556 (94.4%)	text	55	11,689	949.9
			summary	10	65	14.7
	validation	750 (2.8%)	text	118	5,842	1,156.7
			summary	10	30	13.4
	test	757 (2.8%)	text	69	26,794	1,214.4
			summary	10	35	13.4
XL-Sum ⁵ (BBC News)	train	62,243 (80%)	text	19	22,274	682.1
			summary	1	246	29.4
	validation	7,780 (10%)	text	62	1,583	556.9
			summary	8	60	27.9
	test	7,780 (10%)	text	54	1,745	555.8
			summary	8	60	27.9

Table 1: Corpora statistics.
The length in tokens is specified for the *razdel*⁶ tokenizer.

3.2 Extractive methods

For extractive summarization, we used TextRank method from the *summa* library⁷ [2] and LexRank from the *lexrank* library⁸.

¹ <https://www.gazeta.ru>.

² <https://www.mk.ru/news>.

³ <https://www.bbc.com>.

⁴ MLSUM corpus has a very large max length of the texts (26,794 tokens) but it contains only 54 texts (0.2%) with a length of more than 5,000 tokens.

⁵ The training part of XL-Sum has a very large max length of the texts (22,274 tokens) and very salient min and max lengths for train summaries (min=1 and max=246). But it contains only 96 texts (0.15%) with a length exceeding 5,000 tokens and only 5 summaries with a length of less than 3 tokens, and 23 summaries with a length of more than 100 tokens.

⁶ <https://natasha.github.io/razdel>.

⁷ <https://pypi.org/project/summa>.

⁸ <https://github.com/crabcamp/lexrank>.

TextRank [17] is a method used for keyword extraction and extractive summarization. In the method, the text is divided into sentences, between which the similarity is calculated, and the PageRank algorithm [19] is used to obtain sentence scores. The sentences with the highest scores are included in the summary. A measure of the sentence similarity is the number of common words in these sentences.

In the LexRank method [4], the similarity measure of sentences is the cosine similarity of the TF-IDF vectors of these sentences. The method uses the following idea: if a sentence is similar to other sentences, then it is the central sentence of this text, that is, it contains the necessary and sufficient information about the entire text.

3.3 Abstractive methods

For abstractive summarization, we applied mBART, ruGPT3Small, ruGPT3Large, ruT5-base and ruT5-large models.

The BART (Bidirectional and Auto-Regressive Transformer) model is based on the Transformer architecture and includes a bidirectional encoder (like BERT) and an autoregressive decoder (like GPT) [13]. Two model versions are available: BART_{BASE} and BART_{LARGE}. The multilingual version mBART was trained on the Common Crawl corpus⁹ for 25 languages. We used the multilingual mBART_{LARGE}, fine-tuned for text summarization on the Gazeta dataset [8].

Models of the GPT (Generative Pre-trained Transformer) family consist of a Transformer decoder with a different number of layers [22]. The family includes three main models: GPT [22], GPT-2 [23], and GPT-3 [1]. The ruGPT-3 model is a Russian-language model from Sber based on GPT-2 [25]. The model was trained on 80 billion tokens. There are five versions of different sizes: ruGPT3Small, ruGPT3Medium, ruGPT2Large, ruGPT3Large, and ruGPT3XL. In our experiments, we used the ruGPT3Small and ruGPT3Large model.

The T5 (Text-to-Text Transfer Transformer) model was trained on 24 tasks for the English language [24]. The multilingual version mT5 was trained for 101 languages, but on one task – text filling. The ruT5 model is a Russian-language T5 model from Sber, available in two versions: ruT5-base and ruT5-large [26]. The model was trained on the same corpus as ruGPT-3. We used both versions: ruT5-base and ruT5-large.

4 Experiments

4.1 Experimental Setup

The TextRank method was applied with *compression ratio* = 0.2 (default value). For the LexRank method, the length of the summary was limited to three sentences (*summary_size* = 3). The rest of the methods parameters assumed default values.

During the experiments with each corpus, mBART_{LARGE}, ruGPT3Small, ruGPT3Large, ruT5-base and ruT5-large models were fine-tuned on the training part of the given corpus. The validation part of the corpus was used to select the number of training epochs. For the mBART_{LARGE}, ruT5-base and ruT5-large models, the length of the input text was 1,024 tokens, the length of the output data (the length of the generated summary) was limited by the length of the reference summary. The desired size of the summary is often a requirement in real-world problems. Given the availability of reference summary in our experiments, it is logical to use their size as a limitation.

For the ruGPT3Small and ruGPT3Large models, the length of the output data was regulated in the same way, the length of the input data was 2,048 tokens. When fine-tuning, the input of the ruGPT3Small and ruGPT3Large models was given sequences of the form: “Text:*text*[SEP]Summary:*summary*”, where *text* is the input text, *summary* is the reference summary for this text. When testing, the model generated a summary for the following input: “Text:*text*[SEP]Summary:”.

We also tested Lead-3 – it is a strong baseline, where summary is the first three sentences of every text.

We used five automatic metrics: ROUGE-N [14], ROUGE-L [14], BLEU [20], METEOR [12], and BERTScore [30] to evaluate the results. To calculate the ROUGE-N and ROUGE-L metrics, we applied

⁹ <https://commoncrawl.org>.

the *rouge* library¹⁰, for BLEU and METEOR – the *NLTK* library¹¹, *Snowball Stemmer*¹², and the *wiki_ru_wordnet* semantic network¹³. BERTScore uses embeddings from BERT and matches words in generated summaries and reference summaries by cosine similarity. We calculated BERTScore using the *bert-score* library¹⁴ and the Russian-language RuBERT model [10].

4.2 Results and Discussion

Table 2 shows the results of experiments for the three corpora, as well as the average values.

According to the results of experiments (see Table 2), models and methods can be ranked as follows:

1. ruT5-large,
2. mBART,
3. ruT5-base,
4. LexRank,
5. ruGPT3Large,
6. TextRank,
7. ruGPT3Small.

The ruT5-large and mBART models showed the best results, but mBART tends to repeat parts of the source text sentences. Figure 1 shows the average proportion of novel n-grams for three corpora in the summaries of abstractive models. A novel n-gram is an n-gram of the summary that is not contained in the source text. The proportion of novel n-gram is the number of novel n-grams divided by all n-grams of the summary. mBART summaries have the smallest proportion of novel n-grams. The proportion of novel n-gram summaries of ruT5-base and ruT5-large is greater than the proportion of n-grams of mBART summaries, but never more than the proportion of novel n-grams of reference summaries, which is surpassed by ruGPT3Small and ruGPT3Large. The ruGPT3Small and ruGPT3Large summaries contain the largest proportion of novel n-grams, but there are often errors – mismatches between the summary and the source text. Despite the large proportion of novel n-grams, ruT5-large summaries have significantly fewer errors than ruGPT3Large summaries.

Corpus	Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERTScore
Gazeta	Lead-3	31.02	13.44	27.69	10.80	34.44	56.49
	TextRank	21.44	6.27	18.56	3.92	26.31	49.90
	LexRank	23.93	8.00	20.96	5.64	28.17	51.49
	mBART	31.55	13.54	28.22	11.19	34.09	56.56
	ruT5-base	30.45	12.63	27.41	9.54	28.69	56.35
	ruT5-large	32.45	13.97	29.24	10.88	31.21	57.73
	ruGPT3Small	18.84	4.06	16.68	3.13	18.70	44.06
	ruGPT3Large	23.45	6.45	20.73	4.93	23.77	47.76
MLSUM	Lead-3	9.42	1.55	8.47	0.86	12.98	32.15
	TextRank	4.76	0.55	4.39	0.13	7.51	29.22
	LexRank	10.22	1.42	7.36	0.90	11.28	31.83
	mBART	11.48	1.95	10.26	1.49	10.52	37.89
	ruT5-base	12.35	1.86	11.22	1.58	9.68	38.67
	ruT5-large	14.06	2.86	12.69	2.81	11.84	39.92
	ruGPT3Small	9.14	0.60	8.13	0.40	6.66	34.27
	ruGPT3Large	9.36	0.99	8.17	0.73	7.44	35.00
XL-Sum	Lead-3	16.14	3.38	13.57	1.63	22.70	46.29
	TextRank	14.04	3.14	11.81	1.05	21.45	45.80
	LexRank	16.22	3.16	12.69	2.14	17.20	43.83
	mBART	26.47	10.95	22.67	7.51	27.16	54.24
	ruT5-base	26.52	10.67	22.79	6.58	25.35	52.89
	ruT5-large	28.42	11.98	24.41	7.93	28.31	56.06

¹⁰ <https://pypi.org/project/rouge>.

¹¹ <https://www.nltk.org>.

¹² <https://snowballstem.org>.

¹³ <https://wiki-ru-wordnet.readthedocs.io/en/latest>.

¹⁴ https://github.com/Tiiiger/bert_score.

Corpus	Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERTScore
XL-Sum	ruGPT3Small	16.19	3.28	13.68	2.25	15.94	40.12
	ruGPT3Large	19.37	5.17	16.48	3.74	19.63	42.74
Average	Lead-3	18.86	6.12	16.58	4.43	23.37	44.98
	TextRank	13.41	3.32	11.59	1.70	18.42	41.64
	LexRank	16.79	4.19	13.67	2.89	18.88	42.38
	mBART	23.17	8.81	20.38	6.73	23.92	49.56
	ruT5-base	23.11	8.39	20.47	5.90	21.24	49.30
	ruT5-large	24.98	9.60	22.11	7.21	23.79	51.24
	ruGPT3Small	14.72	2.65	12.83	1.93	13.77	39.48
ruGPT3Large	17.39	4.20	15.13	3.13	16.95	41.83	

Table 2: Automatic summarization scores on Gazeta, MLSUM and XL-Sum corpora

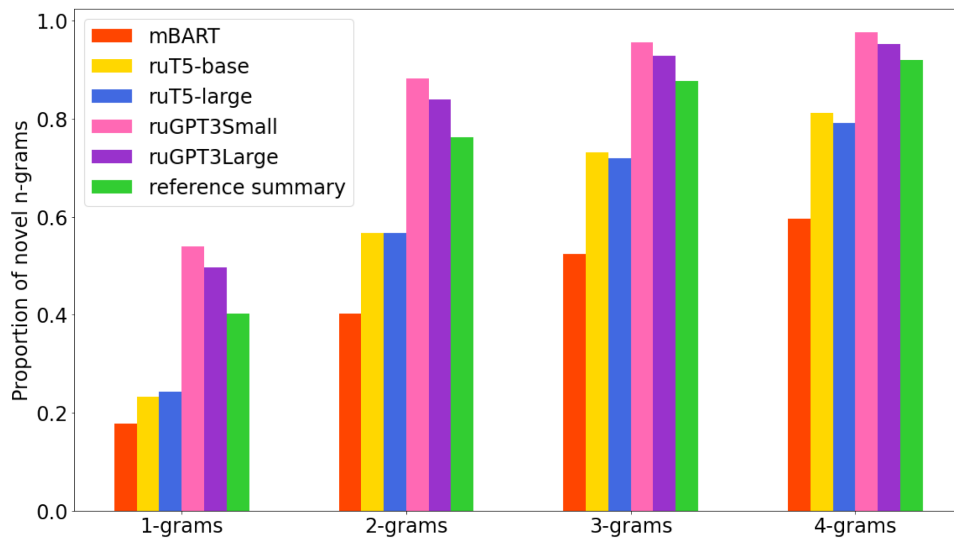


Figure 1: Average proportion of novel n-grams in summaries of abstractive models

The conclusions on the proportion of novel n-grams in summaries are confirmed by the extraction score [6] shown in Table 3. Extraction score is the sum of the normalized lengths of all long non-overlapping common sequences between a text and a summary, which is in the range from 0 to 1. This metric is inversely proportional to the degree of abstractiveness of the constructed summary. The lowest level of abstractiveness is shown by the mBART model, the highest – by ruGPT3Small.

Corpus	Characteristic	Reference summary	mBART	ruGPT3 Small	ruGPT3 Large	ruT5-base	ruT5-large
Gazeta	Average length of summaries in tokens	53.2	59.8	54.0	54.8	42.3	44.5
	Extraction score	0.06	0.39	0.03	0.05	0.24	0.26
MLSUM	Average length of summaries in tokens	13.4	18.8	13.2	13.5	12.4	15.0
	Extraction score	0.09	0.39	0.05	0.06	0.16	0.17
XL-Sum	Average length of summaries in tokens	27.9	22.2	29.0	28.9	19.8	21.5
	Extraction score	0.04	0.10	0.03	0.04	0.08	0.07
Average	Average length of summaries in tokens	31.5	33.6	32.1	32.4	24.8	27.0
	Extraction score	0.06	0.29	0.04	0.05	0.16	0.17

Table 3: Average summaries lengths (in tokens) and extraction score. The length in tokens is specified for the *razdel* tokenizer. Smaller values of extraction score correspond to a greater degree of abstractiveness of summaries

Another problem with abstractive methods is the incompleteness of the generated summaries. The ruGPT3Small and ruGPT3Large generate summaries that are closest in length to the reference ones (Table 3), but often does not complete them, while ruT5-base and ruT5-large, as a rule, complete sentences. Table 4 shows the proportion of summaries that do not end in end-of-sentence punctuation marks: “.”, “!”, “?”. For MLSUM, this value was not calculated, since the reference summaries from which the models were trained do not have punctuation marks at the end of the last sentence.

Corpus	mBART	ruGPT3Small	ruGPT3Large	ruT5-base	ruT5-large
Gazeta	0.10	0.86	0.96	0.09	0.14
XL-Sum	0.42	0.90	0.95	0.19	0.02

Table 4: The proportion of summaries that do not end in one of the punctuation marks “.”, “!”, “?”

With regard to extractive methods, LexRank performed better than TextRank (see Table 2). Figure 2 for extractive methods shows the proportion of extracted sentences according to their position in the source text. TextRank selects sentences from the text more evenly, LexRank tends to select sentences from the beginning of the text. Both methods include the first sentence of the text in summaries more often than others. This is due to the structure of the news article – the main information is contained at the beginning of the text, and then the clarifying facts are indicated.

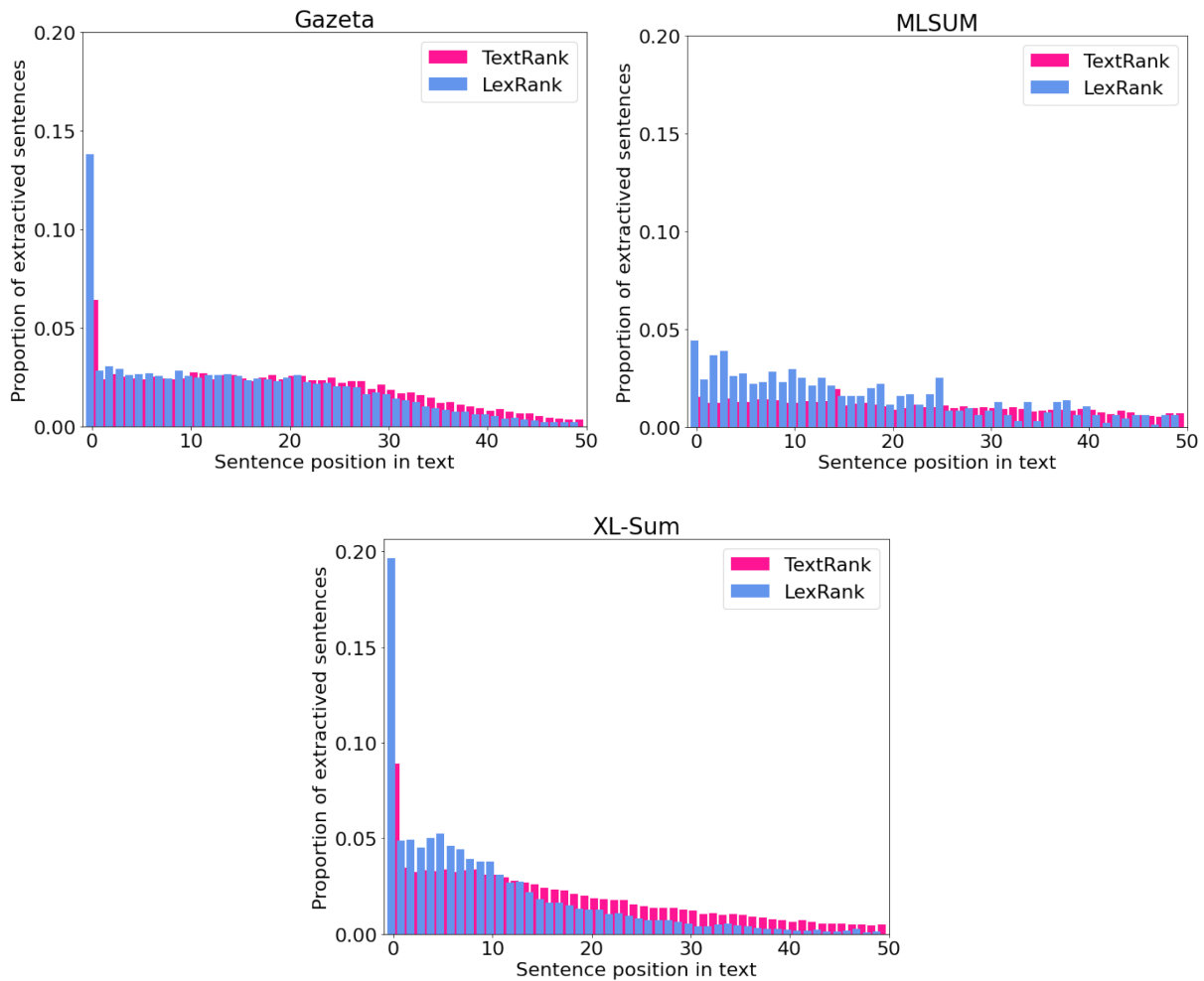


Figure 2: The dependence of the proportion of extracted sentences according to their position in the source text

The comparison between LexRank and ruGPT3Large is ambiguous (see Table 2). For the Gazeta corpus, the LexRank method outperforms ruGPT3Large in all metrics, for the MLSUM corpus it outperforms in 4 out of 6 metrics, for the XL-SUM corpus it is inferior in all metrics, except for BERTScore. On average, LexRank is ahead of ruGPT3Large in terms of METEOR and BERTScore, and for the ROUGE-2 metric the results differ by 0.01. However, we have decided to rank LexRank higher than ruGPT3Large due to the large number of factual errors of the latter, which cannot be in the extractive method. For the same reason, we put TextRank higher than ruGPT3Small.

Figures 3 (Russian version) and 4 (English version) show examples of summaries created by all eight methods. In the reference summary for the text from the Gazeta, two ideas stand out – a description of the new method and its criticism. It is only in the summary obtained by LexRank that there is an attempt to retain both ideas. The remaining methods pay attention to the first idea, while most of the mBART summary repeats the source text, the ruGPT3Large summary is not completed and contains errors.

From Table 2 it can be seen that for MLSUM the metrics are low compared to the other two corpora. This can be explained by the fact that MLSUM is different from the other two datasets: a) MLSUM is 2.3 times smaller than Gazeta and 2.9 times smaller than XLSUM; b) the average length of a summary in MLSUM is much shorter, it is 13.8 tokens, while the average length in Gazeta is 52.2 tokens, in XLSUM it is 28.4 tokens.

We tried to evaluate the adequacy of the automatic metrics. We used four criteria proposed by Fabbri et al. [5]: coherence, consistency, fluency and relevance. We also added included two additional criteria: no excess information (absence of redundancy in the summary) and abstractiveness (absence of sentences from the source text). We randomly selected five texts for each of the corpora and evaluated summaries generated using all eight methods, that is, a total of $5 \cdot 3 \cdot 8 = 120$ examples. During the evaluation, text, title, reference summary and generated summary were available to annotators. The annotator for each generated summary gave a score on a scale from 1 to 5 for each of the 6 criteria. We computed Kendall's tau rank correlations between average manual scores and all the automatic metrics. The correlation values turned out quite high – from 0.7143 (ROUGE-3) to 0.9286 (BERTScore). This confirms the adequacy of using automatic evaluation metrics.

Text / summary	Gazeta_594 ¹⁵
text ¹⁶	Для поимки преступника хватит и пары волосков, заявляют специалисты из Национального института стандартов и технологий в США. Разработанный ими метод идентификации способен выделить специфические для конкретного человека белки из пряди волос длиной всего сантиметр. Для других аналогичных подходов требуется в восемь раз большая длина. О новом способе исследователи рассказали в статье в журнале Journal of Forensic Sciences. ... Поскольку последовательности аминокислот в белках различаются от человека к человеку, такой подход позволит с высокой степенью точности идентифицировать людей и без выделения ДНК...
reference summary	Установить личность преступника можно всего по паре волосков, утверждают американские ученые — это можно сделать даже не выделяя ДНК, а лишь по ряду белковых последовательностей. Однако метод занимает слишком много времени, поэтому неизвестно, найдет ли он применение в криминалистике.
Lead-3	Для поимки преступника хватит и пары волосков, заявляют специалисты из Национального института стандартов и технологий в США. Разработанный ими метод идентификации способен выделить специфические для конкретного человека белки из пряди волос длиной всего сантиметр. Для других аналогичных подходов требуется в восемь раз большая длина.
TextRank	Однако большинство методов исследования белков в волосах подразумевают многократное измельчение и нагрев волос, в результате чего большая часть белков разрушается, а оставшихся может не хватить для получения необходимых данных. Точнее такую возможность просчитать пока сложно — в этом исследовании использовались образцы волос из биобанка и авторы не ставили перед собой цель найти их обладателя. В связи с этим возникает ряд вопросов — как краска для волос и другие косметические продукты влияют на эти белки?
LexRank	Однако большинство методов исследования белков в волосах подразумевают многократное измельчение и нагрев волос, в результате чего большая часть белков разрушается, а оставшихся может не хватить для получения необходимых данных. Разработанный ими метод идентификации способен выделить специфические для конкретного человека белки из пряди волос длиной всего сантиметр. Метод идентификации личности по полученным из волос GVP ранее не применялся в судебной практике, подчеркивает судмедэксперт Монти Миллер.
mBART	В США разработали новый способ идентификации личности по волосам — он позволяет выделить специфические для конкретного человека белки из пряди волос длиной всего сантиметр. Такой подход позволит с высокой степенью точности идентифицировать людей и без выделения ДНК.
ruGPT3Small	Исследование позволяет выявить особенности внешности преступника с помощью двух методов, направленных на выявление возможных причин, вызвавших задержку ДНК в ДНК. Миллионы людей в мире смогут идентифицировать себя по ДНК и другим типам волос
ruGPT3Large	Проверить человека по волосатой персоне в настоящее время можно достаточно просто. Теперь ученые разработали новый метод, позволяющий выявить человека по его волосяным фолликулам. Ранее для установления личности по образцам волос нужно было идентифицировать ДНК из кожи, с которой был
ruT5-base	Ученые выяснили, как определить личность человека по волосам. Для этого им достаточно выделить несколько волосков, чтобы поимка преступника была более точной.
ruT5-large	Американские ученые разработали новый способ идентификации личности по волосам. Для этого достаточно выделить белки из пряди волос длиной всего сантиметр.

Figure 3: Example of summaries received for the text from Gazeta (Russian version). Yellow – repeated text. Blue – the second idea of the reference summary, which is not reflected in any summary of abstractive models. Red – the summary is not completed. Green – the most important and correct information in the summary. Red – errors. Orange – defect (pronouns in summaries of extractive methods)

¹⁵ https://www.gazeta.ru/science/2019/11/24_a_12829868.shtml.

¹⁶ The text is not given in full: its main sentences are included in the example.

Text / summary	Gazeta_594 ¹¹⁷
text ¹⁸	A couple of hairs is enough to catch a criminal, say experts from the National Institute of Standards and Technology in the United States. The identification method they developed is capable of isolating human-specific proteins from a strand of hair just a centimeter long. Other similar approaches require eight times the length. The researchers described the new method in an article in the Journal of Forensic Sciences. ... Since the sequences of amino acids in proteins differ from person to person, this approach will allow people to be identified with a high degree of accuracy without DNA isolation...
reference summary	It is possible to establish the identity of a criminal with just a couple of hairs, according to American scientists — this can be done even without isolating DNA, but only by a number of protein sequences. However, the method takes too much time, so it is not known whether it will find application in forensic science.
Lead-3	A couple of hairs is enough to catch a criminal, say experts from the National Institute of Standards and Technology in the United States. The identification method they developed is capable of isolating human-specific proteins from a strand of hair just a centimeter long. Other similar approaches require eight times the length.
TextRank	However, most methods for studying hair proteins involve repeated grinding and heating of the hair, as a result of which most of the proteins are destroyed, and the remaining ones may not be enough to obtain the necessary data. It is still difficult to calculate this possibility more precisely - in this study, hair samples from the biobank were used and the authors did not set themselves the goal of finding their owner. In this regard, a number of questions arise - how do hair dye and other cosmetic products affect these proteins?
LexRank	However, most methods for studying hair proteins involve repeated grinding and heating of the hair, as a result of which most of the proteins are destroyed, and the remaining ones may not be enough to obtain the necessary data. The identification method they developed is capable of isolating human-specific proteins from a strand of hair just a centimeter long. The method of identifying a person using GVP obtained from hair has not previously been used in judicial practice, emphasizes forensic scientist Monty Miller.
mBART	In the United States, a new method of identifying a person by hair has been developed — it allows to isolate human-specific proteins from a strand of hair just a centimeter long. This approach will allow people to be identified with a high degree of accuracy without DNA isolation.
ruGPT3Small	The study allows you to identify the features of the criminal's appearance using two methods aimed at identifying possible causes that caused the DNA to be retained in DNA. Millions of people in the world will be able to identify themselves by DNA and other types of hair
ruGPT3Large	Checking a person by a hairy person is currently quite simple. Now scientists have developed a new method to identify a person by their hair follicles. Previously, in order to establish an identity from hair samples, it was necessary to identify DNA from the skin with which
ruT5-base	Scientists have figured out how to determine the personality of a person by hair. To do this, it is enough for them to select a few hairs so that the capture of the criminal is more accurate.
ruT5-large	American scientists have developed a new method of identifying a person by hair. To do this, it is enough to isolate proteins from a strand of hair just a centimeter long.

Figure 4: Example of summaries received for the text from Gazeta (English version). Yellow – repeated text. Blue – the second idea of the reference summary, which is not reflected in any summary of abstractive models. Red – the summary is not completed. Green – the most important and correct information in the summary. Red – errors. Orange – defect (pronouns in summaries of extractive methods)

¹⁷ https://www.gazeta.ru/science/2019/11/24_a_12829868.shtml.

¹⁸ The text is not given in full: its main sentences are included in the example.

4.3 Comparison with other works

Comparison of our results with the results of [8, 9, 18] is difficult. Hasan et al. [9] give the values of their own modified ROUGE metric, which considers the language – multilingual rouge scoring¹⁹, while we calculate the standard ROUGE metric [14]. Also, as a metric that takes into account the language, we use METEOR with Russian-language *Snowball Stemmer* and the *wiki_ru_wordnet* semantic network. Gusev [8] uses a different METEOR library. In addition, in [8] the input length was limited to 600 tokens, in our work – to 1024. Nikolich et al. [18] calculate BERTScore using the multilingual BERT model [7], we use RuBERT [10]. The values of the parameters for language models are different in our work and, for example, in [18].

In this regard, we show our results along with the results of [8, 18, 21] only on ROUGE and BLEU metrics (Table 5). The Gazeta is the only corpus, which these works investigate. To emphasize the difficulty of direct comparison, we did not highlight the best results in Table 5.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
ruT5-large (our work, Table 2)	32.45	13.97	29.24	10.88
mBART [8]	32.11	14.2	27.9	12.4 ²⁰
ruGPT3Large (our work, Table 2)	23.45	6.45	20.73	4.93
ruGPT3Small (our work, Table 2)	18.84	4.06	16.68	3.13
ruGPT3Small [18]	11.4	1.4	10.0	23.1 ²¹
[21]	35.6	14.2	32.4	–

Table 5: The comparison of our results with other works for the Gazeta corpus

5 Conclusion

In the study, we compared several models and methods within the framework of abstractive and extractive approaches on the corpora of news articles Gazeta, MLSUM and XL-Sum.

Based on the experimental results, we ranked the methods (from best to worst) as follows: ruT5-large, mBART, ruT5-base, LexRank, ruGPT3Large, TextRank, ruGPT3Small.

During the analysis of summaries obtained by different methods, we identified several features:

- mBART has the lowest level of abstractiveness compared to ruGPT3Large and ruT5-large;
- ruGPT3Small and ruGPT3Large generate summaries that are closest in length to the reference ones, but often does not complete them and makes errors;
- ruT5-base and ruT5-large summaries are also close to the reference length, rather abstract and contain fewer errors than summaries of ruGPT3Small and ruGPT3Large;
- TextRank more evenly selects sentences from the text;
- LexRank tends to select sentences from the beginning of the text.

In further research, we intend to compare the considered methods on the Russian-language part of the WikiLingva corpus [11], formed on WikiHow articles, which differ in their structure from news articles.

Acknowledgements

This work was supported by Russian Science Foundation, project № 22-21-00885, <https://rscf.ru/en/project/22-21-00885>.

¹⁹ https://github.com/csebuettl/xl-sum/tree/master/multilingual_rouge_scoring.

²⁰ Gusev [8] made a mistake in calculating BLEU. Updated scores: <https://arxiv.org/pdf/2006.11063.pdf>

²¹ In [18], the BLEU value is compared with the erroneous results of [8], it is probably also incorrect.

References

- [1] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. et al. (2020), Language Models are Few-Shot Learners, Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), pp. 1877–1901.
- [2] Barrios, F., López, F., Argerich, L., Wachenchauser, R. (2016), Variations of the Similarity Function of TextRank for Automated Summarization, arXiv preprint arXiv:1602.03606.
- [3] El-Kassas, W. S., Salama, C. R., Rafea, A. A., Mohamed, H. K. (2021), Automatic text summarization: A comprehensive survey, Expert Systems with Applications, Vol. 165.
- [4] Erkan, G., Radev D. R. (2004), Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence, Vol. 22, pp. 457–479.
- [5] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., Radev, D. (2021), SummEval: Re-evaluating Summarization Evaluation, Transactions of the Association for Computational Linguistics, Vol. 9, pp. 391–409.
- [6] Cibils, A., Musat, C., Hossmann, A., Baeriswyl, M. (2018), Diverse beam search for increased novelty in abstractive summarization, arXiv preprint arXiv:1802.01457.
- [7] Devlin, J., Chang, M-W., Lee, K., Toutanova, K. (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, Vol. 1, pp. 4171–4186.
- [8] Gusev, I. (2020), Dataset for Automatic Summarization of Russian, Proceedings of 9th Conference Artificial Intelligence and Natural Language (AINL 2020), pp. 122–134.
- [9] Hasan T., Bhattacharjee, A., Islam, M., Samin, K., Li, Y., Kang, Y., Rahman, M. S., Shahriyar R. (2021), XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages, Findings of the Association for Computational Linguistics: ACL-IJCNLP, pp. 4693–4703.
- [10] Kuratov Y., Arkipov M. (2019), Adaptation of deep bidirectional multilingual transformers for Russian language, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2019», No. 18 (24), pp. 333–340.
- [11] Ladhak, F., Durmus, E., Cardie, C., McKeown, K. (2020), WikiLingua: A New Benchmark Dataset for Multilingual Abstractive Summarization, Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4034–4048.
- [12] Lavie, A., Agarwal, A. (2007), Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231.
- [13] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer L. (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, pp. 7871–7880.
- [14] Lin, C. (2004), ROUGE: A Package for Automatic Evaluation of Summaries, Association for Computational Linguistics, Vol. Text Summarization Branches Out, pp. 74–81.
- [15] Liu, U., Lapata, M. (2019), Text Summarization with Pretrained Encoders, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3730–3740.
- [16] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L. (2020), Multilingual Denoising Pre-training for Neural Machine Translation, Transactions of the Association for Computational Linguistics, Vol. 8, pp. 726–742.
- [17] Mihalcea, R., Tarau, P. (2004), TextRank: Bringing Order into Text, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411.
- [18] Nikolich, A., Osljakova, I., Kudinova, T., Kappusheva, I., Puchkova A. (2021), Fine-tuning GPT-3 for Russian Text Summarization, Proceedings of the 5th Computational Methods in Systems and Software, Vol. 2, pp. 748–757.
- [19] Page, P., Brin, S., Motwani, R., Winograd T. (1998), The PageRank Citation Ranking: Bringing Order to the Web.
- [20] Papineni, K., Roukos, S., Ward, T., Zhu W. (2002), BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318.
- [21] Polyakova, I., Pogoreltsev, S. (2021), Extractive Russian Text Summarization as Greedy Sentence Sequence Continuation Search with Probabilities from Pretrained Language Models, Proceedings of the Open Semantic Technologies for Intelligent Systems (OSTIS-2021), pp. 303–308.
- [22] Radford, A., Narasimhan, K., Saliman, T., Sutskever I. (2018), Improving Language Understanding by Generative Pre-Training. OpenAI Technical report.

- [23] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019), Language Models are Unsupervised Multitask Learners, OpenAI Technical report.
- [24] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu P. (2020), Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research*, Vol. 21, pp. 1–67.
- [25] Sber, GitHub – sberbank-ai/ru-gpts: Russian GPT3 models. Retrieved May 10, 2022, from <https://github.com/sberbank-ai/ru-gpts>.
- [26] Sber, GitHub – sberbank-ai/model-zoo: NLP model zoo for Russian. Retrieved May 10, 2022, from <https://github.com/sberbank-ai/model-zoo>.
- [27] Scialom, T., Dray, P. A., Lamprier, S., Piwowarski, B., Staiano, J. (2020), MLSUM: The Multilingual Summarization Corpus, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8051–8067.
- [28] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N. et al. (2017), Attention is all you need, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008.
- [29] Xue, L., Constant, N., Roberts, F., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel C. (2021), mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498.
- [30] Zhang, T., Kishore, V., Wu, F., Weinberger K., Q., Artzi, Y. (2020), BERTScore: Evaluating Text Generation with BERT, *Proceedings of the International Conference on Learning Representations (ICLR 2020)*.

Russian noun *prichina* ‘cause, reason’ revisited: constructions with *что* ‘that’-clause

Goncharov Alexander A.

Federal Research Center
“Computer Science and Control” of
the Russian Academy of Sciences
(FRC CSC RAS), Moscow, Russia
a.gonch48@gmail.com

Kobozeva Irina M.

Lomonosov Moscow State University,
Moscow, Russia;
Institute of Linguistics, Russian
Academy of Sciences, Moscow, Russia
kobozeva@list.ru

Abstract

The paper deals with Russian syntactic constructions *prichina, что P* (lit. ‘cause that P’) and *prichina того, что P* with expletive pronoun *то* in genitive case, that are in most cases interchangeable. The purpose is to describe the syntactic and semantic properties of these constructions in Russian using corpus data. The semantics of the noun *prichina* ‘cause, reason’ and its argument structure were critically examined, and a representative set of examples where this noun is a part of one of these constructions was analyzed. The following results were obtained: 1) the description of the argument structure of the word *prichina* was supplemented; 2) the constraint on the use of *что*-complements with deverbative nouns observed in [10] was shown to cover non-deverbative noun-predicate *prichina* ‘cause’; 3) it was shown that the semantic role of the *что*-clause may be either Cause or Effect depending on the syntactic function of the construction; 4) it was determined that both meanings of the noun *prichina* (objective vs subjective cause) may be expressed in both constructions — with and without *то* — irrespective of syntactic function of the construction.

Keywords: cause; that-clause; semantics; syntax

DOI: 10.28995/2075-7182-2022-21-236-246

Еще раз о существительном *причина*: конструкции с сентенциальным актантом, вводимым союзом *что*

Гончаров А. А.

Федеральный исследовательский
центр «Информатика и управление»
Российской академии наук
(ФИЦ ИУ РАН), Москва, Россия
a.gonch48@gmail.com

Кобозева И. М.

МГУ имени М. В. Ломоносова,
Москва, Россия;
Институт языкознания Российской
академии наук, Москва, Россия
kobozeva@list.ru

Аннотация

Статья посвящена рассмотрению синтаксических конструкций *причина, что P* и *причина того, что P* на корпусном материале. Цель статьи — описать синтаксические и семантические свойства этих конструкций в современном русском языке. Для достижения данной цели были критически рассмотрены имеющиеся описания семантики существительного *причина* и его актантной структуры, после чего была проанализирована представительная выборка примеров, где это существительное употреблено в составе указанных конструкций. В ходе исследования были получены следующие результаты: 1) дополнено описание модели управления существительного *причина*; 2) показано, что ограничение на *что*-придаточное без *то* при отглагольных именах, выявленное в [10], действует и в случае непроизводного (в современном русском языке) имени *причина*; 3) установлено, какую валентность — Причины или Следствия — заполняет *что*-придаточное при существительном *причина* в зависимости от синтаксической функции конструкций; 4) показано, что в обеих конструкциях могут реализоваться оба возможные значения существительного *причина* — объективной и субъективной причины.

Ключевые слова: причина; *что*-придаточное; семантика; синтаксис

1 Введение

Известно, что одной из актуальных проблем в области автоматической обработки естественного языка является проблема внутритекстовых отношений разных типов (см., напр., [19]). Особое внимание нередко уделяется отношениям, которые устанавливаются на уровне «локальной» структуры текста (в терминологии из [9]), а также средствам их выражения. В то время как часть отношений рассматривается лишь в некоторых теориях, отношение причины (с разной степенью детальности) изучается в рамках всех основных подходов к связности текста [5]. В настоящей статье предлагается проанализировать некоторые случаи, где причинное отношение выражено с использованием существительного *причина*, в частности при помощи синтаксических конструкций *причина, что P* и *причина того, что P*¹.

Тема данного исследования актуальна также в контексте изучения ограничений, налагаемых на сентенциальные актанты существительных с предикатной семантикой: как на саму возможность таковых (см. [7], [14]), так и на способ их оформления — с эксплетивным *то* или без (см. [10], [11]). В указанных работах рассматривались только имена, производные от глаголов, предикативов или адъективов (*любовь, открытие, возможность, интерес, уверенность* и т.п.). Существительное *причина* на синхронном уровне непроизводно, являясь при этом двухвалентным предикатным именем в смысле [1: 27]. Анализируя дистрибуцию двух вариантов конструкции — *причина, что P* и *причина того, что P* — мы увидим, подчиняется ли выбор варианта с *то* или без *то* ограничению, сформулированному М. Ю. Князевым: сентенциальный актант предикатного имени может быть оформлен со *что* (без *то*), только если это имя входит в коллокационный контекст (об особых свойствах имен в составе коллокаций см. [15]).

Конструкции рассматриваются с использованием корпусных данных. Источником большинства примеров, приводимых в статье, является Основной корпус Национального корпуса русского языка [17], причем для поиска были использованы тексты, написанные не ранее 1900 г.

В разделе 2 анализируется описание семантики существительного *причина* и его модели управления, данное в исследованиях [3], [4], и отмечаются имеющиеся в нем лакуны; в разделе 3 описываются синтаксические функции конструкции *причина, что P* и семантика сентенциального актанта, выражаемого *что*-придаточным в зависимости от синтаксической функции конструкции; в разделе 4 аналогичное описание дается для конструкций *причина того (тому), что P*; в разделе 5 выясняется, в каком из двух своих значений — объективной или субъективной причины — выступает слово *причина* в конструкциях *причина (того), что P*. В заключении делаются обобщения, касающиеся рассмотренных конструкций.

2 Семантика и актантная структура существительного *причина*: state of the art

Наиболее детальное из известных нам описаний значения и актантной структуры существительного *причина* дано в [3] и [4], где у слова *причина* выделено два значения и, соответственно, две лексемы в терминологии авторов, работающих в рамках Московской семантической школы — *причина 1* и *причина 2*.

Лексеме *причина 1* дается следующее толкование: *X — причина 1 Y-a* = 'ситуация Y имеет место, потому что имеет или ранее имела место ситуация X; причем, если у ситуации X есть субъект, то Y не является его целью' [4: 72]. Для иллюстрации этого значения в [4: 75] приводится, в частности, такой пример:

- (1) *По той же причине — сильный жар и головная боль — я плохо помню, что я делал на Сухаревке.* [В. Каверин]

¹ Один из рецензентов считает, что это — «одна и та же конструкция, точнее, первая последовательность представляет собой результат эллипсиса местоимения *то* в конструкции *причина того, что P*». Мы не можем согласиться с этой точкой зрения, поскольку, во-первых, понятие эллипсиса не относится к феномену наличия / отсутствия соотносительного *то* в главной части сложного предложения с сентенциальным актантом (см. общепринятое определение эллипсиса в [21]), а во-вторых, употребление «первой последовательности» лицензируется особым типом морфосинтаксических контекстов, что позволяет говорить о различии двух синтаксических конструкций с вершинным именем *причина* и его сентенциальным актантом: с *то* и без *то*.

В толковании фигурирует семантический примитив ‘потому что’ отражающий понятие, близкое к «общенаучному понятию причины» [4: 71]². Из этого толкования следует, что *причина 1* имеет только две семантических валентности — «исходной и производной ситуаций» [3: 877, 879]. О выражении валентности исходной ситуации (далее ИС) ничего не сообщается, а о валентности производной (далее ПС) — что она «выражается существительным в родительном падеже» [14: 76]. Однако в целом ряде конструкций валентность ПС слова *причина* может выражаться *что*-придаточным, как показывает пример (2):

(2) *Может быть, это была причина, что царь меня так сухо встретил.* [А. Брусилов (1923)]

Валентность ИС в (2) выражена анафорическим местоимением *это*, отсылающим к ситуации, описанной в предшествующем тексте.

О семантике лексемы *причина 1* также сказано, что она «фактивна» в том смысле, что ПС реализуется обязательно, когда имеет место ИС [4: 75]³.

Причина 2 толкуется следующим образом: *Р является для X-а причиной для Q* = ‘Говорящий или человек X понимает, что существование ситуации Р может быть причиной 1 того, что субъект X делает или испытывает Q, или хочет делать Q’ [3: 877; 4: 76]. Для иллюстрации этого значения в [4: 75] приводится, в частности, такой пример:

(3) *У него была причина обидеться и уйти, но он решил этого не делать.*

Наличие в толковании *причины 2* модального оператора возможности показывает, что, в отличие от *причины 1*, ситуация Р не влечет с необходимостью ситуацию Q, а создает для говорящего или человека X возможность осуществления Q. Реализация же Q зависит от воли субъекта X ситуации Q, и в этом смысле *причина 2* субъективна, в отличие от объективной *причины 1*.

Из толкования *причины 2* следует, что эта лексема имеет 3 семантических валентности: ИС, ПС и субъекта ПС. О валентности субъекта в указанных работах сообщается, что при этом слове она не может быть выражена его непосредственным синтаксическим зависимым, но может выражаться подчиненными лексико-функциональному глаголу предложно-именными группами у + РОД. О валентности ИС в [4] ничего не сообщается, а в [3: 880] дана противоречивая информация: в зоне валентностей сказано, что валентность ИС при лексеме *причина 2* и ее синонимах не выражается, а в зоне сочетаемости приведены примеры (4) и (5), в которых «названия людей» выступают «в роли метонимического обозначения» ИС, то есть выражают валентность ИС:

(4) *Разве вы не допускаете, что я тоже могу служить причиной беспокойства.* [М. Горький (1928)]

(5) *И этот гвардеец <...> был причиной того, что она так изменилась.* [В. Каверин (1926)]

О валентности ПС в указанных трудах сообщается, при лексеме *причина 2* она может быть выражена: 1) существительным в РОД; 2) инфинитивом; 3) придаточным предложением, вводимым союзом *чтобы*, 4) предложно-именными группами *для* + РОД и *к* + ДАТ.

Обратим внимание на то, что во всех приводимых в указанных работах примерах именных или предложно-именных групп (далее — NP и PP), выражающих ПС, кроме (5), ПС представлена полной номинализацией (*его переезд в Англию, разговоры о смерти* и т.п.), то есть синтаксической группой, вершиной которой является отглагольное имя (*переезд, беспокойство* и т.п.), и ко всем этим случаям вполне применима характеристика: ПС выражается существительным в РОД. Но в (5) ПС представлена неполной номинализацией — сложной именной группой (ИГ), состоящей из эксплетивного местоимения *то* и *что*-придаточного (о многочисленных различиях между полной и неполной номинализациями, коррелирующими с семантическим противопоставлением «фактов» и «событий» см. [2: 101 и далее]).

В традиционной русистике неполные номинализации рассматриваются как изъяснительные придаточные в составе сложного предложения, в главной части которого при «опорном слове»

² Поясним: в «общенаучном» смысле *причина* — это явление X, предшествующее явлению Y и обуславливающее его осуществление: если бы не было X, то не было бы Y; ср. [20].

³ Данное требование не входит в общенаучное понимание причины и не соответствует употреблению слова *причина* в языке, на что наше внимание обратил анонимный рецензент, приведя пример (i): *Солнечное затмение явилось причиной массовых ДТП в Ханты-Мансийске*. Действительно, чтобы признать истинность (i) нам не требуется также признать, что всякий раз, когда в районе Ханты-Мансийска происходит солнечное затмение, там происходят массовые ДТП.

(глаголе, имени и др.) «может находиться указательное местоименное слово *то* (*того, тому ...*) или (только при существительном) *тот* (*та, то*) и *такой*» [18: 482]. Это указательное слово, называемое соотносительным местоимением, или коррелятом, считается факультативным элементом, но выделены случаи, при которых оно обязательно [18: 482]. В [12] все предикаты и предикативы, управляющие *что*-придаточным разделены на три субкатегории: не допускающие, требующие и допускающие *то* перед придаточной частью. Аналогичный подход можно применить и к существительным, управляющим изъяснительным *что*-придаточным. В какую категорию в таком случае попадает имя *причина* (в любом из его значений)? Существенно чаще оно встречается с коррелятом (*причина того, что P; та причина, что P*). Но отсутствие коррелята вполне допустимо, и такие примеры, причем без налета просторечности, встречаются в НКРЯ достаточно часто, см. примеры (2) и (6) из Основного корпуса, а также примеры из (7) и (8) из Газетного корпуса:

- (6) <...> *лишь небольшая любовь его сограждан к чтению могла быть **причиной, что** он доселе сохранял добрые отношения со всей губернией.* [«Сибирские огни» (2013)]
- (7) *Пока это единственная **причина, что** мы не можем забрать Мишеньку домой.* [«Комсомольская правда» (2013.09)]
- (8) *Все-таки в чем **причина, что** у региона так долго не было своего губернатора?* [«Известия» (2014.03)]

Независимо от того, какой синтаксический анализ принимается для неполных номинализаций — сложная именная группа или придаточное с коррелятом в главной части — описание актантной структуры существительного *причина* (и его синонимов), данное в указанных работах, должно быть дополнено способом выражения валентностей при помощи *что*-придаточного с эксплетивным *то* (*та*) или без него.

3 Конструкция *причина, что P*: синтактико-семантический анализ

Конструкция *причина, что P* нередко встречается в корпусе. Выборка, формируемая по запросу «*причина, на расстоянии 1 от что, ном*», содержит 1 047 вхождений из 835 документов.

Нами было проанализировано достаточно представительное подмножество, включающее 295 вхождений (получено путем выгрузки поисковых результатов при помощи функции «Скачать несколько первых результатов выдачи»). Из этого подмножества были исключены 33 случая, где *что* не вводит сентенциальный актант при существительном *причина*, и 1 случай, дублирующий уже имеющийся в выборке пример. Оставшиеся примеры (261) распадаются на 2 неравных группы:

- группа I, где имя *причина* входит в состав субъекта или предиката (глагольной группы) клаузы — 10 примеров;
- группа II, где имя *причина* подчинено предлогу *по* или другому причинному предлогу и вводит обстоятельственную клаузу — 251 пример.

Рассмотрим схемы предложений, представленные в группе I.

1. *Причина в составе предиката клаузы: X Copul / *Opeг₁ причина / причиной, что Y*, где Copul — глагол-связка *быть*, а *Opeг₁ — лексико-функциональный глагол типа Opeг₁ или представляющий сложную лексическую функцию ряда Opeг₁ (*являться, служить, стать* и др.); см. (2), (6) из нашей выборки, а также (7) из Газетного корпуса и (9), найденный при помощи сервиса Google Books [6].

- (9) *Я потом думала — может быть, вся эта история во время призыва и была правдой, и, может, она и была **причиной, что** по возвращении с фронта ему поставили ультиматум или стать провокатором, или идти в тюрьму.* [Н. Крупская (2017)]

2. *Причина — вершина субъектной NP:*

2.1. *Причина, что X, Copul такая / следующая / Adj: Y*, где Adj — любое прилагательное, семантически совместимое с именем *причина*.

- (10) *А **причина, что** я их носил, была простая: отец в молодости носил такие же.* [митрополит Питирим (Нечаев) (1993–2003)]

2.1(в) В вопросительном или косвенно-вопросительным варианте этой схемы валентность Y выражена вопросительным местоимением *какой*, как в (11), или фраземой *что (это) за*, как в (12).

(11) *Один вопрос: если ты идешь по Пути, то какие причины, что ты идешь по Пути?* [«Пятое измерение» (2003)]

(12) *а что это за ряд причин, что пришлось оформить на свекровь???* [Женщина + мужчина: Брак (форум) (2004)]

2.2. *Причина, что X , Copul / Func₁ в Y -е / в том, что Y* (конверсив схемы 2.1), где Func₁ — лексико-функциональный глагол (здесь — с сильноуправляемым предлогом *в*), связывающий предикатное имя в функции подлежащего с его первым актантом в функции первого дополнения (*быть / состоять / крыться в* и др.); см. пример (13), найденный при помощи сервиса Google Books [6].

(13) *Причина, что это не было сделано своевременно, состояла в том, что в городе было запрещено готовиться к эвакуации <...>*. [Е. Мосягин (2021)]

2.2(в). В нашей выборке данная схема представлена только в вопросительном и косвенно-вопросительном варианте, см. (14), где валентность P выражена вопросительным местоимением *что*, (15), где валентность P выражена местоимением *это* в сфере действия оператора общего *нели-вопроса*, а также (8) из Газетного корпуса.

(14) *Необходимо проанализировать, насколько успешны были судебные реформы XIX века в дореволюционной России, в чем причина, что тот потенциал, который был в них изначально заложен <...>, не был реализован в долгосрочной перспективе.* [«Арбитражный и гражданский процессы» (2004.06.28)]

(15) *Правда, в законопроекте никак не прописан порядок выявления таких злонамеренных граждан (не в этом ли причина, что «резина» все тянется и тянется?).* [«Огонек» (2013)]

Обратим внимание на то, что все пять видов синтаксических контекстов, в которых встречается конструкция *причина, что P* относятся к типу коллокационных, т.е. имя *причина* синтаксически связано или с лексико-функциональным глаголом матричного предложения, образуя с ним коллокацию, или с глаголом-связкой *быть*⁴.

Таким образом, корпусные данные подтверждают, что ограничение, сформулированное в [10] на употребление *что*-придаточных без эксплетивного *то* при именах, образованных от полнозначных предикатных (= валентных) слов (глаголов, предикативов, прилагательных) действуют и в случае непроемных предикатных лексем, значение которых сродни грамматическому, к каковому относится (на синхронном уровне) *причина*.

Легко видеть, что во всех приведенных схемах группы I *что*-придаточное в конструкции *причина, что P* кодирует ПС. Исключение — схема 3,

3. *В том-то и причина (X-а), что Y* , в которой *что*-придаточное выражает ИС, как показывает пример (16).

(16) *В том-то и причина, что спортивная и педагогическая деятельность — разные вещи.* [Новая тема, которую никто пока не трогает (форум) (2008)]

Схема 3 — экспрессивный трансформ нейтральной 2.2: *Причина, что X , Copul / Func₁ в Y -е / в том, что Y* . Местоимение *то* в составе схемы *в том-то и X, что Y* (*в том-то и беда / дело / причина, что Y*) катафорически отсылает к ситуации Y в придаточном. А поскольку NP, управляемая предлогом *в*, однозначно кодирует ИС, то и *что*-придаточное однозначно соотносится с ИС. В отличие от *то*, местоимение *это* в роли актанта X анафорически отсылает к ранее упомянутой за пределами данной клаузы ИС не только в нейтральной, но и в экспрессивной конструкции с выносом предложной группы в начало клаузы, а *что*-придаточное по общему правилу выражает ПС (следствие):

(17) *Они так давно были вместе, что могли считаться семейной парой. Но, может, в этом и причина, что жениться Рома не спешил.* [«Комсомольская правда» (2013.10)]

⁴ Об отсутствии четкой границы (схождении синтаксических свойств) между связками, полусвязочными и полувспомогательными (= лексико-функциональными) глаголами см. [21: 248].

В Группе II имя *причина* в *причина, что P* подчинено предлогу *по* или другому предлогу с причинным значением, который и образует контекст релевантный для синтаксических и семантических свойств конструкции:

- (18) *Ну, и получилось так, Георгий, что по причине, что тебя найти и опознать не смогли, ты тут остался, прижился, работаешь...* [«Знамя» (2006)]

При этом только в 5 случаях из 251 имя *причина* выступает без адъективного эксплетивного *то* в функции определителя, как в (18). Во всех остальных случаях определитель *то* присутствует:

- (19) *На «южном» же, т.е. верхневолжском маршруте <...>, подобных названий нет именно по той причине, что кривичская миграция на восток и юго-восток из смоленско-полоцких пределов осуществлялась позже <...>.* [«Вопросы языкознания» (2010. №2)]

Грамматическая трактовка конструкции *по той причине, что* заслуживает отдельного рассмотрения. В [18: 578] *по той причине, что* (как и *по причине того, что*) рассматривается как составной союз. Однако существительное *причина* в составе конструкции *по той причине, что P* еще не грамматикализовалось: оно сохранило свое лексическое значение, изменяется по числу, допускает при себе разнообразные определения, в том числе связанные сочинительной связью, а также вставки ограничительных и вопросительных частиц, вводных слов и оборотов. Наиболее частотным в нашей выборке оказалось определение *простой*: на конструкцию *по той простой причине, что* приходится 80 примеров из 246. Однако встречаются и другие — гораздо менее частотные, но крайне разнообразные случаи, см. Табл. 1.

	по		той		причине	, что	147
	по		тем		причинам	, что	1
не	по		той	ли	причине	, что	1
	по		той	еще	причине	, что	1
	по		той	лишь	причине	, что	2
	по	одной	той		причине	, что	1
	по	одной лишь	той		причине	, что	1
	по		той	единственной	причине	, что	1
	по		той	, вероятно,	причине	, что	1
	по		той	, как они мне объяснили,	причине	, что	1
	по		той	простой	причине	, что	80
	ввиду		той	простой	причины	, что	1
	в силу		той	простой	причины	, что	1
	по		той	простой, но важной	причине	, что	1
	по		той	простейшей	причине	, что	1
	по		той	деликатной	причине	, что	1
	по		той	наглядной	причине	, что	1
	по		той	парадоксальной	причине	, что	1
	по		той	печальной	причине	, что	1
	по		той	уважительной	причине	, что	1
							246

Табл. 1. Распределение вариантов сочетаний конструкции *причина, что P* с определителем *то* и управляющим причинным предлогом.

Как видно из таблицы, возможным оказывается даже употребление предлогов, отличных от *по*: ср. (20) с *ввиду* и (21) с *в силу*.

- (20) *«Не бывает идеальных людей» (без слабостей и недостатков) ввиду той простой причины, что все мы вынуждены жить среди подобных.* [«Coach» (2008)]

- (21) *Суверенных прав на ведение внутренней политики тогда не касались в силу той простой причины, что ограничивать т. Сталина по части подавления гражданских прав было нереально, да и у прочих держав <...> рыльце тоже было отчасти в пушку.* [«Эксперт» (2004.12.06)]

Наблюдаемые факты не позволяют отнести выражение *по той причине, что* к союзам и поставить их в один ряд с *потому(,) что* или *поскольку*, точно так же, как словосочетания типа *в этот день* не относят к наречиям и не ставят в один ряд с наречием *сегодня*. И с синтаксической, и с семантической точек зрения данное выражение — просто наиболее частотное в ряду других предложных групп, каждая из которых состоит из причинного предлога — *по, ввиду, вследствие, в силу* — и сложной ИГ (complex noun phrase) т. е. ИГ с вложенным (= придаточным) предложением. В вершине этой ИГ находится родовое имя, так называемый «классификатор» [16: 325], которое характеризует вложенную в ИГ пропозицию *P* или по ее логическому типу (напр., *факт в тот факт, что P*) или по ее роли в соответствующем логико-семантическом отношении, напр., *цель в с той целью, чтобы P* (*P* выступает в роли цели в отношении «действие — его цель»), *условие в при условии, что P* (*P* выступает в роли условия в отношении «условие — (гипотетическое) следствие») или *причина в по (той) причине, что P* (*P* выступает в роли причины в причинно-следственном отношении).

При таком подходе все *что*-придаточные из примеров выше оказываются придаточными разъяснительными (или дополнительными) при существительном *причина*, выражающими валентность содержания *причины*. Иными словами, они описывают конкретную ситуацию, выступающую в роли причины, или исходную (ИС) по принятой нами терминологии, по отношению к другой конкретной ситуации — производной (ПС), задаваемой главным предикатом матричного предложения. Классификатор *причина* в силу своего лексического значения «конкретизирует» семантику многозначного предлога *по*, участвуя в кодировании причинно-следственного отношения между двумя ситуациями — ситуацией, описанной в главном предложении, и ситуацией в *что*-придаточном. С точки зрения дискурс-анализа, для которого важно прежде всего установление логико-семантических отношений (ЛСО) между элементарными единицами дискурса, выражения из Табл. 1 относятся к функциональному классу коннекторов, выражающих ЛСО причины, и входят в кластер коннекторов, формируемых на базе имени *причина*. При этом *по причине, что P* представляет собой базовый (минимальный) вариант коннектора, *по той причине, что P* — основной (= самый частотный) вариант, а все прочие модификации данных выражений — их речевые реализации [8], [13].

Итак, *что*-придаточное в конструкции *причина, что P* может выражать как валентность ИС, так и валентность ПС имени *причина* в зависимости от синтаксической функции этого имени: ИС, если *причина* входит в группу причинного предлога или употреблено в эмфатической схеме 3, и ПС, если *причина* входит в группу субъекта или предиката предложения.

4 Конструкции *причина того, что P* и *причина тому, что P*

Конструкция *причина того, что P* также достаточно частотна. Выборка, формируемая по запросу «Слово 1: *причина*; Слово 2: *то* на расстоянии 1 от Слова 1; Слово 3: *что* на расстоянии 1 от Слова 2», содержит 613 вхождений из 518 документов.

Как и для *причина, что P*, нами было проанализировано 295 первых примеров, полученных путем выгрузки поисковых результатов при помощи функции «Скачать несколько первых результатов выдачи». Из этого подмножества были исключены 10 случаев, где *что*-придаточное не является сентенциальным актантом имени *причина*, и 3 случая, дублирующие уже имеющиеся в выборке примеры. Оставшиеся примеры (282) распадаются на 3 группы:

- группа I, где *причина того, что P* входит в состав либо субъекта, либо предиката клаузы — 201 пример;
- группа II, где *причина того, что P* образует комплемент предлога *по* (других причинных предлогов в выборке не встретилось) — 75 примеров;
- группа III, где в составе субъекта или предиката клаузы выступает конструкция *причина тому, что P* (6 примеров), выявленная благодаря тому, что в запросе была не оговорена падежная форма эксплетива *то*.

Группы I и II для *причина того, что P* соответствуют группам I и II для *причина что P*. Обращаем внимание на тот факт, что количественные соотношения групп с одним и тем же номером для конструкций с *то* и без *то* существенно различаются: если в группе I для *причина, что P* было всего 10 примеров против 251 примера в группе II ($\approx 4\%$ от общего числа примеров), то для

причина того, что P группа I содержит максимальное число примеров — 201 против 75 в группе II ($\approx 73\%$ от общего числа примеров в обеих группах). Такое соотношение не случайно. Оно предсказывается ограничением сформулированным М. Ю. Князевым: синтаксический акт предикатного имени может быть оформлен со *что* (без *то*), только если это имя входит в коллокационный контекст. Синтаксический акт с эксплетивным *то* не связан этим ограничением, то есть имя *причина* с таким актантом может вступать в свободное сочетание с полнозначными предикатными лексемами разных частей речи, как в (22) и (23):

(22) *В этом, скорее всего, надо искать причину того, что люди так любят бражничать* <...> [Н. Дежнев (2009)]

(23) *Меня терзало полное непонимание причин того, что вдруг произошло.* [Б. Васильев (1998)]

В нашей выборке встретилось 19 полнозначных глаголов, девербативов, предикативов и адъективов, валентность которых заполняло имя *причина*. В основном они относятся к одному из трех семантических классов предикатов:

- бытийному — *быть* (= 'существовать, иметься');
- ментальному — *видеть_2, искать_2, узнать, непонимание, (быть) известным;*
- речевому — *говорить, называть, объяснять, обсуждать, ставить вопрос (о).*

Однако таких свободных сочетаний в группе I меньшинство — всего 31 из 201 ($\approx 15\%$). Все прочие примеры группы I содержат имя *причина* в коллокационных контекстах, аналогичных тем, которые были рассмотрены в разделе 3. То есть, даже в коллокационных контекстах, допускающих конструкцию *причина, что P*, предпочитается конструкция с эксплетивным *то*.

Как и следовало ожидать, низкочастотная конструкция *причина тому, что P*, не выходящая, однако за пределы литературной нормы, благодаря наличию эксплетива *то* встречается как в свободных сочетаниях (24), так и коллокационных контекстах (25) (группа 3):

(24) <...> *причин тому, что столь важный и первостепенный вопрос обеспокоил премьер-министра Касьянова, может быть много.* [«Дело» (Самара), 2002.10.23]

(25) *Все это и послужило одной из причин тому, что широко разрекламированный пистолет Маузер образца 1896 г. официально не был поставлен на вооружение ни в одной стране мира.* [«Боевое искусство планеты» (2004.09.09)]

В синтаксической функции субъекта клаузы или в составе предиката (= глагольной группы) клаузы синтаксический акт предикатного имени *причина*, выражаемый сложной ИГ *того / тому, что P* всегда заполняет валентность ПС, в том числе в контексте эмфатической схемы 3 из раздела 3:

(26) *В том-то и причина того / тому, что люди так любят бражничать.*

Напомним, что в такой же конструкции актант, выражаемый *что*-придаточным без *то* интерпретируется как ИС, ср. выше пример (16).

В примерах группы 2, где конструкция *причина того, что P* выступает в функции компонента предлога *по, что*-придаточное всегда выражает ИС, и в этом отношении не отличается от конструкции *причина, что P*. Однако у конструкции *причина того, что P* полностью отсутствует та вариативность, которая наблюдалась у конструкции *причина, что P* в контексте причинных предлогов. Отсутствует возможность вставки между предлогом и именем *причина* определителя в виде местоимения *то* (**по той причине того, что P*) или определений (**по простой причине того, что P*), вставки вводных слов между именем *причина* и эксплетивом *того* (**по причине, вероятно, того, что P*), нет возможности заменить предлог *по* на синонимичный причинный предлог (**в силу причины того, что P*). Устойчивость словосочетания, *по причине того, что (P)*, утрата словоформой *того* значения однозначного маркера роли ПС у синтаксического акта предикатного имени *причина*, присущего ей во всех прочих контекстах, позволяют считать это выражение неоднословной лексической единицей — составным союзом, в отличие от *по той причине, что (P)*.

5 Значение существительного *причина* в конструкциях *причина, (того) что P*

В [4] участие *что*-придаточного в оформлении валентностей слова *причина* было проигнорировано, в отличие от *чтобы*-придаточного, которое было отнесено исключительно к средствам выражения валентности ПС у лексемы *причина* 2. Валентности какой *причины* выражает *что*-придаточное (с эксплетивным *то* или без него)?

Начнем анализ с употреблений группы 2, где *причина* подчинена причинному предлогу. В этой группе, как показывает Табл. 1, встречаются примеры с согласованным определением при слове *причина*. Среди таких определений встречаются прилагательные, которые, согласно указанным работам, сочетаются только с *причиной* 2. Это прилагательные, указывающие на субъективную оценку степени значимости причин [4: 77]: *важный* (27), *уважительный* (28).

(27) *Наверное, со времен работы в «Сибирских огнях» не чувствовал Саша себя столь легко и свободно. Возможно, по той простой, но важной причине, что планета детства была Плитченко во всех отношениях очень близка.* [«Сибирские огни» (2013)].

(28) *Аллах возьмет «духов» без молитвы по той уважительной причине, что они не успеют этого сделать.* [«Солдат удачи» (2004.03.10)]

Действительно, любовь писателя к детям не влечет с необходимостью, что писатель будет чувствовать себя легко и свободно, работая в детской газете, а лишь делает это вероятным. Аналогичное рассуждение применимо и к (28). Таким образом, в полном соответствии с толкованием, *причина* 2 возможна в составе конструкции *причина, что P*, подчиненной причинному предлогу. А это значит, что, по крайней мере для этого типа употреблений *что*-придаточное должно быть отнесено к средствам выражения валентности ИС лексема *причина* 2 наряду со *чтобы*-придаточным и другими средствами.

Равно возможна в этом типе употреблений и лексема *причина* 1. Это доказывают примеры, в которых ПС не является ни действием субъекта, ни испытываемым им состоянием, ни его желанием действовать, а значит толкование *причины* 2 к ним не приложимо. Но они вполне соответствуют толкованию *причины* 1:

(29) *<...> поиск в горах ночью малоэффективен, по той причине, что ночи в горах темнее, чем на равнине.* [Особенности разведки в горно-лесистой местности. Часть IV. Тактический поиск (2004)]

Действительно, если бы не имела место ИС ‘ночи в горах темнее, чем на равнине’, т.е. ночи были бы не более темными, чем на равнине, то эффективность поисков при прочих равных условиях (том же составе поисковой группы и той же ее оснащенности) была бы достаточной. Причина здесь объективна, следствие не зависит от воли субъекта. Аналогичным образом и в примере (20) мы имеем дело с *причиной* 1. Как поясняет его автор в дальнейшем тексте, жизнь среди «подобных» (= неидеальных людей) с неизбежностью заставит даже идеального человека, «но с ограниченными ресурсами и возможностями», приспособливаться к неидеальным людям, тем самым он сам станет неидеальным, и в результате идеальных людей не будет.

В той же группе встречается и употребление слова *причина* в значении «логической» причины, или импликации, опровергая мнение, что «русское слово *причина* связывается только с онтологической причинностью» [4: 86]:

(30) *Так, русск. пуп и англ. роор неродственны уже по одной той причине, что английское слово имеет такое же начальное р, как русское (а не f, как было бы при родстве) <...>.* [А. Зализняк (2009)]

Итак, употребления группы 2 слова *причина* могут представлять *причину* 1 (объективную, онтологическую), *причину* 2 (субъективную), и *причину* 3 (логическую). Однозначными маркерами причины 2, в конструкции *по (той) причине, что P* могут быть только прилагательные со значением оценки степени значимости (*важный* и т. п.), которые не столь часты в этом контексте. В общем случае для выбора правильной интерпретации необходимо учитывать семантический тип предиката ПС: если он не может обозначать действие, контролируемое состояние или желание действовать личного субъекта, то это не *причина* 2. В прочих случаях выбор интерпретации определяется экстралингвистическими знаниями о каузальных связях между ситуациями.

Среди употреблений группы 1, где имя *причина* выступает в составе субъекта или предиката клаузы, оно также может быть либо *причиной* 1, либо *причиной* 2.

Так, сравнивая примеры (7) и (5), которые относятся к схеме 1 (с эксплетивом и без него соответственно), мы видим, что в (7) выступает объективная *причина* 1: если бы не было решения суда (ИС), родители могли бы забрать ребенка домой из «Дома малютки» (ПС), а (5) это пример на субъективную *причину* 2 из [3: 880]. Проверить правильность определения значения в конструкциях данного типа можно путем замены (*то*), *что P* на диагностическую для субъективной *причины* 2 конструкцию с предлогом *для*. В предложениях типа (7), с объективной *причиной* 1,

такая замена кардинально поменяет смысл предложения, ср.: *Это [решение суда] — единственная причина для того, чтобы мы не могли взять Мишеньку домой.* А в предложениях типа (5) — даст синонимичную перифразу, ср. *И этот гвардеец был причиной для тех больших изменений, которые в ней произошли.* Аналогично, в других конструкциях сложная ИГ *причина, что Р* может обозначать как объективную, так и субъективную *причину*. Так, в примере (10), построенном по модели 2, фигурирует *причина 2*, что доказывает ее синонимичная перифраза *А причина для того, чтобы их [усы] носить, была простая: отец в молодости носил такие же.* В сконструированном нами по той же модели примере (31) выступает уже *причина 1*:

- (31) **Причина, что** планета была обнаруженной, была такая: ее орбитальная скорость была точно как наша.

Замена в (31) *что* на *для того, чтобы* даст прагматически аномальное предложение:

- (32) [?]**Причина для того, чтобы** планета была обнаруженной, была такая: ее орбитальная скорость была точно как наша.

Проведенный нами анализ показал, что *причина* в сочетании *причина (того), что Р*, выступающем в любом возможном для него синтаксическом контексте, при отсутствии явных маркеров *причины 2* (прилагательных оценки значимости) и при соблюдении ограничений, налагаемых *причиной 2* на семантический тип предиката Р (действие, контролируемое состояние или желание действовать) может быть как *причиной 1*, так и *причиной 2*.

6 Заключение

В начале статьи мы продемонстрировали неполноту имеющихся сведений о семантике и морфосинтаксической сочетаемости существительного *причина*. С целью заполнения лакун был проведен семантико-синтаксический анализ представительной выборки примеров, в которых семантические валентности имени *причина* выражались при помощи *что*-придаточного. На основе корпусных данных были описаны схемы (модели) предложений, в которых конструкции *причина, что Р* и *причина того, что Р* выступают в разных синтаксических функциях, и было показано, что в функции компонента причинных предлогов сентенциальный актант имени *причина* заполняет валентность исходной ситуации (причины), а в других синтаксических функциях — производной (следствия). Исключением является схема *В том-то и причина, что Х*, в которой *что*-придаточное заполняет валентность исходной ситуации. В ходе анализа было подтверждено сформулированное М. Ю. Князевым ограничение на сентенциальные актанты существительных: *что*-придаточное без эксплетивного *то* возможно, только если существительное входит в коллокацию. Кроме того, установлено, что *причина* в конструкциях *причина (того), что Р* может иметь как значение объективной причины, так и значение субъективной причины.

Следующий вопрос, естественно вытекающий из проведенного анализа: так как не только *чтобы*-придаточное, но и *что*-придаточное возможны в качестве способов кодирования валентности следствия при лексеме *причина 2*, то какими факторами определяется выбор одного из них и в чем состоит семантическое различие между ними? Этому вопросу мы надеемся посвятить отдельное исследование.

Благодарности

Исследование выполнено при поддержке гранта РФФИ № 22-18-00528 «Связь пропозициональных единиц в предложении и в тексте: семантика и пути грамматикализации» в Институте языкознания РАН. Авторы благодарят фонд за поддержку. Авторы также благодарны анонимным рецензентам «Диалога» за полезные замечания, которые позволили устранить ряд неточностей. Авторы признательны Е. А. Лютиковой за обсуждение отдельных положений статьи и ценные рекомендации.

References

- [1] Apresyan Yu. D. Studies in Semantics and Lexicography. Volume I. Paradigmatics [Issledovaniya po semantike i leksikografii. Tom I. Paradigmatika]. Moscow: Yazyki slavyanskikh kul'tur, 2009.
- [2] Arutyunova N. D. Types of meanings: Evaluation. Event. Fact [Tipy yazykovykh znacheniy: Otsenka. Sobytie. Fakt] Moscow: Nauka, 1988.
- [3] Boguslavskaya O. Yu. Cause 2, basis 5, reason 1 [Prichina 2, osnovanie 5, rezon 1] // Apresyan Yu. D. (ed.) The New Explanatory Dictionary of Russian Synonyms 2nd ed. [Novyy ob'yasnitel'nyy slovar' sinonimov russkogo yazyka. 2-e izd]. Moscow: Shkola «Yazyki slavyanskoy kul'tury», 2003. P. 877–882.
- [4] Boguslavskaya O. Yu., Levontina I. B. Meanings 'cause' and 'purpose' in natural language [Smysly 'prichina' i 'tsel' v estestvennom yazyke] // Topics in the study of language [Voprosy yazykoznanija]. 2004. № 2. P. 68–88.
- [5] Bunt H., Prasad R. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations // Proceedings of the LREC 2016 Workshop “ISA-12: 12th Joint ACL — ISO Workshop on Interoperable Semantic Annotation”. Portorož, 2016. P. 45–54.
- [6] Google Books. Access mode: <https://books.google.ru/?hl=ru>.
- [7] Grimshaw J. Argument structure. Cambridge, MA: MIT Press, 1990.
- [8] In'kova O. Yu. Language-specific connectives: description methods and parameters [Lingvospetsifichnost' konnektorov: metody i parametry opisaniya] // In'kova O. Yu. (ed.) Semantics of connectives: contrastive study [Semantika konnektorov: kontrastivnoe issledovanie]. Moscow: TORUS PRESS, 2018. P. 5–23.
- [9] Kibrik A. A. Discourse analysis in cognitive perspective [Analiz diskursa v kognitivnoy perspektive]. Doct. diss. Moscow: Institute of Linguistics, 2003.
- [10] Knyazev M. Licensing clausal complements: The case of čto-clauses. Doct. diss. Utrecht: LOT Publications, 2016.
- [11] Knyazev M. Yu. Studying a weak grammatical violation with experimental syntax methods: the case of sentential complements of nouns with the complementizer čto [Issledovanie «slabogo» grammaticheskogo ograničeniya metodami eksperimental'nogo sintaksisa: primer pridatochnykh s soyuzom čto v funktsii sententsial'nogo aktanta sushchestvitel'nogo] // Rhema. 2017. № 1. P. 22–40.
- [12] Kobozeva I. To for completive that-clause: conditions of use [Usloviya upotrebleniya to pered pridatochnym iz'yasnitel'nyim s soyuzom čto] // In'kova O. (ed.) Du mot au texte. Études slavo-romanes. Berne: Peter Lang, 2013. P. 129–148.
- [13] Kobozeva I. M. Connectives of immediate precedence in Russian and French based on parallel corpus data [Konnektory kontaktnogo predshestvovaniya vo frantsuzskom i russkom yazykakh po dannym parallel'nogo korpusa] // Contrastive Linguistics. 2017. Vol. 17, № 4. P. 48–62.
- [14] Letuchiy A. B. Complement clauses of nouns in Russian [Sententsial'nye aktanty imen v russkom yazyke] // Typology of morphosyntactic parameters [Tipologiya morfosintaksicheskikh parametrov]. 2. P. 46–66.
- [15] Lyutikova E. A. Towards the categorial status of noun phrases in Russian [K voprosu o kategorial'nom statuse imennykh grupp v russkom yazyke] // Moscow State University Bulletin. Series 9. Philology [Vestnik Moskovskogo universiteta. Seriya 9. Filologiya]. 2010. №6. P. 36–76.
- [16] Paducheva E. V. Predicative nouns in terms of lexicography [Predikatnye imena v leksikograficheskom aspekte] // Paducheva E. V. Collection of articles [Stat'i raznykh let]. Moscow: Yazyki slavyanskikh kul'tur, 2009. P. 317–335.
- [17] Russian National Corpus [Natsional'nyy korpus russkogo yazyka]. Access mode: <https://ruscorpora.ru/new/>.
- [18] Shvedova N. Yu. (ed.) Russian grammar. Vol. II. Syntax [Russkaya grammatika. Tom II. Sintaksis]. Moscow: Nauka, 1980.
- [19] Stede M. Discourse processing, San Rafael: Morgan & Claypool Publ., 2012.
- [20] Stepin V. S. (ed.) The New Encyclopedia of Philosophy: In 4 vols. [Novaya filosofskaya entsiklopediya: V 4 t.]. Moscow: Mysl', 2001.
- [21] Testeleys Ya. G. Introduction to general syntax [Vvedenie v obshchii sintaksis]. Moscow: RGGU, 2001.

Suffixal imperfectivization of prefixed verbs: Record breakers and outsiders (based on the dictionary, corpus and Runet)

Elena V. Gorbova

Herzen State Pedagogical University of
Russia

elenagorbova12@gmail.com

Oksana Iu. Chuikova

Herzen State Pedagogical University of
Russia

oxana.chuykova@gmail.com

Abstract

The paper sums up the results of the study on the imperfectivability of the Russian prefixed verb using such sources of linguistic data as the Small Academic Dictionary, the corpus (RNC – ruscorpora.ru) and Runet. The focus is on those subsets of the set of prefixed perfectives that show specificity in relation to suffixal imperfectivization, differing in increased (record breakers) and decreased (outsiders) imperfectivability compared to the average level. The former are represented by denominatives and prefixed perfectives derived from perfective simplex stems, while the latter are represented by most Aktionsarten and verbs with the *-(zi)rova-* formant. Systemic and morphological explanations for the specifics of these subsets are proposed.

Keywords: Russian language; prefixed verb; suffixal imperfectivization; imperfectivability; imperfectivability scale and its poles

DOI: 10.28995/2075-7182-2022-21-247-263

Суффиксальная имперфективизация приставочных глаголов: рекордсмены и аутсайдеры (в словаре, корпусе и Рунете)

Елена Викторовна Горбова

РГПУ им. А. И. Герцена
elenagorbova12@gmail.com

Оксана Юрьевна Чуйкова

РГПУ им. А. И. Герцена
oxana.chuykova@gmail.com

Аннотация

Статья подводит итоги исследования имперфективизируемости русского приставочного глагола с использованием таких источников языкового материала, как словарь (МАС) корпус (НКРЯ) и Рунет. В фокусе внимания находятся те подмножества совокупности приставочных перфективов, которые обнаруживают специфику по отношению к суффиксальной имперфективизации, отличаясь повышенной (рекордсмены) и пониженной (аутсайдеры) имперфективизируемостью в сопоставлении со средним уровнем. Первые репрезентированы отыменными и отперфективными дериватами, а вторые – большинством способов действия и глаголами с формантом *-и(зи)рова-*. Предложены системные и морфонологические объяснения специфике этих подмножеств.

Ключевые слова: русский язык; приставочный глагол; суффиксальная имперфективизация; имперфективизируемость; шкала имперфективизируемости и ее полюса

1 Введение

Задачей статьи является обсуждение двух крайних точек на шкале имперфективизируемости (способности к суффиксальной имперфективизации типа *про-жи-ть*^{СВ} → *про-жи-ва-ть*^{НСВ}) русских приставочных глаголов совершенного вида (СВ, перфективов): групп (или подгрупп) приставочных глаголов с максимально высокой степенью имперфективизируемости (рекордсменов) и минимально низкой (аутсайдеров). Эта задача решается на основе результатов проекта по изучению имперфективизируемости (далее — ИМПФ) приставочных глаголов, реализованного в 2019–2020 гг.

Работа строится следующим образом. Раздел 2 посвящен характеристике используемых понятий, описанию материала исследования и его методике. В разделе 3 представлен обзор общих результатов исследования ИМПФ по всем приставочным группам (ПГ) как фон для обсуждаемых вопросов, а также ставится задача отдельно рассмотреть два полюса на шкале ИМПФ – максимальный и минимальный. Разделы 4 и 5 посвящены множествам ПГ, оказавшимся на максимальном (отыменные и отперфективные) и на минимальном (способы действия и глаголы на *-и(зи)рова-*) полюсах ИМПФ. В разделе 6 подводятся итоги.

2 Используемые понятия, материал и методика исследования

2.1 Используемые понятия

Все множество русских глаголов может быть разделено на подмножества производных глаголов-симплексов (далее – *симплексы*): *спать, дать, лечь, махать*, и производных от них глаголов типа *про-спать, за-дать, у-лечь-ся, мах-ну-ть*.

Подмножество симплексов в зависимости от их аспектуального функционирования делятся на глаголы НСВ, или имперфективы, типа *спать, махать*, и глаголы СВ, или перфективы, типа *дать, лечь*. Первые будут обозначены как первичные имперфективы, или НСВ₁, вторые – как первичные перфективы, или СВ₁.

Из производных глаголов нас будет интересовать лишь одно из подмножеств – те дериваты, словообразовательный формант которых на последнем шаге деривации является префиксом или включает префикс наряду с суффиксом, но не с суффиксом имперфективации (*есть*→*на-есть-ся*)¹. Такие производные глаголы будем называть приставочными.

Они различаются видом и далее обозначаются как *вторичные перфективы*, СВ₂ (*про-спать, за-дать, у-лечь-ся, при-лечь, под-махнуть*), и *вторичные имперфективы*, НСВ₂ (*про-сып-а-ть, пере-да-ва-ть, под-мах-ива-ть*). В случае множественной префиксации с увеличением количества префиксов соответственно увеличивается индекс: *за-дать*^{СВ₂}→*пере-за-дать*^{СВ₃}. Соответственно, *пере-за-да-ва-ть* следовало бы индексировать как НСВ₃, однако поскольку эта форма образуется от *перезадавать*^{СВ₃} при помощи того же суффиксального механизма, который использован в *задавать*^{НСВ₂}←*задать*^{СВ₂}, продолжим использовать традиционный для русской аспектологии термин «вторичная имперфективация» в том числе при образовании имперфектива от полипрефиксального глагола: [*пере-[за-[да]*^{СВ₁}]^{СВ₂}]^{СВ₃}*ва]*^{НСВ₂}*ть*.

Итак, *вторичная* (или суффиксальная) *имперфективация* – это аффиксальный механизм образования от приставочного перфектива приставочного же имперфектива путем введения в основу суффиксального формата *-(и/ы/е)ва-* (или более архаичного *-а-*), т. е. СВ₂ (СВ₃, СВ_n)→НСВ₂².

Имперфективируемость понимается как способность приставочных перфективов к суффиксальной имперфективации.

Шкала имперфективируемости – распределение множеств и подмножеств русских приставочных глаголов по уровню выявленной ИМПФ. Максимально возможный уровень равен 100%; это такое положение, при котором каждый приставочный перфектив образует НСВ₂: *за-кры-ть*^{СВ₂}→*за-кры-ва-ть*^{НСВ₂}.

2.2 Материал исследования

Предметом изучения являются приставочные, или префигированные, перфективы, которые 1) являются приставочными дериватами от исходных симплексов: (*жи-ть*^{НСВ₁})→*про-жи-ть*^{СВ₂}→*про-жи-ва-ть*^{НСВ₂}; (*да-ть*^{СВ₁})→*от-да-ть*^{СВ₂}→*от-да-ва-ть*^{НСВ₂}; и 2) их префиксы-перфективаторы выделяются синхронно (в соответствии с [4]).

¹ Префикс в глагольной основе может быть один, в этом случае глагол монопрефиксальный, либо их более одного, тогда глагол полипрефиксальный.

² Отметим два момента. 1) Термин «вторичная имперфективация» здесь условен, поскольку неточен в двух отношениях: для образования НСВ₂ механизм имперфективации применяется впервые (имперфективность симплекса типа *спать* обусловлена его акциональностью, она не маркируется показателем); в случаях типа *пере-за-дать*^{СВ₃}→*пере-за-да-ва-ть*^{НСВ₂} происходит «понижение» индекса; осознавая условность термина, оставляем его как традиционный и привычный. 2) Мы признаем «запрет на третичную имперфективацию», обоснованный в [6].

Перечень анализируемых приставочных СВ получен методом сплошной выборки по МАС [3]. Единицей в созданном перечне (наборе перечней, организованном по левопериферийной приставке) выступает не вокабула, а моносемичная лексема, понимаемая как глагол в одном из значений, выделенных в МАС. В качестве примера работы с моносемичными лексемами приведем лексему *отдуть* 2 «Отколотить, избить». В МАС НСВ₂ у этой лексемы отсутствует³. Однако реализацию имперфектива этой лексемы можно обнаружить как в НКРЯ, см. (1), так и в Рунете, см. (2).

(1) — *Вас папа просит, — почти закричал он на него: — там я хлопочу одну девушку определить к нам в кастаньяши, и если вы отговорите папа, я вас отдую^{НСВ₂} за то! — заключил Николая и показал кулак Феодосию Ивановичу.*
— *Да погодите еще отдувать^{НСВ₂}-то! — ответил тот ему и пошел в присутствие.*
[А. Ф. Писемский. В водвороте (1871)]

(2) *Почему русские так любят отдувать^{НСВ₂} евреев? quoted1. Ну на этот вопрос лучше Веллера никто не ответит.* (https://www.politforums.net/world/1368974664_2.htm)

2.3 Методика исследования

Используемая методика представляет собой существенно расширенный и дополненный вариант методики, примененной для анализа так называемых *про*-глаголов в [7]. Расширение обусловлено охватом всех ПГ русского глагола, выделенных методом сплошной выборки по МАС, а дополнение – введением в круг источников языкового материала НКРЯ⁴ и Рунета (через поисковые системы Яндекс и Google).

Работа с каждой ПГ проводилась в несколько этапов: 1) сплошная выборка приставочных моносемичных перфективных лексем по МАС и установление для них НСВ₂ по трем источникам: словарю (в таблицах столбец «в МАС»), словарю и НКРЯ (столбец «в НКРЯ»), словарю и Рунету («в Рунете»); 2) выделение в полученном множестве ПГ подмножеств: отыменных дериватов, отперфективных перфективов (деривация СВ→СВ), глаголов на *-и(зи)рова-*, способов действия (СД; префикс – формант или часть форманта СД); 3) формирование коллекций выборок, в число которых входят максимальная (тах-выборка: исходная выборка без отыменных) и минимальная (min-выборка: тах-выборка за вычетом отперфективных, глаголов на *-и(зи)рова-*, СД)⁵; 4) статистическая обработка полученных данных в рамках ПГ.

После чего осуществлялась общая оценка регулярности суффиксальной имперфективации приставочных перфективов по всем ПГ в совокупности. Представление результатов этого этапа и является нашей актуальной задачей.

Предваряя вопросы, прокомментируем еще один момент: перечень ПГ глаголов соотносим с перечнем глагольных приставок. Последний был сформирован по результатам более раннего корпусного исследования на базе НКРЯ, [2]. Запрос в основном корпусе НКРЯ включал в себя грамматические признаки: глагол, НСВ, наличие приставки. Соответственно, на выходе были получены ПГ, подвергшиеся вторичной имперфективации. Именно эти приставки и составили рабочий перечень из 21 приставки: *в(о)-, в(о)з(о)/в(о)с-, вы-, до-, за-, из(о)-/ис-, на-, над(о)-, недо-, о-/об(о)-, от(о)-, пере-, по-, под(о)-, пре-, пред(о)-, при-, про-, раз(о)/с-, с(о)-, у-* (при раздельном рассмотрении *о-* и *об(о)-* — 22 приставки).

3 Имперфективируемость русских приставочных глаголов: шкала и два полюса

В этом разделе приводятся данные по установленной ИМПФ приставочных глаголов. В самом общем виде они представлены в Табл. 1.

³ При этом у некоторых лексем этой вокабулы НСВ₂ в МАС имеются; в частности, *отдуть* 1 («Сов. к отдувать», толкуемый как «Дуновением, дуя, относить, отгонять и т. п. назад, в сторону»).

⁴ URL: <https://ruscorpora.ru/new/search-main.html>

⁵ В дальнейшем приводятся в основном min-выборки ПГ, поскольку именно они «очищены» от различных выбросов – подмножеств внутри ПГ со спецификой ИМПФ.

	в МАС	% в МАС	в НКРЯ	% в НКРЯ	в Рунете	% в Рунете
<i>по-</i>	253	31,12	354	43,54	561	69
<i>из(о)-/ис-</i>	357	68,52	399	76,58	450	86,37
<i>на-</i>	549	70,03	603	76,91	697	88,9
<i>с(о)-</i>	1040	75,2	1094	79,1	1257	90,89
<i>в(о)з(о)-/в(о)с-</i>	250	66,14	288	76,19	344	91,01
<i>про-</i>	629	68,82	708	77,46	839	91,79
<i>за-</i>	1078	76,95	1177	84,01	1299	92,72
<i>вы-</i>	777	82,92	849	90,61	889	94,88
<i>пере-</i>	415	81,21	454	88,85	485	94,91
<i>от(о)-</i>	692	85,64	734	90,84	770	95,3
<i>о(б)-/об(о)-</i>	1073	80,86	1158	87,26	1267	95,48
<i>у-</i>	778	85,87	801	88,41	875	96,58
<i>в(о)-</i>	289	92,63	301	96,47	305	97,76
<i>над(о)-</i>	39	86,67	45	100	44	97,78
<i>под(о)-</i>	406	95,53	418	98,35	418	98,35
<i>раз(о)-/рас-</i>	877	89,86	926	94,88	962	98,57
<i>при-</i>	544	90,97	571	95,48	593	99,16
<i>до-</i>	96	93,2	97	94,17	103	100
<i>недо-</i>	37	88,1	41	97,62	42	100
<i>пред(о)-</i>	24	100	24	100	24	100
<i>пре-</i>	46	97,87	47	100	47	100
Среднее	488,05	81,34	528,05	87,46	584,33	94,26
Медиана	415	85,64	454	90,61	561	95,48

Таблица 1: Имперфективируемость ПГ (min-выборки, сортировка по возрастанию в столбце «% в Рунете»)

Как можно видеть по Табл. 1, уровень ИМПФ (в процентах) в целом по всем ПГ в совокупности и, как правило, в каждой ПГ, постепенно повышается от такого источника, как МАС, до суммарных данных МАС+НКРЯ и МАС+Рунет: от 81% через 87% к 94% (см. предпоследнюю строку; с округлением до целых).

При этом хорошо заметны сильно различающиеся уровни ИМПФ различных ПГ: в Табл. 1 вверху приведены *по-*перфективы с наименьшим (на 18 процентных пунктов отличающимся от ближайшего соседа) уровнем ИМПФ в 69%, а внизу – четыре ПГ (*до-*, *недо-*, *пред(о)-* и *пре-*перфективы) со 100% по МАС+Рунет. Тем самым ПГ в своей совокупности также являют собой шкалу ИМПФ.

Однако еще более яркое различие по уровню ИМПФ демонстрируют отдельные подмножества приставочных перфективов, организованные не по единству приставки, а по другим основаниям, и пересекающие все или большинство ПГ.

Таковыми являются подмножества: отыменных перфективов (ОИП), отперфективных перфективов (ОПП), глаголов на *-и(зи)рова-*, СД. Из них два – ОИП и ОПП – показывают более высокий уровень ИМПФ (в среднем по всем ПГ в сравнении со средним уровнем по min-выборкам), и еще два – глаголов на *-и(зи)рова-* и СД – обнаруживают сниженный уровень ИМПФ. Тем самым соответствующие показатели образуют шкалу ИМПФ, где на минимальном полюсе расположены перфективы на *-и(зи)рова-*, за ними идут СД, далее располагаются ПГ, очищенные от специализированных подмножеств (часть из которых близка к максимальному полюсу), и – полюс максимальной ИМПФ, где располагаются ОПП и ОИП.

Ниже по отдельности рассматриваются подмножества обоих полюсов ИМПФ.

4 Рекордсмены имперфективируемости: ОИП и ОПП

Отыменные и отперфективные перфективы обнаруживают сходство: в деривационной истории входящих в них лексем отсутствуют НСВ₁, т.е. суффиксальная имперфективация от таких СВ в обоих случаях ни в одном из известных смыслов не является вторичной.

4.1 Отыменные перфективы

Это префигированные глаголы СВ, образованные непосредственно от именной (напр., *прошляпить*←*шляпа*, *облесить*←*лес*) или адвербиальной (напр., *переиначить*←*иначе*) основы. После установления отыменности такие глаголы исключались из исходных выборок, вследствие чего они не входят ни в max-, ни в min-выборки (решение из [7]).

ОИП обнаружены в 18 из 21 ПГ. Однако приставки участвуют в деривации отыменных СВ в неравной степени, образуя диапазон от невозможности такой деривации (*до-*, *недо-* и *пред(о)-*) до доли в 18–21% отыменных дериватов в ПГ (*пре-* и *о(б)-/об(о)-*), см. Табл. 2.

	в МАС	% в МАС	в НКРЯ	% в НКРЯ	в Рунете	% в Рунете	количество	% в исходной выборке
<i>до-</i>	0	0	0	0	0	0	0	0
<i>недо-</i>	0	0	0	0	0	0	0	0
<i>пред(о)-</i>	0	0	0	0	0	0	0	0
<i>про-</i>	20	74,07	22	81,48	24	88,89	27	1,88
<i>на-</i>	8	88,89	8	88,89	8	88,89	9	0,6
<i>по-</i>	28	87,5	28	87,5	29	90,63	32	1,8
<i>за-</i>	42	70	44	73,33	56	93,3	60	2,78
<i>о(б)-/об(о)-</i>	280	81,63	296	86,3	325	94,75	343	21,33
<i>у-</i>	109	93,97	112	96,55	112	96,55	116	10,53
<i>с(о)-</i>	46	79,31	47	81,03	56	96,55	58	1,09
<i>в(о)-</i>	24	100	24	100	24	100	24	6,27
<i>в(о)з(о)-/в(о)с-</i>	1	50	1	50	2	100	2	0,42
<i>вы-</i>	8	100	8	100	8	100	8	0,77
<i>из(о)-/ис-</i>	9	100	9	100	9	100	9	1,56
<i>над(о)-</i>	1	100	1	100	1	100	1	2,08
<i>от(о)-</i>	19	100	19	100	19	100	19	1,71
<i>пере-</i>	10	100	10	100	10	100	10	0,8
<i>под(о)-</i>	6	66,67	8	88,89	9	100	9	1,26
<i>пре-</i>	16	100	16	100	16	100	16	18,39
<i>при-</i>	32	86,49	35	94,59	37	100	37	4,28
<i>раз(о)-/рас-</i>	27	96,43	27	96,43	28	100	28	1,87
Среднее	38,11	87,50	39,72	90,28	42,94	97,20	44,89	4,41
Медиана	19,5	91,43	20,5	95,51	21,5	100	21,5	1,84

Таблица 2: Имперфективируемость ОИП (сортировка по возрастанию в столбце «% в Рунете»)

Табл. 2 показывает, что средние показатели ИМПФ ОИП (при учете данных по ПГ с ненулевым количеством ОИП) выше, чем по ПГ в целом (min-выборки, см. Табл. 1)⁶.

В Табл. 3 представлены результаты применения статистического критерия согласия Пирсона (хи-квадрат с поправкой Йетса, двусторонний вариант) к суммарным данным по ИМПФ ОИП и глаголов, входящих в min-выборки (по 18 ПГ, за исключением *до-*, *недо-* и *пред(о)-*глаголов).

⁶ То же верно при сравнительном рассмотрении средних показателей по ОИП и min-выборкам 18 ПГ, содержащих ОИП (см. Приложение А).

		Есть НСВ2	Нет НСВ2	Всего	доля НСВ2
МАС	ОИП	686	122	808	84,90%
	Min-выборки	10092	2994	13086	77,12%
	Всего	10778	3116	13894	
	Различие статистически значимо: $\chi^2=26,034$, $p \ll 0,01$				
МАС+НКРЯ	ОИП	715	93	808	88,49%
	Min-выборки	10927	2159	13086	83,50%
	Всего	11642	2252	13894	
	Различие статистически значимо: $\chi^2=13,580$, $p=0,0002$				
МАС+Рунет	ОИП	773	35	808	95,67%
	Min-выборки	12102	984	13086	92,48%
	Всего	12875	1019	13894	
	Различие статистически значимо: $\chi^2=10,915$, $p=0,001$				

Таблица 3: Имперфективируемость ОИП в сравнении с min-выборками (по 18 ПГ)

Табл. 3 подтверждает, что по совокупным данным всех ПГ уровень ИМПФ ОИП выше соответствующих показателей для min-выборок по всем рассматриваемым источникам языкового материала: МАС, МАС+НКРЯ, МАС+Рунет.

Единственной ПГ, для которой последовательно, по всем источникам, в отличие от других ПГ по отдельности, наблюдается статистически значимо более высокий уровень ИМПФ в подмножестве ОИП по сравнению с соответствующей min-выборкой, являются *по*-глаголы (см. Приложение А). Возможные интерпретации: 1) в очередной раз подтверждается особый статус *по*-глаголов среди других ПГ; 2) напротив, подмножество отыменных *по*-глаголов в меньшей степени отличается от других приставочных перфективов, поскольку именно min-выборка *по*-глаголов демонстрирует необычно низкий уровень ИМПФ (см. первую строку в Табл. 1). Оставляем обе интерпретации в качестве гипотез.

4.2 Отперфективные перфективы

Это префигированные перфективы, на последнем шаге деривации образованные от глагола СВ, т.е. от перфектива. Могут быть как монопрефиксальными (*про-пустить*^{СВ2} ← *пустить*^{СВ1}), так и полипрефиксальными (*по-вы-бить*^{СВ3} ← *вы-бить*^{СВ2}).

ОПП присутствуют во всех без исключения ПГ, их доля в тах-выборке колеблется от 2,81% (*из(о)-/ис*-глаголы) до 59,32% (*пред(о)*-глаголы).

В Табл. 4 приводятся данные об ИМПФ ОПП.

Данные в Табл. 4 не позволяют сделать однозначный вывод о повышенном уровне ИМПФ ОПП в сопоставлении с показателями по min-выборкам (см. Табл. 1): если сопоставить такой показатель, как среднее арифметическое, то ИМПФ по min-выборкам окажется даже несколько выше, чем по отперфективным. Таким образом, в общем случае ОПП не столь однозначно отличаются повышенной ИМПФ, как ОИП. Приведем, однако, дополнительные наблюдения.

- 1) Имеет место довольно пестрая картина: отдельные ПГ статистически значимо демонстрируют более высокий уровень ИМПФ как ОПП, так и соответствующих min-выборок (см. Приложение Б).
- 2) Единственная ПГ, в которой по всем источникам статистически значимо наблюдается сниженный уровень ИМПФ ОПП в сопоставлении с min-выборкой, – это *по*-глаголы.

При изъятии *по*-глаголов из общих данных ПГ суммарный уровень ИМПФ оставшихся ОПП, оказывается выше, чем в рамках min-выборки по всем источникам материала, см. Табл. 5.

	в МАС	% в МАС	в НКРЯ	% в НКРЯ	в Рунете	% в Рунете	количество	% в max-выборке
<i>по-</i>	23	12,3	52	27,81	93	49,73	187	10,7
<i>из(о)-/ис-</i>	12	75	13	81,25	14	87,5	16	2,81
<i>о(б)-/об(о)-</i>	84	83,17	86	85,15	90	89,11	101	6,9
<i>в(о)з(о)-/в(о)с-</i>	72	77,42	77	82,8	85	91,4	93	19,57
<i>с(о)-</i>	117	87,31	122	91,04	124	92,54	134	8,42
<i>от(о)-</i>	98	89,09	103	93,64	102	92,73	110	10,11
<i>до-</i>	13	86,67	13	86,67	14	93,33	15	3,73
<i>на-</i>	67	88,16	70	92,11	71	93,42	76	5,2
<i>у-</i>	49	81,67	53	88,33	57	95	60	6,09
<i>под(о)-</i>	106	81,54	115	88,46	124	95,38	130	18,47
<i>пре-</i>	22	91,67	22	91,67	23	95,83	24	33,8
<i>в(о)-</i>	49	94,23	50	96,15	50	96,15	52	14,25
<i>раз(о)-/рас-</i>	112	85,5	119	90,84	127	96,95	131	8,93
<i>при-</i>	144	81,82	157	89,2	172	97,73	176	21,28
<i>за-</i>	95	90,48	99	94,29	103	98,1	105	5,01
<i>вы-</i>	74	93,67	76	96,2	78	98,73	79	7,75
<i>про-</i>	96	92,31	100	96,15	103	99,04	104	7,4
<i>пере-</i>	188	92,16	197	96,57	203	99,51	204	16,45
<i>над(о)-</i>	0	0	1	50	2	100	2	4,26
<i>недо-</i>	8	88,89	9	100	9	100	9	17,65
<i>пред(о)-</i>	34	97,14	35	100	35	100	35	59,32
Среднее	69,67	79,53	74,71	86,59	79,95	93,44	87,76	13,72
Медиана	72	87,31	76	91,04	85	95,83	93	8,93

Таблица 4: Имперфективируемость ОПП (сортировка по возрастанию в столбце «% в Рунете»)

		Есть НСВ2	Нет НСВ2	Всего	доля НСВ2
МАС	ОПП	1440	216	1656	86,96%
	Min-выборки	9996	2446	12442	80,34%
	Всего	11436	2662	14098	
Различие статистически значимо: $\chi^2=41,331$, $p<<0,01$					
МАС+НКРЯ	ОПП	1517	139	1656	91,61%
	Min-выборки	10735	1707	12442	86,28%
	Всего	12252	1846	14098	
Различие статистически значимо: $\chi^2=35,964$, $p<<0,01$					
МАС+Рунет	ОПП	1586	70	1656	95,77%
	Min-выборки	11710	732	12442	94,12%
	Всего	13296	802	14098	
Различие статистически значимо: $\chi^2=7,167$, $p=0,0074$					

 Таблица 5: Имперфективируемость ОПП в сравнении с min-выборками (за вычетом *по-*глаголов)

Итак, вновь можно отметить специфику *по-*перфективов (в части подмножества отперфективных *по-*дериватов) на общем фоне русских префигированных перфективов (более подробно см. в [1]).

5 Аутсайдеры имперфективируемости: СД и глаголы на *-и(зи)рова-*

Ниже рассмотрим группы глаголов – СД и глаголы на *-и(зи)рова-*, – показывающие в целом сниженный уровень ИМПФ относительно средних данных по ПГ.

5.1 СД и шкала имперфективируемости

Согласно одному из определений, «[с]пособами глагольного действия принято называть различные типы семантических модификаций глагола, выраженные определенными формальными средствами» [8: 110]. При составлении перечня СД использовались классификации из [8: 110–135] и [5: 141–167], несколько модифицированные с опорой на положения:

1) СД обязательно имеет формальный показатель; СД, выделяемые в классификации Шелякина исключительно на семантических основаниях, не рассматривались;

2) СД не может занимать полный объем ПГ; так, в перечень рассматриваемых СД не включались выделяемые в классификации Шелякина недостаточно-нормативный СД (с показателем *недо-*) и проспективный СД (с *пред(о)-*);

3) при принятии решения о (не)объединении глаголов в один СД соблюдалось тождество знака: один показатель и разные значения – омонимия (разные СД), один показатель при наличии нетривиального семантического сходства значений – полисемия (один СД), разные показатели при сходстве семантики – синонимия (разные СД); напр., аттенуативный СД с префиксом *по-* и аттенуативный СД с префиксом *при-* учитываются как отдельные синонимичные СД; однако различаемые в [5] финально-комплетивный, комплетивно-партитивный и терминативно-локальный СД с префиксальным показателем *до-* (*долететь, дочитать, досыпать*) объединены в **комплетивный** СД, а финально-отрицательный и длительно-усилительный (*до...ся: доиграться и долежаться*) – в единый СД, условно названный **дуративно-негативным**.

В Табл. 6 приводится перечень СД с указанием формального показателя и данные об их ИМПФ.

способы действия	в МАС	% в МАС	в НКРЯ	% в НКРЯ	в Рунете	% в Рунете	% в тах-выборке
<i>от(о)-</i> : финитивный	0	0	25	17,24	47	32,41	13,31
<i>по-</i> : дистрибутивный	0	0	26	15,48	62	36,9	9,61
<i>по-</i> : ингрессивный	0	0	9	13,04	26	37,68	3,95
<i>за-</i> : начинательный	11	2,29	43	8,96	190	39,58	22,89
<i>по-</i> : аттенуативный	0	0	11	17,74	30	48,39	3,54
<i>про-</i> : пердуративный	16	4,58	81	23,21	183	52,44	24,8
<i>из-...-ся</i> : чрезмерно-кратный	8	28,57	10	35,71	16	57,14	4,95
<i>на-...-ся</i> : сатуративный/сативный	38	20,54	54	29,19	130	70,27	12,65
<i>в(о)з(о)-/в(о)с-</i> : делимитативный	0	0	2	28,57	5	71,43	1,47
<i>по-</i> : делимитативный	117	20,49	248	43,43	417	73,03	32,65
<i>раз(о)/с-...(-ся)</i> : ингрессивно-интенсивный	74	36,1	116	56,59	169	82,44	13,97
<i>на-</i> : кумулятивный	257	61,93	280	67,47	347	83,61	28,39
<i>пере-</i> : дистрибутивный	174	53,7	221	68,21	274	84,57	26,13
<i>за-...-ся</i> : чрезмерно-длительный	26	50	30	57,69	44	84,62	2,48
<i>у-...-ся</i> : чрезмерно-интенсивный	8	44,44	9	50	16	88,89	1,83
<i>до-...-ся</i> : дуративно-негативный	21	53,85	22	56,41	37	94,87	9,7
<i>при-</i> : аттенуативный	72	77,42	78	83,87	89	95,7	11,25
<i>о(бо)-</i> : дистрибутивно-суммарный	24	77,42	29	93,55	30	96,77	2,12
<i>под-</i> : аттенуативный	149	81,87	160	87,91	177	97,25	25,85
<i>до-</i> : комплетивный	229	91,98	234	93,98	244	97,99	61,94
<i>раз(о)-/рас-</i> : дистрибутивно-суммарный	141	89,81	150	95,54	154	98,09	10,7
<i>пере-</i> : репетитивный	235	87,69	245	91,42	267	99,63	21,61
Среднее:	72,73	40,12	94,68	51,60	134,27	73,80	15,72
Медиана	25	40,27	48,5	53,21	109,5	83,03	11,95

Таблица 6: Имперфективируемость СД (сортировка по возрастанию в столбце «% в Рунете»)

Данные в Табл. 6 показывают, что средние показатели ИМПФ СД значительно ниже аналогичных показателей по min-выборкам (см. Табл. 1). Уточним, что СД обнаруживаются в составе 14 из 21 ПГ, поэтому корректно будет сопоставить данные по ИМПФ СД со средними данными по 14 ПГ, содержащим СД (см. Табл. 7).

	в МАС	% в МАС	в НКРЯ	% в НКРЯ	в Рунете	% в Рунете
<i>по-</i>	253	31,12	354	43,54	561	69
<i>из(о)-/ис-</i>	357	68,52	399	76,58	450	86,37
<i>на-</i>	549	70,03	603	76,91	697	88,9
<i>в(о)з(о)-/вос-</i>	250	66,14	288	76,19	344	91,01
<i>про-</i>	629	68,82	708	77,46	839	91,79
<i>за-</i>	1078	76,95	1177	84,01	1299	92,72
<i>пере-</i>	415	81,21	454	88,85	485	94,91
<i>от(о)-</i>	692	85,64	734	90,84	770	95,3
<i>о(б)-/об(о)-</i>	1073	80,86	1158	87,26	1267	95,48
<i>у-</i>	778	85,87	801	88,41	875	96,58
<i>под(о)-</i>	406	95,53	418	98,35	418	98,35
<i>раз(о)-/рас-</i>	877	89,86	926	94,88	962	98,57
<i>при-</i>	544	90,97	571	95,48	593	99,16
<i>до-</i>	96	93,2	97	94,17	103	100
Среднее	571,21	77,48	620,57	83,78	690,21	92,72
Медиана	546,5	81,04	587	87,84	645	95,10

Таблица 7: Имперфективируемость ПГ (min-выборки ПГ, содержащих СД, сортировка по возрастанию в столбце «% в Рунете»)

Гипотеза о сниженном уровне ИМПФ СД в сопоставлении с данными min-выборки подтверждается и на суммарных данных (на основании 14 ПГ), см. Табл. 8.

	Есть НСВ2	Нет НСВ2	Всего	доля НСВ2
МАС				
СД	1600	2497	4097	39,05%
Min-выборки	7997	2468	10465	76,42%
Всего	9597	4965	14562	
Различие статистически значимо: $\chi^2=1827,576$, $p<<0,01$				
МАС+НКРЯ				
СД	2083	2014	4097	50,84%
Min-выборки	8688	1777	10465	83,02%
Всего	10771	3791	14562	
Различие статистически значимо: $\chi^2=1581,477$, $p<<0,01$				
МАС+Рунет				
СД	2954	1143	4097	72,10%
Min-выборки	9663	802	10465	92,34%
Всего	12617	1945	14562	
Различие статистически значимо: $\chi^2=1039,968$, $p<<0,01$				

Таблица 8: Имперфективируемость глаголов СД в сравнении с min-выборками (по 14 ПГ)

Табл. 8 подтверждает, что суммарный уровень ИМПФ СД статистически значимо ниже соответствующего показателя min-выборки по всем рассматриваемым источникам. То же верно для большинства взятых по отдельности СД в сопоставлении с данными по ПГ, в которые они входят: по МАС – в 18 из 22 групп, по МАС+НКРЯ – в 17, по МАС+Рунет – в 12. Противоположная же тенденция наблюдается только для одного СД (репетитивный с *пере-*) по данным МАС и МАС+Рунет (см. Приложение В).

5.2 Глаголы на *-и(зи)рова-*

Глаголы на *-и(зи)рова-* (*про-информ-иров-ыва-ть*^{НСВ2} ← *про-информ-ирова-ть*^{СВ}) – самое малочисленное подмножество, специфичное в отношении суффиксальной имперфективации. Специализированные выборки такого типа выделяются в составе 13 из 21 ПГ, их доля в *max*-выборках незначительна: от 0,2–0,3% для *у-*, *в(о)-* и *вы-* глаголов до 4,71% для *с(о)-* глаголов.

В Табл. 9 средние показатели рассчитаны по 13 ПГ с ненулевым количеством глаголов данного множества.

	в МАС	% в МАС	в НКРЯ	% в НКРЯ	в Рунете	% в Рунете	% в <i>max</i> -выборке
<i>в(о)з(о)-/вос-</i>	0	0	0	0	0	0	0
<i>до-</i>	0	0	0	0	0	0	0
<i>из(о)-/ис-</i>	0	0	0	0	0	0	0
<i>над(о)-</i>	0	0	0	0	0	0	0
<i>недо-</i>	0	0	0	0	0	0	0
<i>пре-</i>	0	0	0	0	0	0	0
<i>пред(о)-</i>	0	0	0	0	0	0	0
<i>при-</i>	0	0	0	0	0	0	0
<i>по-</i>	0	0	0	0	2	16,67	0,69
<i>про-</i>	0	0	1	2,5	7	17,5	2,84
<i>с(о)-</i>	5	6,67	9	12	31	41,33	4,71
<i>на-</i>	4	25	4	25	9	56,25	1,09
<i>о-/об(о)-</i>	3	42,86	4	57,14	4	57,14	0,48
<i>за-</i>	12	18,18	13	19,7	45	68,18	3,15
<i>от(о)-</i>	5	20	7	28	24	96	2,3
<i>пере-</i>	13	40,63	15	46,88	31	96,88	2,58
<i>в(о)-</i>	0	0	0	0	1	100	0,27
<i>вы-</i>	0	0	1	33,33	3	100	0,3
<i>у-</i>	0	0	0	0	2	100	0,2
<i>раз(о)-/рас-</i>	6	35,29	7	41,18	17	100	1,16
<i>под(о)-</i>	3	50	3	50	6	100	0,85
Среднее	3,92	18,36	4,92	24,29	14,00	73,07	1,59
Медиана	3	18,18	4	25	7	96	1,09

Таблица 9: Имперфективируемость перфективов на *-и(зи)рова-* (сортировка по возрастанию в столбце «% в Рунете»)

Сравнение средних показателей ИМПФ глаголов на *-и(зи)рова-* с данными по *min*-выборкам в 13 ПГ (см. Табл. 10) позволяет сделать вывод: в целом уровень ИМПФ глаголов на *-и(зи)рова-* понижен. При этом для нескольких ПГ резкое возрастание уровня ИМПФ при привлечении данных НКРЯ и Рунета связано скорее с малым объемом данных групп, поэтому к полученной картине следует относиться с осторожностью.

	в МАС	% в МАС	в НКРЯ	% в НКРЯ	в Рунете	% в Рунете
<i>по-</i>	253	31,12	354	43,54	561	69
<i>на-</i>	549	70,03	603	76,91	697	88,9
<i>с(о)-</i>	1040	75,2	1094	79,1	1257	90,89
<i>про-</i>	629	68,82	708	77,46	839	91,79
<i>за-</i>	1078	76,95	1177	84,01	1299	92,72
<i>вы-</i>	777	82,92	849	90,61	889	94,88
<i>пере-</i>	415	81,21	454	88,85	485	94,91
<i>от(о)-</i>	692	85,64	734	90,84	770	95,3
<i>о(б)/об(о)-</i>	1073	80,86	1158	87,26	1267	95,48
<i>у-</i>	778	85,87	801	88,41	875	96,58
<i>в(о)-</i>	289	92,63	301	96,47	305	97,76
<i>под(о)-</i>	406	95,53	418	98,35	418	98,35
<i>раз(о)-/рас-</i>	877	89,86	926	94,88	962	98,57
Среднее	681,23	78,20	736,69	84,36	817,23	92,70
Медиана	692	81,21	734	88,41	839	94,91

Таблица 10: Имперфективируемость ПГ (min-выборки ПГ, содержащих глаголы на *-и(зи)рова-*, сортировка по возрастанию в столбце «% в Рунете»)

Табл. 11 подтверждает, что суммарный уровень ИМПФ глаголов на *-и(зи)рова-* оказывается существенно ниже (различие статистически значимо) показателей по min-выборкам (на основании 13 ПГ, включающих глаголы на *-и(зи)рова-*). Показательно также, что при рассмотрении данных по отдельным ПГ не фиксируется ни одного случая значимого превышения уровня ИМПФ глаголов на *-и(зи)рова-* в сопоставлении с min-выборками (см. Приложение Г).

	Есть НСВ2	Нет НСВ2	Всего	доля НСВ2
МАС				
<i>-и(зи)рова-</i>	51	251	302	16,89%
Min-выборки	8856	2641	11497	77,03%
Всего	8907	2892	11799	
Различие статистически значимо: $\chi^2=572$, $p < 0,01$				
МАС+НКРЯ				
<i>-и(зи)рова-</i>	64	238	302	21,19%
Min-выборки	9577	1920	11497	83,30%
Всего	9641	2158	11799	
Различие статистически значимо: $\chi^2=755,402$, $p < 0,01$				
МАС+Рунет				
<i>-и(зи)рова-</i>	182	120	302	60,26%
Min-выборки	10624	873	11497	92,41%
Всего	10806	993	11799	
Различие статистически значимо: $\chi^2=390,265$, $p < 0,01$				

Таблица 11: Имперфективируемость глаголов на *-и(зи)рова-* в сравнении с min-выборками (по 13 ПГ)

6 Итоги

Итак, русские приставочные перфективы, показывая в среднем уровень ИМПФ, достаточный для квалификации суффиксальной имперфективации как регулярного словоизменительного механизма (в понимании [9: 25–26]), представляют собой набор разного рода подмножеств с различной склонностью к имперфективации. Среди этих подмножеств есть те, которые демонстрируют повышенный уровень ИМПФ, и те, которые отличаются сниженным уровнем. К первым относятся отыменные и, в меньшей степени, отперфективные перфективы, к последним – большинство СД и все глаголы с *-и(зи)рова-*. Первые две группы находятся на максимальном полюсе шкалы ИМПФ, последние – на минимальном.

Объяснением повышенной ИМПФ ОИП и ОПП может служить отсутствие в деривационной истории получаемых путем регулярной имперфективации НСВ₂ стадии НСВ₁, т.е. отсутствие глагола-симплекса, который мог бы образовать функциональный видовой противочлен для приставочного перфектива: у *прошляпнуть*^{СВ} нет **ишляпнуть*^{НСВ}, соответственно при необходимости образовать НСВ возможен только путь через суффиксальный имперфектив *прошляпливать* (то же для *выдать*^{СВ}: только *выдавать*^{НСВ}, потому что симплекс – перфектив: *дать*^{СВ}). Таким образом, высокий уровень имперфективации этих подмножеств является следствием давления системы.

Низкий уровень ИМПФ приставочных перфективов с *-и(зи)рова-* имеет, с высокой степенью вероятности, морфонологическую природу: глаголы с многосложным формантом *-и(зи)рова-* в основе неохотно присоединяют суффикс имперфективации в дополнение к этому сегменту (ср. зафиксированные формы типа *про-информ-иров-ыва-ть*). Наименее ясна некоторая резистентность к суффиксальной имперфективации глаголов ряда СД. Здесь необходимы дополнительные изыскания.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-012-00143 «Суффиксальная имперфективация приставочных перфективов в современном русском языке». Авторы признательны Софье Шарыгиной – за участие в сборе и первичном анализе языковых данных, и Екатерине Еникеевой – за консультации в области их статистической обработки, а также за создание открытой базы данных проекта, доступной по ссылке: <https://www.rusimpdb.ru>.

References

- [1] Chuikova Oksana Iu. (2020), On the secondary imperfectivization of *po*-perfectives in Russian [Ob osobennostyakh vtorichnoj imperfektivatsii glagolov s prefiksom *po*- v russkom yazyke], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”, Vol. 19 (26), pp. 160–176.
- [2] Gorbova Elena V. (2015), Aspectual formation of the Russian verb: Prefication or suffixation? [Vidoobrazovanie russkogo glagola: prefiksaciya i/ili suffiksaciya?], Topics in the study of language [Voprosy jazykoznanija], Vol. 1, pp. 7–37.
- [3] Evgenieva Anastasija P. (ed.) (1981–1984), Dictionary of Russian Language in 4 volumes [Slovar' russkogo jazyka v 4 tomah], 2nd ed. Academy of Science of USSR, Institute of Russian Language, Moscow. Access mode: <http://www.slovari.ru/default.aspx?s=0&p=240>.
- [4] Kuznecova Ariadna I., Efremova Tatjana F. Dictionary of Russian morphemes [Slovar' morfem russkogo jazyka]. — Moscow: Russian Language [Russkij yazyk], 1986.
- [5] Shelyakin Mikhail A. Category of aspectuality of the Russian verb [Kategoriya aspektual'nosti russkogo glagola], — Moscow: URSS Publ., 2008.
- [6] Tatevosov Sergey G. Actionality in lexicon and grammar. The verb and the event structure [AktSIONal'nost' v leksike i grammatike. Glagol i struktura sobytiya]. — Moscow: Languages of Slavic Cultures [Yazyki Slavyanskikh Kul'tur], 2015.
- [7] Xrakovskij Viktor S. (2018), Aspect's “sore points” revisited once more. Topics in the study of language [Voprosy jazykoznanija], Vol. 1, pp. 105–118.
- [8] 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], 2000.
- [9] Zaliznyak Andrey A. «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 yazykoznaniju], — Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul'tury], 2002.

Приложение А. Соотношение количества имперфективируемых и неимперфективируемых лексем в ОИП и в min-выборках⁷

МАС	ОИП				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	24	0	24	100%	289	23	312	92,63%	7,37	
в(о)з(о)-/в(о)с-	1	1	2	50%	250	128	378	66,14%	-16,14	
вы-	8	0	8	100%	777	160	937	82,92%	17,08	
за-	42	18	60	70%	1078	323	1401	76,95%	-6,95	
из(о)-/ис-	9	0	9	100%	357	164	521	68,52%	31,48	
на-	8	1	9	88,89%	549	235	784	70,03%	18,86	
над(о)-	1	0	1	100%	39	6	45	86,67%	13,33	
о(б)-/об(о)-	280	63	343	81,63%	1073	254	1327	80,86%	0,77	
от(о)-	19	0	19	100%	692	116	808	85,64%	14,36	
пере-	10	0	10	100,00%	415	96	511	81,21%	18,79	
по-	28	4	32	87,50%	253	560	813	31,12%	56,38	Фишера, p<0,0001
под(о)-	6	3	9	66,67%	406	19	425	95,53%	-28,86	Фишера, p=0,0078
пре-	16	0	16	100%	46	1	47	97,87%	2,13	
при-	32	5	37	86,49%	544	54	598	90,97%	-4,48	
про-	20	7	27	74,07%	629	285	914	68,82%	5,26	
раз(о)-/рас-	27	1	28	96,43%	877	99	976	89,86%	6,57	
с(о)-	46	12	58	79,31%	1040	343	1383	75,20%	4,11	
у-	109	7	116	93,97%	778	128	906	85,87%	8,09	χ ² =5,191, p=0,0227
Всего	686	122	808	84,90%	10092	2994	13086	77,12%	7,78	χ²=26,034, p<0,0001
Среднее	38,11	6,78	44,89	87,50%	560,67	166,33	727	79,27%	8,23	
Медиана	19,5	1	21,5	91,43%	546,5	128	796	82,07%	9,36	

МАС+НКРЯ	ОИП				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	24	0	24	100%	301	11	312	96,47%	3,53	
в(о)з(о)-/в(о)с-	1	1	2	50%	288	90	378	76,19%	-26,19	
вы-	8	0	8	100%	849	88	937	90,61%	9,39	
за-	44	16	60	73,33%	1177	224	1401	84,01%	-10,68	χ ² =4,032, p=0,0446
из(о)-/ис-	9	0	9	100%	399	122	521	76,58%	23,42	
на-	8	1	9	88,89%	603	181	784	76,91%	11,98	
над(о)-	1	0	1	100%	45	0	45	100,00%	0,00	
о(б)-/об(о)-	296	47	343	86,30%	1158	169	1327	87,26%	-0,97	
от(о)-	19	0	19	100%	734	74	808	90,84%	9,16	
пере-	10	0	10	100%	454	57	511	88,85%	11,15	
по-	28	4	32	87,50%	354	459	813	43,54%	43,96	Фишера, p<0,0001
под(о)-	8	1	9	88,89%	418	7	425	98,35%	-9,46	
пре-	16	0	16	100%	47	0	47	100,00%	0,00	
при-	35	2	37	94,59%	571	27	598	95,48%	-0,89	
про-	22	5	27	81,48%	708	206	914	77,46%	4,02	
раз(о)-/рас-	27	1	28	96,43%	926	50	976	94,88%	1,55	
с(о)-	47	11	58	81,03%	1094	289	1383	79,10%	1,93	
у-	112	4	116	96,55%	801	105	906	88,41%	8,14	Фишера, p=0,0059
Всего	715	93	808	88,49%	10927	2159	13086	83,50%	4,99	χ²=13,580, p=0,0002
Среднее	39,72	5,17	44,89	90,28%	607,06	119,94	727	85,83%	4,45	
Медиана	20,5	1	21,5	95,51%	587	89	796	88,63%	6,88	

МАС+Runet	ОИП				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	24	0	24	100%	305	7	312	97,76%	2,24	
в(о)з(о)-/в(о)с-	2	0	2	100%	344	34	378	91,01%	8,99	
вы-	8	0	8	100%	889	48	937	94,88%	5,12	
за-	56	4	60	93,33%	1299	102	1401	92,72%	0,61	
из(о)-/ис-	9	0	9	100%	450	71	521	86,37%	13,63	
на-	8	1	9	88,89%	697	87	784	88,90%	-0,01	
над(о)-	1	0	1	100%	44	1	45	97,78%	2,22	
о(б)-/об(о)-	325	18	343	94,75%	1267	60	1327	95,48%	-0,73	
от(о)-	19	0	19	100%	770	38	808	95,30%	4,70	
пере-	10	0	10	100%	485	26	511	94,91%	5,09	
по-	29	3	32	90,63%	561	252	813	69%	21,62	Фишера, p=0,0094
под(о)-	9	0	9	100%	418	7	425	98,35%	1,65	
пре-	16	0	16	100%	47	0	47	100%	0,00	
при-	37	0	37	100%	593	5	598	99,16%	0,84	
про-	24	3	27	88,89%	839	75	914	91,79%	-2,91	
раз(о)-/рас-	28	0	28	100%	962	14	976	98,57%	1,43	
с(о)-	56	2	58	96,55%	1257	126	1383	90,89%	5,66	
у-	112	4	116	96,55%	875	31	906	96,58%	-0,03	
Всего	773	35	808	95,67%	12102	984	13086	92,48%	3,19	χ²=10,915, p=0,001
Среднее	42,94	1,94	44,89	97,20%	672,33	54,67	727	93,30%	3,90	
Медиана	21,5	0	21,5	100%	645	36	796	95,10%	4,90	

⁷ В Приложениях А–Г приводятся данные о распределении имперфективируемых (демонстрирующих способность к суффиксальной имперфективации) и неимперфективируемых лексем в каждой из рассматриваемых специализированных выборок и в min-выборках для каждой ПГ в отдельности и для совокупных данных по всем ПГ, включающим в свой состав то или иное специализированное подмножество. В последнем столбце таблиц приводятся результаты применения критерия согласия Пирсона (хи-квадрат с поправкой Йетса, двусторонний вариант) либо точного критерия Фишера (двусторонний вариант; используется, если хотя бы одно из значений не превышает 5). Значения критериев даны только при наличии статистической значимости (при $p < 0,05$). Зеленая заливка используется для выделения случаев, которые соответствуют гипотезе о повышенном (для ОИП и ОПП) или сниженном (для СД и глаголов на *-и(зи)рова-*) уровне ИМПФ специализированных подмножеств в сопоставлении с данными по min-выборкам, а красная – для случаев, противоречащих предсказываемым гипотезой результатам.

Приложение Б. Соотношение количества имперфективируемых и неимперфективируемых лексем в ОПП и в min-выборках

МАС	ОПП				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	49	3	52	94,23%	289	23	312	92,63%	1,60	
в(о)з(о)-/в(о)с- вы-	72	21	93	77,42%	250	128	378	66,14%	11,28	χ ² =3,886, p=0,0487 Фишера, p=0,0105
до-	74	5	79	93,67%	777	160	937	82,92%	10,75	
за-	13	2	15	86,67%	96	7	103	93,20%	-6,54	
из(о)-/ис-	95	10	105	90,48%	1078	323	1401	76,95%	13,53	χ ² =9,614, p=0,0019
на-	12	4	16	75,00%	357	164	521	68,52%	6,48	
над(о)-	67	9	76	88,16%	549	235	784	70,03%	18,13	χ ² =10,335, p=0,0013 Фишера, p=0,0259
недо-	0	2	2	0,00%	39	6	45	86,67%	-8,67	
недо- о(б)-/об(о)- от(о)-	8	1	9	88,89%	37	5	42	88,10%	0,79	
пере-	84	17	101	83,17%	1073	254	1327	80,86%	2,31	
по-	98	12	110	89,09%	692	116	808	85,64%	3,45	
под(о)-	188	16	204	92,16%	415	96	511	81,21%	10,94	χ ² =12,402, p=0,0004
пре-	23	164	187	12,30%	253	560	813	31,12%	-18,82	χ ² =26,014, p<0,0001 χ ² =25,341, p<0,0001
пред(о)-	106	24	130	81,54%	406	19	425	95,53%	-13,99	
при-	22	2	24	91,67%	46	1	47	97,87%	-6,21	
про-	34	1	35	97,14%	24	0	24	100%	-2,86	
раз(о)-/рас- с(о)-	144	32	176	81,82%	544	54	598	90,97%	-9,15	χ ² =10,623, p=0,0011
у-	96	8	104	92,31%	629	285	914	68,82%	23,49	χ ² =24,001, p<0,0001
Всего	112	19	131	85,50%	877	99	976	89,86%	-4,36	
Среднее	117	17	134	87,31%	1040	343	1383	75,20%	12,11	χ ² =9,248, p=0,0024
Медиана	49	11	60	81,67%	778	128	906	85,87%	-4,21	
Всего	1463	380	1843	79,38%	10249	3006	13255	77,32%	2,06	χ²=3,828, p=0,0504
Среднее	69,67	18,10	87,76	79,53%	488,05	143,14	631,19	81,34%	-1,81	
Медиана	72	10	93	87,31%	415	116	598	85,64%	1,67	

МАС+НКРЯ	ОПП				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	50	2	52	96,15%	301	11	312	96,47%	-0,32	
в(о)з(о)-/в(о)с- вы-	77	16	93	82,80%	288	90	378	76,19%	6,61	
до-	76	3	79	96,20%	849	88	937	90,61%	5,59	
за-	13	2	15	86,67%	97	6	103	94,17%	-7,51	
из(о)-/ис-	99	6	105	94,29%	1177	224	1401	84,01%	10,27	χ ² =7,194, p=0,0073
на-	13	3	16	81,25%	399	122	521	76,58%	4,67	
над(о)-	70	6	76	92,11%	603	181	784	76,91%	15,19	χ ² =8,526, p=0,0035 Фишера, p=0,0426
недо-	1	1	2	50,00%	45	0	45	100,00%	-50,00	
недо- о(б)-/об(о)- от(о)-	9	0	9	100,00%	41	1	42	97,62%	2,38	
пере-	86	15	101	85,15%	1158	169	1327	87,26%	-2,12	
по-	103	7	110	93,64%	734	74	808	90,84%	2,79	
под(о)-	197	7	204	96,57%	454	57	511	88,85%	7,72	χ ² =9,744, p=0,0018
пре-	52	135	187	27,81%	354	459	813	43,54%	-15,73	χ ² =14,962, p<0,0001 χ ² =23,053, p<0,0001
пред(о)-	115	15	130	88,46%	418	7	425	98,35%	-9,89	
при-	22	2	24	91,67%	47	0	47	100%	-8,33	
про-	35	0	35	100%	24	0	24	100%	0,00	
раз(о)-/рас- с(о)-	157	19	176	89,20%	571	27	598	95,48%	-6,28	χ ² =8,504, p=0,0035
у-	100	4	104	96,15%	708	206	914	77,46%	18,69	Фишера, p<0,0001
Всего	119	12	131	90,84%	926	50	976	94,88%	-4,04	
Среднее	122	12	134	91,04%	1094	289	1383	79,10%	11,94	χ ² =10,215, p=0,0014
Медиана	53	7	60	88,33%	801	105	906	88,41%	-0,08	
Всего	1569	274	1843	85,13%	11089	2166	13255	83,66%	1,47	χ²=2,487, p=0,1148
Среднее	74,71	13,05	87,76	86,59%	528,05	103,14	631,19	87,46%	-0,88	
Медиана	76	6	93	91,04%	454	74	598	90,61%	0,44	

МАС+Рунет	ОПП				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	50	2	52	96,15%	305	7	312	97,76%	-1,60	
в(о)з(о)-/в(о)с- вы-	85	8	93	91,40%	344	34	378	91,01%	0,39	
до-	78	1	79	98,73%	889	48	937	94,88%	3,86	
за-	14	1	15	93,33%	103	0	103	100%	-6,67	
из(о)-/ис-	103	2	105	98,10%	1299	102	1401	92,72%	5,38	Фишера, p=0,0425
на-	14	2	16	87,5%	453	71	524	86,45%	1,05	
над(о)-	71	5	76	93,42%	697	87	784	88,90%	4,52	
недо-	2	0	2	100%	44	1	45	97,78%	2,22	
недо- о(б)-/об(о)- от(о)-	9	0	9	100%	42	0	42	100%	0,00	
пере-	90	11	101	89,11%	1267	60	1327	95,48%	-6,37	χ ² =6,768, p=0,0093
по-	102	8	110	92,73%	770	38	808	95,30%	-2,57	
под(о)-	203	1	204	99,51%	485	26	511	94,91%	4,60	Фишера, p=0,0018
пре-	93	94	187	49,73%	561	252	813	69%	-19,27	χ ² =23,346, p<0,0001
пред(о)-	124	6	130	95,38%	418	7	425	98,35%	-2,97	
при-	23	1	24	95,83%	47	0	47	100%	-4,17	
про-	35	0	35	100%	24	0	24	100%	0,00	
раз(о)-/рас- с(о)-	172	4	176	97,73%	593	5	598	99,16%	-1,44	
у-	103	1	104	99,04%	839	75	914	91,79%	7,24	Фишера, p=0,0047
Всего	127	4	131	96,95%	962	14	976	98,57%	-1,62	
Среднее	124	10	134	92,54%	1257	126	1383	90,89%	1,65	
Медиана	57	3	60	95%	875	31	906	96,58%	-1,58	
Всего	1679	164	1843	91,10%	12274	984	13258	92,58%	-1,48	χ²=4,815, p=0,0282
Среднее	79,95	7,81	87,76	93,44%	584,48	46,86	631,33	94,26%	-0,83	
Медиана	85	2	93	95,83%	561	31	598	95,48%	0,35	

МАС	ОПП без по-				Мин-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	49	3	52	94,23%	289	23	312	92,63%	1,60	
в(о)з(о)-/в(о)с-	72	21	93	77,42%	250	128	378	66,14%	11,28	χ ² =3,886, p=0,0487
вы-	74	5	79	93,67%	777	160	937	82,92%	10,75	Фишера, p=0,0105
до-	13	2	15	86,67%	96	7	103	93,20%	-6,54	
за-	95	10	105	90,48%	1078	323	1401	76,95%	13,53	χ ² =9,614, p=0,0019
из(о)-/ис-	12	4	16	75,00%	357	164	521	68,52%	6,48	
на-	67	9	76	88,16%	549	235	784	70,03%	18,13	χ ² =10,335, p=0,0013
над(о)-	0	2	2	0,00%	39	6	45	86,67%	-86,67	Фишера, p=0,0259
недо-	8	1	9	88,89%	37	5	42	88,10%	0,79	
о(б)-/об(о)-	84	17	101	83,17%	1073	254	1327	80,86%	2,31	
от(о)-	98	12	110	89,09%	692	116	808	85,64%	3,45	
пере-	188	16	204	92,16%	415	96	511	81,21%	10,94	χ ² =12,402, p=0,0004
под(о)-	106	24	130	81,54%	406	19	425	95,53%	-13,99	χ ² =25,341, p<0,0001
пре-	22	2	24	91,67%	46	1	47	97,87%	-6,21	
пред(о)-	34	1	35	97,14%	24	0	24	100%	-2,86	
при-	144	32	176	81,82%	544	54	598	90,97%	-9,15	χ ² =10,623, p=0,0011
про-	96	8	104	92,31%	629	285	914	68,82%	23,49	χ ² =24,001, p<0,0001
раз(о)-/рас-	112	19	131	85,50%	877	99	976	89,86%	-4,36	
с(о)-	117	17	134	87,31%	1040	343	1383	75,20%	12,11	χ ² =9,248, p=0,0024
у-	49	11	60	81,67%	778	128	906	85,87%	-4,21	
Всего	1440	216	1656	86,96%	9996	2446	12442	80,34%	6,62	χ²=41,331, p<0,0001
Среднее	72,00	10,80	82,80	82,89%	499,80	122,30	622,10	83,85%	-0,96	
Медиана	73	9,5	86	87,74%	480	108	560	85,76%	1,98	

МАС+НКРЯ	ОПП без по-				Мин-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	50	2	52	96,15%	301	11	312	96,47%	-0,32	
в(о)з(о)-/в(о)с-	77	16	93	82,80%	288	90	378	76,19%	6,61	
вы-	76	3	79	96,20%	849	88	937	90,61%	5,59	
до-	13	2	15	86,67%	97	6	103	94,17%	-7,51	
за-	99	6	105	94,29%	1177	224	1401	84,01%	10,27	χ ² =7,194, p=0,0073
из(о)-/ис-	13	3	16	81,25%	399	122	521	76,58%	4,67	
на-	70	6	76	92,11%	603	181	784	76,91%	15,19	χ ² =8,526, p=0,0035
над(о)-	1	1	2	50,00%	45	0	45	100,00%	-50,00	Фишера, p=0,0426
недо-	9	0	9	100,00%	41	1	42	97,62%	2,38	
о(б)-/об(о)-	86	15	101	85,15%	1158	169	1327	87,26%	-2,12	
от(о)-	103	7	110	93,64%	734	74	808	90,84%	2,79	
пере-	197	7	204	96,57%	454	57	511	88,85%	7,72	χ ² =9,744, p=0,0018
под(о)-	115	15	130	88,46%	418	7	425	98,35%	-9,89	χ ² =23,053, p<0,0001
пре-	22	2	24	91,67%	47	0	47	100%	-8,33	
пред(о)-	35	0	35	100%	24	0	24	100%	0,00	
при-	157	19	176	89,20%	571	27	598	95,48%	-6,28	χ ² =8,504, p=0,0035
про-	100	4	104	96,15%	708	206	914	77,46%	18,69	Фишера, p<0,0001
раз(о)-/рас-	119	12	131	90,84%	926	50	976	94,88%	-4,04	
с(о)-	122	12	134	91,04%	1094	289	1383	79,10%	11,94	χ ² =10,215, p=0,0014
у-	53	7	60	88,33%	801	105	906	88,41%	-0,08	
Всего	1517	139	1656	91,61%	10735	1707	12442	86,28%	5,33	χ²=35,964, p<0,0001
Среднее	75,85	6,95	82,80	89,53%	536,75	85,35	622,10	89,66%	-0,13	
Медиана	76,5	6	86	91,36%	513	66	560	90,72%	0,63	

МАС+Runet	ОПП без по-				Мин-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	50	2	52	96,15%	305	7	312	97,76%	-1,60	
в(о)з(о)-/в(о)с-	85	8	93	91,40%	344	34	378	91,01%	0,39	
вы-	78	1	79	98,73%	889	48	937	94,88%	3,86	
до-	14	1	15	93,33%	103	0	103	100%	-6,67	
за-	103	2	105	98,10%	1299	102	1401	92,72%	5,38	Фишера, p=0,0425
из(о)-/ис-	14	2	16	87,5%	450	71	521	86,37%	1,13	
на-	71	5	76	93,42%	697	87	784	88,90%	4,52	
над(о)-	2	0	2	100%	44	1	45	97,78%	2,22	
недо-	9	0	9	100%	42	0	42	100%	0,00	
о(б)-/об(о)-	90	11	101	89,11%	1267	60	1327	95,48%	-6,37	χ ² =6,768, p=0,0093
от(о)-	102	8	110	92,73%	770	38	808	95,30%	-2,57	
пере-	203	1	204	99,51%	485	26	511	94,91%	4,60	Фишера, p=0,0018
под(о)-	124	6	130	95,38%	418	7	425	98,35%	-2,97	
пре-	23	1	24	95,83%	47	0	47	100%	-4,17	
пред(о)-	35	0	35	100%	24	0	24	100%	0,00	
при-	172	4	176	97,73%	593	5	598	99,16%	-1,44	
про-	103	1	104	99,04%	839	75	914	91,79%	7,24	Фишера, p=0,0047
раз(о)-/рас-	127	4	131	96,95%	962	14	976	98,57%	-1,62	
с(о)-	124	10	134	92,54%	1257	126	1383	90,89%	1,65	
у-	57	3	60	95%	875	31	906	96,58%	-1,58	
Всего	1586	70	1656	95,77%	11710	732	12442	94,12%	1,66	χ²=7,167, p=0,0074
Среднее	79,30	3,50	82,80	95,62%	585,50	36,60	622,10	95,52%	0,10	
Медиана	81,5	2	86	95,99%	539	29	560	96,03%	-0,03	

Приложение В. Соотношение количества имперфективируемых и неимперфективируемых лексем в группе СД и в min-выборках

MAC	СД				ПП	Min-выборки				РАЗНИЦА	
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		Есть НСВ2	Нет НСВ2	Всего	% в НСВ2	%	Φ/Фишера
СД с формальными показателями											
в(о)и(о)-в(о)е-: делимитативный	0	7	7	0%	в(о)и(о)-в(о)е-	250	128	378	66,14%	-66,14	Фишера, p=0,0006
до-: комплетивный	229	20	249	91,97%	до-	96	7	103	93,20%	-1,24	
до...-св: дуративно-негативный	21	18	39	53,85%	до-	96	7	103	93,20%	-39,36	χ²=27,556, p<0,0001
за-: начинательный	11	469	480	2,29%	за-	1078	323	1401	76,95%	-74,65	χ²=814,301, p<0,0001
за...-св: чрезмерно-длительный	26	26	52	50%	за-	1078	323	1401	76,95%	-26,95	χ²=18,498, p<0,0001
нз...-св: чрезмерно-кратный	8	20	28	28,57%	нз-	357	164	521	68,52%	-39,95	χ²=17,282, p<0,0001
на-: кумулятивный	257	158	415	61,93%	на-	549	235	784	70,03%	-8,10	χ²=7,713, p=0,0055
на...-св: сатуративный/сативный	38	147	185	20,54%	на-	549	235	784	70,03%	-49,48	χ²=151,417, p<0,0001
о(о)и(о)-: дистрибутивно-суммарный	24	7	31	77,42%	о(о)и(о)-	1073	254	1327	80,86%	-3,44	
от(о)-: финитивный	0	145	145	0%	от(о)-	692	116	808	85,64%	-85,64	Фишера, p<0,0001
пере-: дистрибутивный	174	150	324	53,70%	пере-	415	96	511	81,21%	-27,51	χ²=706888, p<0,0001
пере-: репетитивный	235	33	268	87,69%	пере-	415	96	511	81,21%	6,47	χ²=4,873, p=0,0273
по-: аттенуативный	0	62	62	0%	по-	253	560	813	31,12%	-31,12	Фишера, p<0,0001
по-: делимитативный	117	454	571	20,49%	по-	253	560	813	31,12%	-10,63	χ²=18,808, p<0,0001
по-: дистрибутивный	0	168	168	0%	по-	253	560	813	31,12%	-31,12	Фишера, p<0,0001
по-: ингрессивный	0	69	69	0%	по-	253	560	813	31,12%	-31,12	Фишера, p<0,0001
под-: аттенуативный	149	33	182	81,87%	под-	406	19	425	95,53%	-13,66	χ²=28,643, p<0,0001
при-: аттенуативный	72	21	93	77,42%	при-	544	54	598	90,97%	-13,55	χ²=13,905, p=0,0002
про-: пердуративный	16	333	349	4,58%	про-	629	285	914	68,82%	-64,23	χ²=414,451, p<0,0001
раз(о)/рас-: дистрибутивно-суммарный	141	16	157	89,81%	раз(о)/рас-	877	99	976	89,86%	-0,05	
раз(о)/рас...(-св): ингрессивно-интенсивный	74	131	205	36,10%	раз(о)/рас-	877	99	976	89,86%	-53,76	χ²=308,792, p<0,0001
у...-св: чрезмерно-интенсивный	8	10	18	44,44%	у-	778	128	906	85,87%	-41,43	χ²=20,693, p<0,0001
Всего	1600	2497	4097	39,05%	Сумма	7997	2468	10465	76,42%	-37,36	χ²=1827,576, p<0,0001
Среднее	72,73	113,50	186,23	40,12%	Среднее:	571,21	176,29	747,50	77,48%	-37,36	
Медиана	25	47,5	162,5	40,27%	Медиана:	479,5	146	810,5	81,04%	-40,77	

MAC+НКРЯ	СД				ПП	Min-выборки				РАЗНИЦА	
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		Есть НСВ2	Нет НСВ2	Всего	% НСВ2	%	Φ/Фишера
СД с формальными показателями											
в(о)и(о)-в(о)е-: делимитативный	2	5	7	28,57%	в(о)и(о)-в(о)е-	288	90	378	76,19%	-47,62	Фишера, p=0,0114
до-: комплетивный	234	15	249	93,98%	до-	97	6	103	94,17%	-0,20	
до...-св: дуративно-негативный	22	17	39	56,41%	до-	97	6	103	94,17%	-37,76	χ²=24,151, p<0,0001
за-: начинательный	43	437	480	8,96%	за-	1177	224	1401	84,01%	-75,05	χ²=880,286, p<0,0001
за...-св: чрезмерно-длительный	30	22	52	57,69%	за-	1177	224	1401	84,01%	-26,32	χ²=22,859, p<0,0001
нз...-св: чрезмерно-кратный	10	18	28	35,71%	нз(о)-/нз-	399	122	521	76,58%	-40,87	χ²=21,260, p<0,0001
на-: кумулятивный	280	135	415	67,47%	на-	603	181	784	76,91%	-9,44	χ²=11,986, p=0,0005
на...-св: сатуративный/сативный	54	131	185	29,19%	на-	603	181	784	76,91%	-47,72	χ²=153,980, p<0,0001
о(о)и(о)-: дистрибутивно-суммарный	29	2	31	93,55%	о(о)и(о)-	1158	169	1327	87,26%	6,28	
от(о)-: финитивный	25	120	145	17,24%	от(о)-	734	74	808	90,84%	-73,60	χ²=406,232, p<0,0001
пере-: дистрибутивный	221	103	324	68,21%	пере-	454	57	511	88,85%	-20,64	χ²=53,184, p<0,0001
пере-: репетитивный	245	23	268	91,42%	пере-	454	57	511	88,85%	2,57	
по-: аттенуативный	11	51	62	17,74%	по-	354	459	813	43,54%	-25,80	χ²=14,729, p<0,0001
по-: делимитативный	248	323	571	43,43%	по-	354	459	813	43,54%	-0,11	
по-: дистрибутивный	26	142	168	15,48%	по-	354	459	813	43,54%	-28,07	χ²=45,040, p<0,0001
по-: ингрессивный	9	60	69	13,04%	по-	354	459	813	43,54%	-30,50	χ²=23,186, p<0,0001
под-: аттенуативный	160	22	182	87,91%	под(о)-	418	7	425	98,35%	-10,44	χ²=28,283, p<0,0001
при-: аттенуативный	78	15	93	83,87%	при-	571	27	598	95,48%	-11,61	χ²=17,036, p<0,0001
про-: пердуративный	81	268	349	23,21%	про-	708	206	914	77,46%	-54,25	χ²=314,764, p<0,0001
раз(о)/рас-: дистрибутивно-суммарный	150	7	157	95,54%	раз(о)/рас-	926	50	976	94,88%	0,66	
раз(о)/рас...(-св): ингрессивно-интенсивный	116	89	205	56,59%	раз(о)/рас-	926	50	976	94,88%	-38,29	χ²=235,537, p<0,0001
у...-св: чрезмерно-интенсивный	9	9	18	50%	у-	801	105	906	88,41%	-38,41	χ²=20,656, p<0,0001
Всего	2083	2014	4097	50,84%	Всего	8688	1777	10465	83,02%	-32,18	χ²=1581,477, p<0,0001
Среднее	94,68	91,55	186,23	51,60%	Среднее:	620,57	126,93	747,50	83,78%	-32,18	
Медиана	48,50	37,00	162,50	53,21%	Медиана:	587,00	97,50	796,00	87,84%	-34,63	

MAC+Рунет	СД				ПП	Min-выборки				РАЗНИЦА	
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		Есть НСВ2	Нет НСВ2	Всего	% НСВ2	%	Φ/Фишера
СД с формальными показателями											
в(о)и(о)-в(о)е-: делимитативный	5	2	7	71,43%	в(о)и(о)-в(о)е-	344	34	378	91,01%	-19,58	
до-: комплетивный	244	5	249	97,99%	до-	103	0	103	100,00%	-2,01	
до...-св: дуративно-негативный	37	2	39	94,87%	до-	103	0	103	100,00%	-5,13	
за-: начинательный	190	290	480	39,58%	за-	1299	102	1401	92,72%	-53,14	χ²=608,665, p<0,0001
за...-св: чрезмерно-длительный	44	8	52	84,62%	за-	1299	102	1401	92,72%	-8,10	
нз...-св: чрезмерно-кратный	16	12	28	57,14%	нз(о)-/нз-	450	71	521	86,37%	-29,23	χ²=15,486, p<0,0001
на-: кумулятивный	347	68	415	83,61%	на-	697	87	784	88,90%	-5,29	χ²=6,281, p=0,0122
на...-св: сатуративный/сативный	130	55	185	70,27%	на-	697	87	784	88,90%	-18,63	χ²=40,074, p<0,0001
о(о)и(о)-: дистрибутивно-суммарный	30	1	31	96,77%	о(о)и(о)-	1267	60	1327	95,48%	1,30	
от(о)-: финитивный	47	98	145	32,41%	от(о)-	770	38	808	95,30%	-62,88	χ²=392,234, p<0,0001
пере-: дистрибутивный	274	50	324	84,57%	пере-	485	26	511	94,91%	-10,34	χ²=24,408, p<0,0001
пере-: репетитивный	267	1	268	99,63%	пере-	485	26	511	94,91%	4,71	Фишера, p=0,0003
по-: аттенуативный	30	32	62	48,39%	по-	561	252	813	69,00%	-20,62	χ²=10,248, p=0,0014
по-: делимитативный	417	154	571	73,03%	по-	561	252	813	69,00%	4,03	
по-: дистрибутивный	62	106	168	36,90%	по-	561	252	813	69,00%	-32,10	χ²=60,521, p<0,0001
по-: ингрессивный	26	43	69	37,68%	по-	561	252	813	69,00%	-31,32	χ²=26,643, p<0,0001
под-: аттенуативный	177	5	182	97,25%	под(о)-	418	7	425	98,35%	-1,10	
при-: аттенуативный	89	4	93	95,70%	при-	593	5	598	99,16%	-3,46	Фишера, p=0,0227
про-: пердуративный	183	166	349	52,44%	про-	839	75	914	91,79%	-39,36	χ²=250,848, p<0,0001
раз(о)/рас-: дистрибутивно-суммарный	154	3	157	98,09%	раз(о)/рас-	962	14	976	98,57%	-0,48	
раз(о)/рас...(-св): ингрессивно-интенсивный	169	36	205	82,44%	раз(о)/рас-	962	14	976	98,57%	-16,13	χ²=250,848, p<0,0001
у...-св: чрезмерно-интенсивный	16	2	18	88,89%	у-	875	31	906	96,58%	-7,69	
Всего	2954	1143	4097	72,10%	Всего	9663	802	10465	92,34%	-20,23	χ²=1039,968, p<0,0001
Среднее	134,27	51,95	186,23	73,80%	Среднее:	690,21	57,29	747,50	92,72%	-18,92	
Медиана	109,50	22,00	162,50	83,03%	Медиана:	645,00	36,00	796,00	95,10%	-12,08	

Приложение Г. Соотношение количества имперфективируемых и неимперфективируемых лексем в группе глаголов на -и(зи)рова- и в min-выборках

МАС	-и(зи)рова-				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	0	1	1	0%	289	23	312	92,63%	-92,63	
вы-	0	3	3	0%	777	160	937	82,92%	-82,92	Фишера, p=0,0051
за-	12	54	66	18,18%	1078	323	1401	76,95%	-58,76	χ ² =110,930, p<0,0001
на-	4	12	16	25%	549	235	784	70,03%	-45,03	Фишера, p=0,0003
о(б)-/об(о)-	3	4	7	42,86%	1073	254	1327	80,86%	-38,00	Фишера, p=0,0293
от(о)-	5	20	25	20%	692	116	808	85,64%	-65,64	Фишера, p<0,0001
пере-	13	19	32	40,63%	415	96	511	81,21%	-40,59	χ ² =27,337, p<0,0001
по-	0	12	12	0%	253	560	813	31,12%	-31,12	Фишера, p=0,0224
под(о)-	3	3	6	50%	406	19	425	95,53%	-45,53	Фишера, p=0,0021
про-	0	40	40	0%	629	285	914	68,82%	-68,82	Фишера, p<0,0001
раз(о)/рас-	6	11	17	35,29%	877	99	976	89,86%	-54,56	χ ² =45,112, p<0,0001
с(о)-	5	70	75	6,67%	1040	343	1383	75,20%	-68,53	Фишера, p<0,0001
у-	0	2	2	0%	778	128	906	85,87%	-85,87	Фишера, p=0,0204
Всего	51	251	302	16,89%	8856	2641	11497	77,03%	-60,14	χ²=572, p<0,0001
Среднее	3,92	19,31	23,23	18,36%	681,23	203,15	884,38	78,20%	-59,85	
Медиана	3	12	16	18,18%	692	160	906	81,21%	-63,03	

МАС+НКРЯ	-и(зи)рова-				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	0	1	1	0%	301	11	312	96,47%	-96,47	Фишера, p=0,0383
вы-	1	2	3	33,33%	849	88	937	90,61%	-57,27	Фишера, p=0,0255
за-	13	53	66	19,70%	1177	224	1401	84,01%	-64,31	χ ² =166,044, p<0,0001
на-	4	12	16	25%	603	181	784	76,91%	-51,91	Фишера, p<0,0001
о(б)-/об(о)-	4	3	7	57,14%	1158	169	1327	87,26%	-30,12	Фишера, p=0,0497
от(о)-	7	18	25	28%	734	74	808	90,84%	-62,84	χ ² =91,182, p<0,0001
пере-	15	17	32	46,88%	454	57	511	88,85%	-41,97	χ ² =41,571, p<0,0001
по-	0	12	12	0%	354	459	813	43,54%	-43,54	Фишера, p=0,0017
под(о)-	3	3	6	50%	418	7	425	98,35%	-48,35	Фишера, p=0,0002
про-	1	39	40	2,50%	708	206	914	77,46%	-74,96	Фишера, p<0,0001
раз(о)/рас-	7	10	17	41,18%	926	50	976	94,88%	-53,70	χ ² =75,678, p<0,0001
с(о)-	9	66	75	12%	1094	289	1383	79,10%	-67,10	χ ² =170,287, p<0,0001
у-	0	2	2	0%	801	105	906	88,41%	-88,41	Фишера, p=0,0138
Всего	64	238	302	21,19%	9577	1920	11497	83,30%	-62,11	χ²=755,402, p<0,0001
Среднее	4,92	18,31	23,23	24,29%	736,69	147,69	884,38	84,36%	-60,08	
Медиана	4	12	16	25%	734	105	906	88,41%	-63,41	

МАС+Runet	-и(зи)рова-				Min-выборки				РАЗНИЦА, %	χ ² /Фишера
	Есть НСВ2	Нет НСВ2	Всего	% НСВ2	Есть НСВ2	Нет НСВ2	Всего	% НСВ2		
в(о)-	1	0	1	100%	305	7	312	97,76%	2,24	
вы-	3	0	3	100%	889	48	937	94,88%	5,12	
за-	45	21	66	68,18%	1299	102	1401	92,72%	-24,54	χ ² =46,263, p<0,0001
на-	9	7	16	56,25%	697	87	784	88,90%	-32,65	χ ² =170,287, p<0,0001
о(б)-/об(о)-	4	3	7	57,14%	1267	60	1327	95,48%	-38,34	Фишера, p=0,0031
от(о)-	24	1	25	96%	770	38	808	95,30%	0,70	
пере-	31	1	32	96,88%	485	26	511	94,91%	1,96	
по-	2	10	12	16,67%	561	252	813	69%	-52,34	Фишера, p=0,0003
под(о)-	6	0	6	100%	418	7	425	98,35%	1,65	
про-	7	33	40	17,50%	839	75	914	91,79%	-74,29	χ ² =202,877, p<0,0001
раз(о)/рас-	17	0	17	100%	962	14	976	98,57%	1,43	
с(о)-	31	44	75	41,33%	1257	126	1383	90,89%	-49,56	χ ² =164,840, p<0,0001
у-	2	0	2	100%	875	31	906	96,58%	3,42	
Всего	182	120	302	60,26%	10624	873	11497	92,41%	-32,14	χ²=390,265, p<0,0001
Среднее	14,00	9,23	23,23	73,07%	817,23	67,15	884,38	92,70%	-19,63	
Медиана	7	1	16	96%	839	48	906	94,91%	1,09	

Russian Texts Detoxification with Levenshtein Editing

Илья Гусев

Moscow Institute of Physics and Technology

Moscow, Russia

ilya.gusev@phystech.edu

Abstract

Text detoxification is a style transfer task of creating neutral versions of toxic texts. In this paper, we use the concept of text editing to build a two-step tagging-based detoxification model using a parallel corpus of Russian texts. With this model, we achieved the best style transfer accuracy among all models in the RUSSE Detox shared task, surpassing larger sequence-to-sequence models.

Keywords: detoxification, style transfer, BERT, T5, tagging, text editing

DOI: 10.28995/2075-7182-2022-21-264-272

Преобразование оскорбительных текстов на русском языке с помощью предсказания редакционных предписаний

Илья Гусев

Московский физико-технический институт

Москва, Россия

ilya.gusev@phystech.edu

Аннотация

Детоксикация текста — это задача создания нейтральных версий оскорбительных текстов. В этой статье мы используем концепцию преобразования текста с помощью предсказания редакционных предписаний для построения двухэтапной модели детоксикации русских текстов при наличии параллельного корпуса. С помощью этой модели мы добились наилучшей точности передачи стиля среди всех участников дорожки RUSSE Detox, превзойдя более крупные sequence-to-sequence модели.

Ключевые слова: оскорбительные тексты, детоксификация, перенос стиля, BERT, T5, тегирование, редакционные предписания

1 Introduction

There is a vast amount of user-generated content on the Internet containing hate speech, profanity, toxicity, and aggression. It may not be appropriate for some platforms to show toxic texts. Some countries can even consider illegal writing or showing such content.

There are several ways to combat this problem. The obvious solution is to censor all toxic messages. Such texts can be deleted completely, covered with a warning, or placed at the very bottom of the page. However, it is ethically questionable apparent censorship.

Another way is to prevent writing such messages by suggesting alternative neutral options to a user. We will refer to the task of making such neutral variants of toxic texts as detoxification. It is a style transfer task where the source style is toxic, and the target style is neutral. The goal of this work was to build a system to solve this task.

Why is this task difficult?

1. Indistinct boundaries of what to consider toxic
2. Obfuscations that hide the meaning of words

3. Occasions of some rare insults
4. Sarcasm and other issues that require external world knowledge

Toxicity is a broad term that includes hate speech, obscene or condescending language, aggression, or grave insults. An instruction for annotators should define the particular rules for it. From the perspective of the shared task organizers, a text should contain «insults or obscene and rude words» to be considered toxic.

From a scientific perspective, it is a curious sequence-to-sequence task, where a target text is almost the same as a source one but with a different style. It allows specific methods that rely on the similarity of source and target texts.

This work is a part of the RUSSE Detox shared task (Dementieva et al., 2022), organized by a group of researchers as a part of the Dialogue-2022 conference. The goal of the shared task was to build a detoxification model with provided parallel corpus. Organizers also provided several baselines.

Our contributions:

1. We adopt a concept from the LEWIS paper (Reid and Zhong, 2021) to build a two-step tagging-based detoxification model using a parallel corpus of Russian texts.
2. We compare this tagging-based model with sequence-to-sequence baselines trained on the same corpus.
3. We propose a better model for toxicity classification.
4. We achieve the best style transfer accuracy among all models in the shared task.

Our code¹ and models²³⁴ are available online.

2 Related work

2.1 Toxicity classification

Nobata et al. (2016) made one of the first attempts to formulate the task of toxicity classification and collect a unified test dataset for it. They used comments posted on Yahoo Finance and News and rated by their in-house workers. The model was Vowpal Wabbit’s regression over different manual NLP features.

Gordeev (2016) selected anonymous imageboards (4chan.org, 2ch.hk) as the material for their corpus for the task of analysis of aggression. Authors utilized convolutional neural networks to detect the state of aggression in English and Russian texts.

Andrusyak et al. (2018) collected a dataset from Russian YouTube comments in an unsupervised way using a seed dictionary of abusive words and an iterative process updating this dictionary.

Smetanin (2020) used the Russian Language Toxic Comments Dataset (RTC dataset) from Kaggle⁵. It is the collection of annotated comments from 2ch⁶ and Pikabu⁷ websites. Fine-tuned RuBERT (Kuratov and Arkhipov, 2019) was the best model from this paper.

Zueva et al. (2020) introduced a novel corpus of 100000 comments posted on a major Russian social network (VK). As their primary model, they used a self-attentive encoder to get interpretable weights for each input token. They also used several tweaks, such as identity dropout and multi-task learning.

Saitov and Derczynski (2021) utilized Russian subtitles from the «South Park» TV show (RSP dataset) and the RTC dataset. Crowdsourcers annotated these subtitles for toxicity. Again, RuBERT held the best result.

Pronoza et al. (2021) focused on ethnicity-targeted hate speech detection in Russian texts. The authors composed a dataset of 5600 texts with 12000 mentioned instances of different ethnic groups. They named it RuEthnoHate. One more time, the modified RuBERT with additional linguistic features was the best model.

¹<https://github.com/IlyaGusev/rudetox>

²https://huggingface.co/IlyaGusev/rubertconv_toxic_clf

³https://huggingface.co/IlyaGusev/rubertconv_toxic_editor

⁴https://huggingface.co/IlyaGusev/sber_rut5_filler

⁵<https://www.kaggle.com/blackmoon/russian-language-toxic-comments>

⁶<https://2ch.hk/>

⁷<https://pikabu.ru/>

We used several of the mentioned datasets for toxicity classification to fine-tune a conversational RuBERT model.

2.2 Style transfer

Li et al. (2018) started the whole field of research, proposing a set of simple baselines for unsupervised text style transfer. The baselines were based on detecting style tokens with n-gram statistics and replacing them with altered retrieved similar sentences with the target style.

Wu et al. (2019) introduced a way to augment texts without breaking the label compatibility. They trained a **conditional BERT** (Devlin et al., 2019) using a conditional MLM task on a labeled dataset. Aside from data augmentation, they proposed to use this method as a part of a style transfer system, using an attention-based method to find style words and conditional BERT to replace them.

Krishna et al. (2020) suggested **STRAP**, **Style Transfer via Paraphrasing**. First, they generated a pseudo-parallel corpus. They started with styled texts and applied paraphrasers to normalize these texts in terms of style. The diversity of paraphrasing was promoted by filtering outputs heavily. Then, they fine-tuned style-changing inverse paraphrasers on this pseudo-parallel corpus. GPT2 (Radford et al., 2018) language model was used to implement both the paraphrasers and inverse paraphrasers. This scheme can also be used to augment an existing parallel corpus.

They also criticized existing style transfer evaluation methods and proposed an evaluation scheme based on transfer accuracy, semantic similarity, and fluency that we use in this work.

Malmi et al. (2020) introduced **Masker**, a system that used two language models to detect style tokens and padded masked language models to replace them. They tested it on sentence fusion and sentiment transfer. As for supervised tasks, they created **LaserTagger** (Malmi et al., 2019), a sequence tagging approach that casts text generation as a text editing task.

Krause et al. (2021) used **GeDi (Generative Discriminator)** to control generation towards the desired style. They use three language models: a base one, one for the desired style, and one for the undesired anti-style. The Bayes rule is applied during generation to compute style modifiers for every token from a vocabulary. Then these modifiers are applied to predictions of the base language model. This method allows computationally effective style-guided generation, but there is no source sequence, unlike the style transfer task. Dale et al. (2021) introduced the **ParaGeDi** method that applies GeDi for style transfer using a paraphrasing model instead of the base language model.

Dementieva et al. (2021) introduced the first study of automatic detoxification of Russian texts. They proposed two methods, the unsupervised one based on condBERT and the supervised one based on fine-tuning pretrained language GPT-2 model on a small manually created parallel corpus.

Reid and Zhong (2021) proposed **LEWIS (Levenshtein Editing With unsupervised Synthesis)**, the editing and synthesis framework for text style transfer. They had no parallel data, so the first task was to create a pseudo-parallel corpus. They used an attention-based detector of style words and two style-specific BART (Lewis et al., 2020) masked language models to replace these style words. Then they filtered resulting pairs with a style classifier, keeping only examples where the language models and the classifier agree.

After obtaining the pseudo-parallel corpus, they trained a RoBERTa-tagger (Liu et al., 2019) on it, predicting coarse edit types: «insert», «keep», «replace» and «delete» (Levenshtein, 1966). Then they trained a fine-grain edit generator to produce the target text, filling in phrases for coarse-grain edit types «insert» and «replace». We use this scheme almost without any modifications, but with a different language, with different base models, and already existing parallel corpus.

3 Evaluation

We built our style classifier by fine-tuning conversational RuBERT (Kuratov and Arkhipov, 2019) instead of the model⁸ proposed by organizers of the shared task. In addition to ok.ru⁹ and 2ch/Pikabu¹⁰ datasets,

⁸https://huggingface.co/SkolkovoInstitute/russian_toxicity_classifier

⁹<https://www.kaggle.com/datasets/alexanderseमितov/toxic-russian-comments>

¹⁰<https://www.kaggle.com/datasets/blackmoon/russian-language-toxic-comments>

Test type	Test description	Skolkovo clf., ER, %	Our clf., ER, %
INV	Replace yo	0.6	0.0
INV	Remove exclamations	0.9	0.4
INV	Add exclamations	0.9	0.3
INV	Captioned sentences to lowercase	73.9	34.8
INV	Remove question marks	4.0	0.2
INV	Add typos	3.6	1.9
INV	Masking of characters in toxic words	5.2	0.5
INV	Add typos to toxic words only	24.2	2.8
MFT	Concatenate non-toxic and toxic texts	15.5	3.1
MFT	Concatenate two non-toxic texts	2.1	0.6
MFT	Add toxic words from a vocabulary	16.3	0.1

Table 1: Error rates on different tests for two toxic classification models

Model	AUC, %	Accuracy, %	F1, %
Skolkovo classifier	66.2	86.4	37.2
Our classifier	73.5	90.3	51.3

Table 2: Metrics of toxicity classifiers on unseen crowdsourced test set: 3642 unique texts, 355 of them are toxic

we used Russian Persona Chat dataset¹¹ as a reliable source of non-toxic sentences.

We also tested models using a «checklist» (Ribeiro et al., 2020) methodology and augmented the resulting dataset with all the transformations. Test results are in Table 1. There are invariance (INV) and minimum functionality tests (MFT). Invariance tests ensure that a label will not change after a transformation, and MF tests have a fixed label to be predicted. It is clear from the table that our model has much lower error rates. From the user’s perspective, it is harder to pick up an adversarial example for our model than for the default one.

Two models have different dataset splits, so comparing them on their native test sets is wrong. However, we used crowdsourcing to evaluate the style transfer model, so we can use these annotations as an independent test set, keeping in mind that these samples are adversarial. Results for this new set are in Table 2. Our classifier shows better results in this setting.

We used models provided by organizers of the shared task for measuring semantic similarity¹² and fluency¹³. They have similar problems, but we did not come up with better options.

However, automatic metrics are not reliable, especially when being used with near-adversarial examples. Table 2 gives a glance at how unreliable they can be. To overcome this, we arranged our in-house annotation process with crowdsourcing through the Toloka¹⁴ platform in addition to the final evaluation provided by organizers of the shared task. We measured only style accuracy and semantic similarity, as fluency was much harder to define. Annotation instructions were close as possible to ones provided by the organizers and are available in the repository. Five workers annotated every sample. Samples were aggregated by majority vote. The average agreement was 90% for the style accuracy project, with Krippendorff’s alpha of 46%. For the similarity project, the average agreement was 88%, with Krippendorff’s alpha of 49%.

¹¹<https://toloka.ai/ru/datasets>

¹²<https://huggingface.co/cointegrated/LaBSE-en-ru>

¹³<https://huggingface.co/SkolkovoInstitute/rubert-base-corruption-detector>

¹⁴<https://toloka.ai>



Figure 1: Data generation for training a tagger and a generator.

4 Model

We see text detoxification as a two-step process. In the first step, a model should determine what words should be deleted or replaced. We can explicitly do it through tagging. In the second step, a generator replaces words or adds new ones. From this perspective, any classical sequence-to-sequence model has a trivial first step, as all words can be replaced.

4.1 Tagger — first step

4.1.1 Based on interpretation of a classifier

One way to find style tokens is to interpret a classification model. As for attention-based models, one can find such tokens using attention distribution. Tokens with high attention scores correlate with tokens that manifest style. Many researchers used this method (Xu et al., 2018; Wu et al., 2019; Hoover et al., 2020; Reid and Zhong, 2021).

It is also possible to use models that allow interpretation by design. Dementieva et al. (2021) utilized logistic regression and its weights for each word from the vocabulary for this task, and Li et al. (2018) used a Naive Bayes classifier.

4.1.2 Based on language models

Another way is to use two language models, one trained on texts of one style and another trained on texts of a different style. We can calculate the proportion of their predictions for every token if we have such models. If a prediction of the first model is much higher than that of the second model, then a corresponding token can be style-loaded. For instance, Masker (Malmi et al., 2020) used a similar approach.

4.1.3 Based on tags from parallel corpus

Finally, if we have a parallel corpus, we can directly compute edits required to transform source texts into target texts, convert these edits to tags, and then predict these tags with a token classification model.

4.2 Generator — second step

4.2.1 Based on MLM models

One way of filling the gaps is to use models pretrained for masked language modeling tasks (MLM) such as BERT (Devlin et al., 2019) or T5 (Raffel et al., 2020). It is their original task, but one can fine-tune them on texts of the required style (Wu et al., 2019). However, these models have no access to original words, so they can rely solely on context remained after masking.

4.2.2 Based on pairs from parallel corpus

The most direct way is to fine-tune a sequence-to-sequence model on a parallel corpus. Inputs are templates from the tagger, and outputs are masked words from the target sentence. The whole process of data generation for training is in Figure 1. We also concatenate a source sentence with the generated template, as in Reid and Zhong (2021), to provide access to the original masked words.

Architecture	Generator	STA, %	SIM, %
Seq2seq	T5 baseline	83.5	87.0
Seq2seq	cointegrated/ruT5-base	73.0	83.5
Seq2seq	sber/ruT5-base	81.0	85.0
LEWIP	cointegrated/ruT5-base	93.5	76.0
LEWIP	sber/ruT5-base	94.5	81.5

Table 3: In-house manual annotation with crowdsourcing, 200 random samples from the test set

4.3 Final model

Our final model uses tagger and generator, both based on a parallel corpus, so we will call it LEWIP (Levenshtein editing with parallel corpus), following the LEWIS (Reid and Zhong, 2021) scheme, as there is no «unsupervised synthesis» step.

We use conversational RuBERT as a base model for tagger and two versions of the ruT5-base model for generator^{15,16}, with the final submission based on the Sber model. We did not use ruT5-large in a shared task submission, as it did not fit into our GPU memory.

Organizers of the shared task provided a parallel corpus of 11090 pairs for training, a development set of 800 samples, and a test set of 875 samples. We used only that data for the style transfer model.

5 Results

5.1 In-house annotation and automatic metrics

Some results of in-house annotation are in Table 3. In addition to the T5 baseline provided by the organizers, we fine-tuned two ruT5-base models on our own. LEWIP with the Sber generator shows the best style transfer accuracy among all tested models but a lower similarity score than the baseline. Sber ruT5-base is better than the original ruT5-base for this task in both setups.

Examples where LEWIP failed to generate non-toxic texts according to in-house annotation:

- Original:** сколько же е**нутых в россии в месте с тобой
Translation: how many f**ked up people in russia are there with you
Generated: сколько же неадекватных в россии в месте с тобой
Translation: how many inadequate people in russia are there with you
- Original:** какие же эти люди сволочи!!!
Translation: what kind of bastards are these people!!!
Generated: какие же эти люди плохие !
Translation: how bad are these people!

Examples where LEWIP failed to generate similar texts according to in-house annotation:

- Original:** у курки е**ные не оболщатесь это миф.
Translation: f**king s toners, don't be fooled, it's a myth.
Generated: у вас не оболденьтесь это миф
Translation: you don't go crazy, it's a myth.
- Original:** Только хотел спросить, что за завалы. Е**ть хреновые в Рашке плотники
Translation: Just wanted to ask what are these obstructions. The carpenters in Russia are f**king bad
Generated: Только хотел спросить, что за завалы. в Рашке плотники
Translation: Just wanted to ask what are these obstructions. The carpenters in Russia

It seems that the tagger works well in most cases, and problems are mostly in the generator.

¹⁵<https://huggingface.co/cointegrated/rut5-base>

¹⁶<https://huggingface.co/sberbank-ai/ruT5-base>

Architecture	Model	Our STA, %	SIM, %	FL, %	J, %
Seq2seq	T5 baseline	86.3	82.7	83.7	59.3
Seq2seq	cointegrated/ruT5-base	78.8	85.0	83.9	55.2
Seq2seq	sber/ruT5-base	83.8	83.6	83.4	57.8
LEWIP	cointegrated/ruT5-base	93.6	79.7	88.4	66.1
LEWIP	sber/ruT5-base	93.1	79.8	88.5	65.8

Table 4: Automatic metrics on the test set

Team	STA, %	SIM, %	FL, %	J, %
Human References	88.8	82.4	89.4	65.3
T5 baseline	79.1	82.2	92.5	60.6
SomethingAwful	79.4	87.2	90.3	63.3
FRC CSC RAS	73.4	86.5	91.8	59.8
Our system	82.4	79.1	84.6	58.2

Table 5: Final results of the shared task, human evaluation, 3 top teams out of 10

Automatic metrics for the same set of models are in Table 4. Joint and STA scores for both LEWIP models are higher than the baseline.

5.2 Final human evaluation

The final results are in Table 5. Our model’s style transfer accuracy is much worse than our in-house annotation. We explain it with different instructions and annotation protocols. Still, our model has the best style transfer accuracy among all other models but with lower semantic similarity than the baseline. Pang and Gimpel (2019) showed that these metrics are complementary and challenging to optimize simultaneously.

We attempted to rank several beam search hypotheses from generators with automatic metrics to find different trade-offs, and we were successful in the sense of these automatic metrics. Nevertheless, it did not yield better human assessments. Generators were coming up with adversarial examples that were wrong for humans but good for automatic metrics.

5.3 Computational effectiveness

In some cases, we do not need the second step of the system. For 238 examples of 875 (27%) in the test set, the generator model was not run because there were no «replace» or «insert» tags. Sequence-to-sequence models are much more computationally expensive than encoder-only taggers. Moreover, a generator requires fewer steps than a raw sequence-to-sequence model, as it only fills the gaps. Overall, our system is more computationally effective than a T5 baseline.

6 Conclusions

- Thorough testing of a classification model helps in building data augmentations and, eventually, a much more stable model.
- Current automatic metrics are not reliable for evaluating systems trained on parallel corpora. They can work in a range of low values, e.g., for unsupervised style transfer, but there are too unstable to work with accurate models.
- Text editing models can perform at least as well as pure sequence-to-sequence models. They have the inductive bias based on the assumption that input and output are very close. They are also more environmental-friendly than pure sequence-to-sequence models.

References

- Bohdan Andrusyak, Mykhailo Rimel, and Roman Kern. 2018. Detection of abusive speech for mixed sociolects of russian and ukrainian languages. // Ales Horák, Pavel Rychlý, and Adam Rambousek, *The 12th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018, Karlova Studanka, Czech Republic, December 7-9, 2018*, P 77–84. Tribun EU.
- David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 7979–7996, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Methods for detoxification of texts for the russian language. *Multimodal Technologies and Interaction*, 5(9).
- Daryna Dementieva, Irina Nikishina, Varvara Logacheva, Alena Fenogenova, David Dale, Irina Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. Russe-2022: Findings of the first russian detoxification task based on parallel corpora.
- 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.
- Denis Gordeev. 2016. Detecting state of aggression in sentences using cnn. // *International conference on speech and computer*, P 240–245. Springer.
- Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, P 187–196, Online, July. Association for Computational Linguistics.
- Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. // *Findings of the Association for Computational Linguistics: EMNLP 2021*, P 4929–4952, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 737–762, Online, November. Association for Computational Linguistics.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.
- Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, feb. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.
- 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.
- Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 1865–1874, New Orleans, Louisiana, June. Association for Computational Linguistics.
- 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*, abs/1907.11692.
- Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. // *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 5054–5065, Hong Kong, China, November. Association for Computational Linguistics.

- Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8671–8680, Online, November. Association for Computational Linguistics.
- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. // *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, P 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. // *Proceedings of the 3rd Workshop on Neural Generation and Translation*, P 138–147, Hong Kong, November. Association for Computational Linguistics.
- Ekaterina Pronoza, Polina Panicheva, Olessia Koltsova, and Paolo Rosso. 2021. Detecting ethnicity-targeted hate speech in russian social media texts. *Information Processing and Management*, 58(6):102674.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.
- Colin Raffel, Noam M. 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. *ArXiv*, abs/1910.10683.
- Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. // *FINDINGS*.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 4902–4912, Online, July. Association for Computational Linguistics.
- Kamil Saitov and Leon Derczynski. 2021. Abusive language recognition in Russian. // *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, P 20–25, Kiyv, Ukraine, April. Association for Computational Linguistics.
- Sergey Smetanin. 2020. Toxic comments detection in russian. // *Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference "Dialogue"*.
- Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. // *ICCS*.
- Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. // *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 979–988, Melbourne, Australia, July. Association for Computational Linguistics.
- Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in Russian hate speech detection. // *Proceedings of the Fourth Workshop on Online Abuse and Harms*, P 65–69, Online, November. Association for Computational Linguistics.

The Role of Paragraph in the Corpora of Annotated Texts

Inkova O. Y.

Institute of Informatics Problems,
FRC CSC RAS, Moscow, Russia;
University of Geneva, Geneva,
Switzerland
Olga.Inkova@unige.ch

Nuriev V. A.

Institute of Informatics Problems,
FRC CSC RAS, Moscow, Russia
nurieff.v@gmail.com

Popkova N. A.

Institute of Informatics Problems,
FRC CSC RAS, Moscow, Russia
Natasha__popkova@mail.ru

Abstract

The paper focuses on the function of paragraph both in text organization and in text annotation from the point of view of coherence. Taking as examples three major types of corpora (the RST, ANNODIS, and PDTB corpora), it shows whether and to what extent the existing approaches account for the paragraph when a discourse relation gets annotated. Then it presents the theoretical principles underlying text annotation in two databases: the Supracorpora database of connectives and the Supracorpora database of hierarchical logical-semantic relations (a new linguistic resource). Text coherence is shown to result from the interaction of various discourse phenomena, acting at the level of local and global structures. In this approach, the paragraph is assigned to the meso-level, positioned between local and global levels. The researcher may analyze the internal organization of the paragraph, limiting oneself to the inter-sentential level. Yet, to analyze and describe how paragraphs follow one another in the text, it is necessary to operate at the supra-sentential level, adopting a conceptual apparatus fundamentally different from the one for the description of local text structure.

Keywords: paragraph; text annotation; corpus; discourse relations; database

DOI: 10.28995/2075-7182-2022-21-273-283

Роль абзаца в корпусах аннотированных текстов

Инькова О. Ю.

ИПИ ФИЦ ИУ РАН, Москва, Россия;
Женевский университет,
Женева, Швейцария
Olga.Inkova@unige.ch

Нуриев В. А.

ИПИ ФИЦ ИУ РАН,
Москва, Россия
nurieff.v@gmail.com

Попкова Н. А.

ИПИ ФИЦ ИУ РАН,
Москва, Россия
Natasha__popkova@mail.ru

Аннотация

Статья рассматривает функции абзаца в структуре текста, а также при аннотировании текстов с точки зрения связности. На примере трех наиболее известных корпусов (корпуса, созданные на основе Теории риторической структуры, корпус ANNODIS и PDTB) авторы анализируют существующие подходы и то, в какой степени абзац учитывается при определении дискурсивного отношения или, наоборот, его отсутствия. Авторы формулируют теоретические принципы, лежащие в основе аннотирования в двух базах данных: надкорпусной базы данных коннекторов и надкорпусной базы данных иерархии логико-семантических отношений, нового

лингвистического ресурса. Показано, что связность текста осуществляется в результате взаимодействия дискурсивных явлений различной природы, действующих на уровне как локальной, так и глобальной структуры. Абзац при таком подходе является единицей мезоуровня, промежуточного между локальным и глобальным. Если внутренняя организация абзаца может быть описана на межфразовом уровне, то следование абзацев в структуре текста должно быть описано на сверхфразовом уровне и в терминах, принципиально отличных от используемых для описания локальной структуры текста.

Ключевые слова: абзац; аннотирование текстов; дискурсивные отношения; база данных

1 Introductory remarks

It is known that nowadays for the annotation of discourse relations there are several approaches available. All of them are directly related to the theoretical approach that underlies the understanding of discourse relation and, more broadly, text coherence. We will start with a brief overview of resources where texts are annotated in terms of discourse relations (the RST, ANNODIS, and PDTB corpora), showing their specifics and annotation theoretical guidelines, primarily focusing on the function of paragraph in text organization. Then we will demonstrate how some theoretical assumptions have been adopted for the text annotation in the new linguistic resource – the Supracorpora database of hierarchical logical-semantic relations.

2 The RST corpora

In the annotation of discourse (or rhetorical) relations, the most common theoretical approach is known to be the Rhetorical Structure Theory (RST, Mann & Thompson 1988). The four main theoretical principles underlying it say that: (1) no piece of discourse should be left out of the analysis (completeness condition); (2) all text fragments are interconnected (connectedness condition); (3) the same discourse units (DUs), i.e. clauses, can be connected by only one relation (uniqueness condition); (4) DUs directly follow each other and cannot overlap between themselves (adjacency condition).

The text annotation has to satisfy these three conditions. Therefore

- the relations connecting DUs, however diverse they may be, are of the same order, be it syntactic (explanatory and relative clauses) or semantic dependence, anaphoric repetition, thematic progression, or text structure (division into paragraphs, chapters, title, author, etc.);
- in the hierarchical text structure, the elements of either global or local structure are connected by the same rhetorical relations, and the use of these relations is recursive;
- the entire text can be presented as a single graph.

That the view of the RST on text organization and coherence is somewhat simplified has been repeatedly pronounced by representatives of different linguistic schools. They have shown text coherence to be built simultaneously at several various levels: the genre of the text, its thematic organization, the communicative intentions of the speaker, the level of propositional content, and the level of discourse relations (that are understood in a narrower sense than in the RST). We will not dwell on this issue. For more on this see, for example, Adam 2012, Inkova 2019, Webber et al. 2012.

Presenting the text as a single graph limits the size of annotated texts. While the founders of the RST claim that the text length does not matter¹, to make a single graph is possible only for small texts. The Ru-RSTreebank Annotation Manual (<https://rstreebank.ru/>), adopting the RST principles, even specifies that graphs are built only within paragraphs.

This comes, in turn, from the consideration that the markers of global and local structures have the same functions, and the minimal unit of global structure, especially in a newspaper article, equals a paragraph. Hence, its function – keeping the text coherent – is comparable to the rhetorical relation that connects the elements of local structure, for example, in the fragment: *Он заболел, поэтому не пришел.* Such role of paragraph in text organization traces back to the works of Kenneth Lee Pike and Robert E. Longacre. Since the languages of the Philippines and Papua and New Guinea are known to have some specific identifiers of the beginning and end of a paragraph, linguistic scholars who follow the ideas of Pike (1982) decided to assign paragraph to the fourth level of grammatical units of surface structure. There are words and syntagmas at the first level, clauses – at the second level, and sentences – at the

¹ “It is insensitive to text size, and has been applied to a wide variety of sizes of text” (Mann, Thompson 1988: 243).

third level. The paragraph thus belongs to the inter-sentential level of linguistic analysis. As for Longacre, he considers “the paragraph as a grammatical unit” (Longacre 1979) and proposes in his latest classification of paragraphs (Longacre 1996) rubrics resembling the rhetorical relations that one might find in the RST.

Later we will return to the functions of paragraph and its role in making the text coherent as we understand it. However, now it is to be noticed that while the connective *поэтому* in our example serves to convey the connection between two minimal DUs, the function of paragraph is to signal, on the contrary, the weakening of this connectedness between two larger DUs.

3 The ANNODIS corpus

The notion of text as a complex multi-level and multi-parameter system reflects in the ANNODIS corpus (<http://redac.univ-tlse2.fr/corpus/annodis/>). This resource builds on the Segmented Discourse Representation Theory (SDRT) (Asher 1993, Ascher, Lascarides 2003) and consists of several independent sub-corpora. It aims to annotate various discourse phenomena, each of which contributes to text organization: rhetorical relations and two types of hierarchical structures, namely thematic chains (sequences of semantic blocks with a common topic) and enumerative structures. Through its annotations, the ANNODIS corpus shows how these three phenomena interact between them, making the text coherent. For example, for enumerative structures, one might see

- what rhetorical relation may hold between the initial sentence that specifies the enumeration and the enumeration itself (as a rule, it is Elaboration or Motivation),
- what relations, besides additive ones, can connect the members of the enumerative series.

The texts differ in length and genre, and they are annotated entirely. The annotation does not take into account the division into paragraphs, however, theoretically important is that the minimal DU can be a language unit less than a clause, and several minimal DUs can fall within its scope. We are talking about the so-called frame expressions, the function of which is to create semantic blocks of sentences that should be interpreted in relation to a single criterion (spatial, temporal, communicative). See (Charolles 1997, Inkova 2021) for details.

- (1) *Согласно креационистской гипотезе, которая имеет самую длинную историю, создание жизни есть акт божественного творения. Свидетельством этому является наличие в живых организмах особой силы, «души», управляющей всеми жизненными процессами. Гипотеза креационизма навеяна религиозными воззрениями и к науке отношения не имеет. (Л.А. Михайлов, Концепции современного естествознания. Учебник для вузов; books.google.it; accessed 12 January 2022)*

The frame expression *согласно креационистской гипотезе* helps refine the interpretation of the first sentence, in which it occurs, opening the paragraph. This expression also refines the interpretation of the second sentence. Otherwise, it would give the impression that the author argues in favor of the correctness of this hypothesis. And the author, on the contrary, refutes its accuracy, explaining it in the third and last sentence of the paragraph.

The function of frame expressions in the text is thus twofold: on the one hand, they serve to integrate, combining minimal DUs into larger ones, and on the other hand, they divide into segments, signaling a weaker connectedness between semantic blocks that should be interpreted “separately” (in this case due to different speakers). As we will see, the paragraph can assume the same functions in the text, often signaling that the scope of the frame expression comes to its end. Both the SDRT and the RST consider “relations” conveyed by frame expressions as “rhetorical” ones (Prévoit et al. 2009, Vieu et al. 2005).

4 The Penn Discourse Treebank

In the Penn Discourse Treebank (PDTB), they annotate, first of all, discourse relations that can potentially be expressed by a connective, i.e. the understanding of “rhetorical relation” is narrower than in the RST. Since the texts in the corpus (articles from the *World Street Journal*) are annotated entirely, the observations have resulted in three theoretical conclusions that distinguish this approach from the RST.

First, if it is impossible to place a connective between adjacent DUs, the relation cannot be qualified as discourse one. We are talking about such cases, “where the second sentence only serves to provide some further description of an entity in the first sentence” (PDTB Research Group 2008: 1). In (2) **EntRel** indicates this state of affairs:

- (2) Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern. **EntRel** Mr. Milgrim succeeds David Berman, who resigned last month. [Example borrowed from (PDTB Research Group 2008: 23)]

Secondly, even this kind of relation may not hold between adjacent sentences; see **NoRel** in (3).

- (3) Jacobs Engineering Group Inc.’s Jacobs International unit was selected to design and build a microcomputer- systems manufacturing plant in County Kildare, Ireland, for Intel Corp. *Jacobs is an international engineering and construction concern.* **NoRel** *Total capital investment at the site could be as much as \$400 million, according to Intel.* [Example borrowed from (PDTB Research Group 2008: 25)]

In the description of **NoRel** cases, importantly, there is a mention of paragraph: “NoRel indicates (of adjacent sentences within a paragraph) that no relation holds between them” (PDTB Research Group 2019: 5). This means by default that the paragraph has two aforesaid functions. It can: (1) create a single semantic block, within which sentences should be interpreted together; (2) create a boundary between semantic blocks, signaling no immediate connection between them.

Thirdly, it is assumed that DUs connected by a discourse relation do not have to immediately follow each other (as opposed to the adjacency condition in the RST).

To visualize, in particular, such cases, the PDTB uses colors, which allows you to see the boundaries of DUs. However, the PDTB corpus has a significant flaw: its annotation does not account for the hierarchy of discourse relations. See Figure 1.

Mr. Hahn began selling non-core businesses, such as oil and gas and chemicals.
He even sold one unit that made vinyl checkbook covers.
At the same time, he began building up the pulp and paper segment of the company while refocusing building products on home repair and remodeling, rather than materials for new-home construction.

Figure 1: The annotation of the explicit “temporal synchronous”² relation in the PDTB

In Figure 1, we see that the temporal synchronous relation expressed by *at the same time* (highlighted in red) connects arguments 1 and 2, highlighted in yellow (argument 1) and blue (argument 2) and separated by another sentence (without highlighting). This latter, in turn, is argument 2 in the implicit conjunction relation (see Figure 2) for the same argument 1, thus included in two relations. In the annotation, *in fact* is the connective conveying the conjunction relation.

Mr. Hahn began selling non-core businesses, such as oil and gas and chemicals.
He even sold one unit that made vinyl checkbook covers.
At the same time, he began building up the pulp and paper segment of the company while refocusing building products on home repair and remodeling, rather than materials for new-home construction.

Figure 2: The annotation of the implicit conjunction relation in the PDTB

² The tag is used in the PDTB.

5 The Supracorpora database of hierarchical logical-semantic relations

A new linguistic resource, the Supracorpora database of hierarchical logical-semantic relations (hereupon referred to as the SDB of hierarchical LSRs), aims to enlarge the annotation capabilities provided by the Supracorpora database of connectives (the SDB of connectives³). Firstly, it shows the boundaries of text fragments connected by explicit and implicit relations. Secondly, it visualizes the relation hierarchy. Thus, the “вопреки ожидаемому” (“contrary to the expected state of affairs”) relation expressed by the Russian connective *но* is annotated in the SDB of connectives as shown in Figure 3.

<p>– Слушай, – сказал он Зосимову, – ты малый славный, но ты, <i>кроме</i> всех твоих скверных качеств, <i>еще и</i> потаскун, это я знаю, <i>да еще</i> из грязных.</p>	<p style="text-align: center;">НО <“вопреки ожидаемому”> <сложное предложение> <начальная> <р CNT q> <CNT> <SuperCNT></p>
---	--

Figure 3: The annotated occurrence of the Russian connective *но* in the SDB of connectives

If the focus is on this relation, the left context of the fragment where the connective (CNT) *но* occurs is wider than needed (the part between the dashes is unnecessary). And in its right context, there are two more LSRs: additive propositional relation (*кроме... еще и*) and additive illocutionary relation (*да еще*). This relation hierarchy is not visible in the annotation. Its only mark is the SuperCNT tag, i.e. *но* is the “embedding connective” (in terms of the SDB), and the italicization of “embedded” connectives that fall within the scope of *но*.

Theoretically, the annotation in the SDB of hierarchical LSRs builds on the principles rather close to those of the PDTB. However, the annotation does not cover the whole text, since it is, first of all, the occurrence of a connective that gets annotated (although the SDB allows annotating implicit relations as well⁴). Therefore, there are no limitations on the length of annotated texts, which is important, as the SDB processes texts of significant length, primarily fictional, scientific, and newspaper ones. Hence, there is no need to resort to the criterion of (typo)graphic paragraph⁵. To our mind, the relations between the elements of local and global text structures are fundamentally different, and the paragraph itself occupies an in-between level, or the “meso-level” (the term is from Adam 2018). The following arguments can prove this position.

1) Regarding its internal organization, the paragraph can, in most cases, be defined through the connection between sentences at the inter-sentential level (morphology, semantics, and syntax). Yet, how paragraphs follow each other in the text is subject to discourse laws and needs to be described at the supra-sentential level in terms other than “rhetorical relations” at the local level. Such relations cannot explain how sentences merge into larger – semantically homogeneous and macrostructural – discourse units. Cf. the terms Longacre (1968) uses to classify paragraphs, resting upon major types of text passages (narrative, explanatory, expository, hortatory, procedural, and dialogue paragraphs).

2) The sentences in a paragraph do not make up a simple chain, since they are discursively heterogeneous. Of the greatest importance are the opening and final sentences, and the graphic paragraph is to emphasize this importance. The opening sentence introduces the topic that will evolve throughout the sentences grouped in the paragraph. And the incomplete line ending the paragraph signals that the previous information is detached from the subsequent information. The psycholinguistic experiments

³ For more details about the architecture of the SDB of connectives, its interface and functionality, see Inkova 2018, Inkova & Popkova 2017. For the architecture of the SDB of hierarchical LSRs and its functional content, see Durnovo et al. 2022.

⁴ An implicit relation gets annotated only if it becomes explicit in the target text, or vice versa, if a relation is explicit in the source text and becomes implicit in translation.

⁵ This criterion is likely to be artificial, which is clear from the Ru-RSTreebank Annotation Manual, already quoted earlier: “If there are less than three clauses in a paragraph, we attach it, depending on the meaning, to the ‘tree’ of the previous or next paragraph. If the text does not show a distinct division into paragraphs, and, for example, there are many quotes from various sources (see news texts) – follow the meaning” (https://docs.google.com/document/d/1wdsgGyIo5AQq2IPj6jWa_QmU0fUohXj48qsfVDgcBs/edit, p. 1).

showing a slowdown in reading speed in these zones prove the integrating and demarcating functions to pertain to the opening and final sentences of the paragraph (Coirier, Gaonac'h & Passerault 1996, ch. 14). The same is true for language data. If a connective marks the border between paragraphs or even chapters, it certainly connects not paragraphs or chapters, but the last and first sentences of consecutive paragraphs. Cf. (4), where *однако ж* begins the third chapter of the second part of the novel. The connective expresses the “contrary to the expected state of affairs” relation that holds between the sentence closing the last paragraph of the previous chapter (“Затем наступило беспмятство”) and the sentence opening the paragraph of the third chapter (“Он не то что уж был совсем в беспмятстве”). Then there follows the description of what Raskolnikov remembered and what he forgot. The paragraph ends with his recovery.

- (4) Она сошла вниз и минуты через две воротилась с водой в белой глиняной кружке; но он уже не помнил, что было дальше. Помнил только, как отхлебнул один глоток холодной воды и пролил из кружки на грудь. Затем наступило беспмятство.

III

Он, *однако ж*, не то чтоб уж был совсем в беспмятстве во всё время болезни: это было лихорадочное состояние, с бредом и полусознанием. Многое он потом припомнил. <...> Наконец он совсем пришел в себя. [Ф. М. Достоевский. Преступление и наказание (1866)]

In general, the internal structure may differ from one paragraph to another; it cannot be described in terms of “rhetorical relations” at the local level of text coherence. Cf., for example, the structure of the argumentative paragraph below.

3) While, as shown above, paragraphs help readers interpret the text, the (typo)graphical paragraph is often unnecessary. Firstly, language signals other than paragraphs also help see the topical unity (anaphoric repetitions, connectives, temporal markers, indicators of topic change, headings, subheadings, etc.). Secondly, many corpora (the Russian National Corpus, Frantext) overlook the division into paragraphs. Thirdly, different editions of the same text may have different divisions into typographical paragraphs (cf. Adam 2018, ch. 4). Fourthly, from the translation perspective, translators can change the paragraphing of the source text (Adam 2018: 66-67, Nuriev 2021: 371-384). Moreover, the concept of the paragraph itself is not universal. While most European languages distinguish between the sentence and the paragraph, languages such as, for example, Japanese, Soddo (Ethiopia), Newar (Nepal) or Godié (Ivory Coast) do not.

4) On the other hand, sometimes paragraph boundaries happen to be somewhat misleading: (typo)graphic paragraphs in the text may not coincide with semantic paragraphs mentally reconstructed by the reader. See the Anglo-Saxon opposition between *orthographic paragraph (o-paragraph)* and *semantic paragraph (s-paragraph)*. In this regard, we recall the well-known experiment of Teun van Dijk (1981: 183-190), who splits eleven graphic paragraphs of a Newsweek article (the news story type) into thirteen semantic paragraphs. The semantic paragraph is semantically coherent, which is usually described in terms of the topical or thematic unity (cf. “thematic paragraph” in Givón 1983: 8, and also, among others, Bain 1867, Albadalejo Mayordomo & Garcia Berrio 1983, Adam 2018: 65-82, Hoey 2005, Hoffmann 1989) and is quite obvious to the readers when they move from one paragraph to another. The (typo)graphical paragraph, on the contrary, can be a mere convention determined by other factors, including those of extralinguistic nature (for example, the text layout strategies or the editorial traditions, etc.).

Without going into details on the relationship between semantic and graphic paragraphs, one can say that the division into paragraphs is rather free from strict formal or grammatical laws. So it would be a clear exaggeration to argue that each new graphic paragraph introduces a new topic breaking the referential unity and that any paragraph has only one topic. “In the paragraph, we have uncovered the specific “play” of mild-level structure, which both builds upon smaller components, and acts as a building-block of much larger object. In this looking both “below” and “above” itself, paragraph enjoys a uniquely central position in the economy of texts” (Algee-Hewitt et al. 2015: 22). To recognize the paragraph as an in-between – meso-level – text unit means to connect it with both local (inter- and super-sentential) text structure and the global one. The latter, notably, establishes a hierarchy of text passages that, in turn, largely depends on the genre and stylistic conventions and the publishing traditions.

Building on these theoretical considerations, the SDB of hierarchical LSRs approaches the question of hierarchical text structure regardless of whether the fragments connected by LSRs are in the same paragraph or different paragraphs or even chapters. For example, let us take an argumentative paragraph and compare its graphic, thematic, and discourse design. Traditionally, it should have a tripartite structure, corresponding to an argumentative passage: the first part introduces a thesis, the second part is an argument, and the third part gives a conclusion. Cf. (5), describing how the “dead civil servant” takes the overcoat from the significant personage:

- (5) 1. «А! так вот ты наконец! 2. наконец я тебя того, поймал за воротник! 3. твоей-то шинели мне и нужно! 4. не похлопотал об моей, 5. да еще и распек, – 6. отдавай же теперь свою!» [Н. В. Гоголь. Шинель (1842)]

DUs 1-3 introduce a thesis, DUs 4-5 explain why the overcoat needs to be taken, and DU 6 makes a conclusion. However, it is not the graphic paragraph that reinforces the unity of the argumentative passage here (there are very few paragraphs in the story). The quotation marks enclose the character’s dialogue cue and draw the boundaries of the passage.

We see a different situation in the following excerpt from Ivan Goncharov’s *Oblomov*. The argumentation that we have to abridge is contained in 15 graphic paragraphs and 18 graphic sentences (labeled by P), i.e. the number of paragraphs and the number of sentences (differing significantly in length and complexity) are almost identical.

- (6) §1. **P1** Утешься, добрая мать: твой сын вырос на русской почве – не в будничной толпе, с бюргерскими коровьими рогами, с руками, ворочающими жернова. **P2** Вблизи была Обломовка: там <...>! **P3** Там <...>; там <...>.
 §2. **P4** Да и в самом Верхлёве стоит, хотя большую часть года пустой, запертой дом, но туда частенько забирается шаловливый мальчик, и там видит он <...>, – видит <...>; видит <...>.
 §3. **P5** Он в лицах проходит <...>; читает <...> ...
 §4. **P6** Года в три раз этот замок вдруг наполнялся народом, <...>.
 §5. **P7** Приезжали князь и княгиня с семейством: князь, <...>; княгиня <...>.
 §6. **P8** Она казалась <...>.
 §7. **P9** Зато в доме, кроме князя и княгини, был целый, такой веселый и живой мир, что Андрюша <...>.
 §8. **P10** Тут были князя Пьер и Мишель, из которых первый <...>.
 §9. **P11** Другой, Мишель, только лишь познакомился с Андрюшей, как <...>.
 §10. **P12** Дня через три Андрей, <...>, разбил ему нос.
 §11. **P13** Были еще две княжны, <...>.
 §12. **P14** Была их гувернантка, <...>. **P15** Она <...>!
 §13. **P16** Потом был немец, <...>, потом учитель музыки, <...>, потом целая шайка горничных, наконец стая собак и собачонок.
 §14. **P17** Все это наполняло дом и деревню шумом, гамом, стуком, кликами и музыкой.
 §15. **P18** С одной стороны Обломовка, с другой – княжеский замок, с широким раздольем барской жизни, встретились с немецким элементом, и не вышло из Андрея ни доброго бурша, ни даже филистера. [И. А. Гончаров. Обломов (1848-1859)]

§1. **P1** introduces the following thesis: Stolz’s mother should take comfort in that he will not become one hundred percent German. **P2-P3** give the first argument: in the neighborhood, there is Oblomovka with its Russian way of life, the description of which is an unmarked enumeration, emphasized only by the parallelism of three *там*. The second argument – the princely house at Verhlyovo with its residents and customs – is introduced by the connective *да и* and separated from the first argument by the graphic paragraph. But due to its complexity, this argument is also divided into paragraphs. Being a kind of macro-argument, it describes both what Andrey sees in the house (§§2-3) and the inhabitants of the house who come once in three years (§§4-14). This description could make up a single graphic paragraph since each graphic paragraph here equals in size a graphic sentence of insignificant length. §15 concludes in support of the thesis (**P1**). Due to limitations in volume, we will not dwell much on the internal organization of the argument and will show its scheme (see Figure 4).

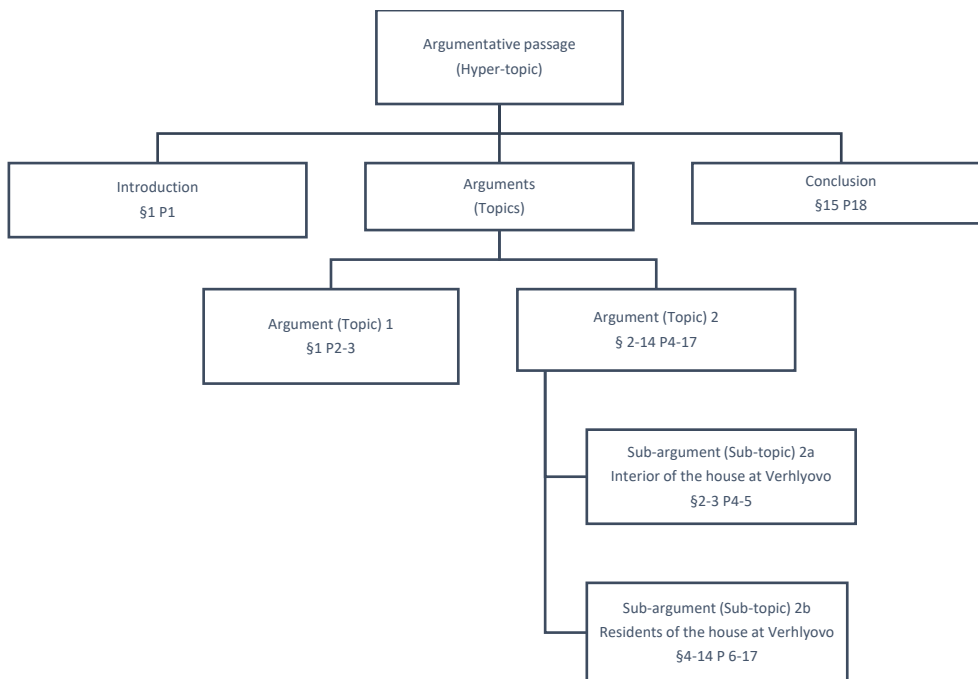


Figure 4: The topical organization of the argumentative passage (6)

Regarding the discourse relations, the SDB of hierarchical LSRs allows building a graph for the connective *da u* (see Figure 5). It pictures the hierarchy of text fragments **P2-P3** (the left context) and **P4-P5** (the right context) falling within its scope and separated from sub-topic 2 by a strong punctuation mark – ellipsis points. The principal difference from the PDTB annotation is the possibility to visualize the relation hierarchy.

Thus, we see that various levels of text organization are not identical; they are somewhat superimposed on each other. If the scopes of diverse means contributing to text coherence coincide or overlap, it leads to greater coherence. As for global structure, its analysis should adopt a different conceptual apparatus that makes it possible: (1) to explain how the units of local structure follow one another, building global text structure, and (2) to account for stylistic and genre criteria.

6 Conclusion

To sum up the whole matter, the corpus annotation, accounting for the complex nature of various discourse elements in text organization, appears to be more thorough and theoretically justified than one that uses the same rhetorical relations to annotate units of all hierarchical levels. The former – the multi-level annotation – does not lead to oversimplification and shows more clearly how different discourse phenomena involved in the creation and interpretation of a coherent text interact between them. Cf. the corpus annotation in the ANNODIS project and the RST annotation.

As for the paragraph, since it is a unit of the level between local and global text structures, it is of little relevance when analyzing discourse phenomena at the level of local text structure. The paragraph is even less relevant, as we have seen, for delineating the boundaries of units at the level of global text structure.

Acknowledgements

Joint Switzerland-Russia Research Preparation Grant “Annotation methodology in a supracorpora database of connectives” of the State Secretariat for Education, Research, and Innovation.

The research was carried out using the infrastructure of the Shared Research Facilities «High Performance Computing and Big Data» (CKP «Informatics») of FRC CSC RAS (Moscow).

References

- [1] Adam J.-M. (2012) The emergentist model in text linguistics [Le modèle émergentiste en linguistique textuelle], *Information grammaticale* 134. Pp. 30–37.
- [2] Adam J.-M. (2018) The paragraph: between the sentences and the text [Le paragraphe : entre phrases et texte]. Paris : Arman Colin.
- [3] Albadalejo Mayordomo T., García Berrio A. (1983) Compositional structure. Macrostructures [Estructura composicional. Macroestructuras], *Estudios de Lingüística. Universidad de Alicante* 1. Pp. 127–180.
- [4] Algee-Hewitt M., Heuser R., Moretti F. (2015) On paragraph. Scale, themes, and narrative form, Pamphlets of the Stanford literary lab, Pamphlet 10. Pp. 1–22.
- [5] Asher N. (1993) *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers: Dordrecht.
- [6] Asher N., Lascarides A. (2003) *Logics of Conversation*. Cambridge: Cambridge University Press.
- [7] Bain A. (1867) *English Composition and Rhetoric. A manual*. New York: Appleton.
- [8] Charolles M. (1997) The framing of discourse [L'encadrement du discours], *Cahier de Recherche Linguistique* 6. Pp. 1–73. URL: <https://hal.archives-ouvertes.fr/hal-00665849>.
- [9] Coirier P., Gaonac'h D., Passerault J.-M. (1996) *Text psycholinguistics [Psycholinguistique textuelle]*. Paris: Armand Colin.
- [10] Durnovo A.A., Inkova O. Yu., Popkova N.A. (2022) Database of hierarchical logical-semantic relations: architecture [Arhitektura bazy dannyh ierarhii logiko-semanticheskikh otnoshenij], *Systems and Means of Informatics [Sistemy i Sredstva Informatiki]* 1. Pp 114–125.
- [11] Givón T. ed. (1983) *Topic continuity in Discourse: A quantitative cross-language study*. Amsterdam/Philadelphia: John Benjamin.
- [12] Hoey M. (2005) *Lexical priming: a new theory of words and language*. London/New York: Routledge.
- [13] Hoffmann T. (1989) Paragraphs & anaphora, *Journal of Pragmatics* 13. Pp. 239–250.
- [14] Inkova O.Yu. (2018) The language-specificity of connectives: methods and parameters of description [Lingvospetsifichnost' konnektorov: metody i parametry opisaniya], *Semantics of connectives: a contrastive study [Semantika konnektorov: kontrastivnoe issledovanie]*, O. Inkova (ed.), Moscow: TORUS PRESS. Pp. 5–23.
- [15] Inkova O. (2019) Logical-semantic relations: classification problems [Logiko-semanticheskie otnosheniya: problemy klassifikatsii], O. Inkova, E. Manzotti, *Text coherence: mereological logical-semantic relations [Svyaznost' teksta: mereologicheskie logiko-semanticheskie otnosheniya]*. Moscow: Izdatel'skii Dom YaSK. Pp. 11–98.
- [16] Inkova O.Yu. (2021) Text incoherence, or some pitfalls of automatic text processing [Nesvyaznost' teksta, ili o nekotorykh podvodnykh kamnyakh na puti avtomaticheskoy obrabotki teksta], *Tomsk State University Journal of Philology [Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya]* 74. In press.
- [17] Inkova O., Popkova N. (2017). Statistical data as information source for linguistic analysis of Russian connectors, *Informatics and applications [Informatika i ee primeneniya]* 11(3). Pp. 123–131.
- [18] Longacre R. E. (1968) *Discourse, paragraph and sentence structure in selected Philippine languages, Vol 2. Sentence structure*. Santa Ana: Summer Institute of Linguistics.
- [19] Longacre R. E. (1979) *The Paragraph as Grammatical Unit*, T. Givón (ed.) *Syntax and semantics. Discourse and Syntax*, vol. 12. New York: Academic Press. Pp. 115–134.
- [20] Longacre R. E. (1996) *The Grammar of Discourse*. 2nd ed. (1st ed. 1983). New York/London: Plenum Press.
- [21] Mann W., Thompson S. (1988) *Rhetorical structure theory: Towards a functional theory of text organization*, *Text* 8. Pp. 243–281.
- [22] Nuriev V. (2021) *Literary translation through the lens of language experiment (syntactic aspect) [Khudozhestvennyj perevod skvoz' prizmu jazykovogo jeksperimenta (sintaksicheskij aspekt)]*. D.Sc. thesis. Moscow: The Military University of the Defense Ministry of the Russian Federation. Pp. 371–384.
- [23] PDTB Research Group (2008) *The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01*. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania; URL: <https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf>.
- [24] PDTB Research Group (2019) *The Penn Discourse Treebank 3.0 Annotation Manual*; URL: <https://doi.org/10.35111/qebf-gk47>.
- [25] Pike K. L. (1982) *Linguistic Concepts: An Introduction to Tagmemics*. Lincoln and London: University of Nebraska Press.

- [26] Prévot L., Vieu L., Asher N. (2009) A more precise formalization for a less confused annotation: the Elaboration relation [Une formalisation plus précise pour une annotation moins confuse : la relation d'élaboration d'entité], *Journal of French Language Studies* 19. Pp. 207–228.
- [27] van Dijk T. A. (1981) Episodes as units of discourse analysis, D. Tannen (ed.) *Analysing Discourse: Text and Talk*. Georgetown: Georgetown University Press. Pp. 177–195.
- [28] Vieu L., Bras M., Asher N., Aurnague M. (2005) Locating adverbials in discourse, *Journal of French Language Studies* 15. Pp. 173–193.
- [29] Webber B. L., Egg M., Kordoni V. (2012) Discourse Structure and Language Technology, *Natural Language Engineering* 18 (4). Pp. 437–490.

“Word-by-word” melodic contour in Russian dialects: quantitative approach

Sergey Knyazev
Vinogradov Russian Language Institute,
Russian Academy of Sciences /
Moscow, Russia
svknia@gmail.com

Maria Evstigneeva
Banks Soft Systems /
Moscow, Russia
masha.ev2012@yandex.ru

Abstract

The paper presents results of quantitative analysis of phrasal tonal structure in two Northern Russian dialects with different types of “word-by-word” melodic contour. These dialects differ from Modern Standard Russian by the quantity of pitch accents since their 60% of words bear pitch accent, thus the prosodic unit in them is not a (phonological) word, but an accent group. In addition, the dialects differ from Standard Russian by regular presence of even tone on the accented vowel (in Arkhangelsk dialect 86% of all accents have it; in Vologda dialect it is less frequent: 33%) and higher frequency of pitch accents with increased interval. The main differences between Arkhangelsk and Vologda dialects are 1) the ratio of rising and falling pitch accents: 2.6% falling in Arkhangelsk dialect and 56% in Vologda dialect, it brings the latter closer to Standard Russian (53%) and 2) the level of the base tone on which the main tonal changes occur (high and medium, respectively). Thus “word-by-word” melodic contour exists at least in two varieties: with rising tonal movement and with a falling tone in the function of an ornamental accent. In general, the intonation system of Vologda dialect, though there are a lot of significant differences, is much closer to Modern Standard Russian than to Arkhangelsk dialect.

Keywords: Russian; dialects; phrase prosody; pitch accent

DOI: 10.28995/2075-7182-2022-21-284-294

«Пословный» тональный контур в русских говорах: количественный аспект

С. В. Князев
Институт русского языка
им. В. В. Виноградова РАН /
Москва, Россия
svknia@gmail.com

М. Ю. Евстигнеева
Банкс Софт Системс /
Москва, Россия
masha.ev2012@yandex.ru

Аннотация

В статье представлены результаты количественного анализа тональной структуры высказывания в двух севернорусских говорах с разными типами «пословного» мелодического контура. Исследованные говоры отличаются от литературного языка вдвое большей частотностью тональных акцентов: они фиксируются на 60% слов, способных нести акцент, тем самым, просодической единицей наименьшей размерности является в них не (фонетическое) слово, а акцентная группа. Кроме того, говоры отличаются от литературного языка регулярным наличием ровного тона на выделенном гласном (в архангельском говоре 86% от всех рассмотренных акцентов обладают соответствующей особенностью; в вологодском говоре 33%), и большей частотностью тональных акцентов с увеличенным интервалом. Основные различия между архангельским и вологодским говорами заключаются 1) в соотношении восходящих и нисходящих тональных акцентов: 2,6% нисходящих в первом и 56% во втором, что сближает вологодский говор с литературным русским языком (53%) и 2) в уровне базового тона, на котором происходят основные тональные изменения (высокий и средний соответственно). Таким образом, «пословный» мелодический контур существует, как минимум, в двух разновидностях: с восходящим тональным движением и с нисходящим тоном в функции орнаментального акцента. В целом, интонационная система вологодского говора, при всех существенных отличиях, ближе литературной русской, чем просодической системе архангельского говора.

Ключевые слова: русский язык; диалект; фонетика; фразовая просодия; тональный акцент

1 Введение

Одной из черт просодической организации высказывания в севернорусских говорах является так называемое *пословное тональное оформление*: в них

«фраза состоит из ряда отрезков с восходящей интонацией» [15: 14];

«ударные слоги выделялись по тону, так что каждое слово имело свой мелодический контур» [3: 262];

«обращает на себя внимание мелодическая выделенность, как бы особая интонационная самостоятельность слов, включенных в общий интонационный контур фразы. Почти каждое слово во фразе получает свое мелодическое оформление, становясь, таким образом, обособленным от соседних слов» [19: 64];

«характерно пословное оформление интонационного контура, в то время как среднерусским говорам свойственно посинтагменное оформление интонации» [19: 78];

«интонационный контур фразы в этих говорах имеет тенденцию к пословному построению. Иными словами, в интонации говора отмечается склонность к выделению каждого самостоятельного слова с примыкающими к нему проклитиками и энклитиками в отдельную синтагму» [20: 94];

«Many <...> utterances show a repeating rising-falling intonation pattern on each phonological word» [22: 46].

Т. М. Николаева, отмечавшая это явление на материале других славянских языков (белорусского, словацкого, македонского), полагала, что оно свойственно наиболее архаичным языковым системам и указывала, что оно сочетается с отсутствием «сильной фразовой просодии» [16: 260, 262].

Изначально считалось, что такое оформление достигается «с помощью тонального подъема на ударном гласном каждого фонетического слова» [9: 42], однако существуют говоры и с нисходящим тоном в «пословном» контуре [11; 4]. Соответствующие примеры приведены ниже на рисунках 1 (восходящий тон на каждом фонетическом слове) и 2 (нисходящий)¹.

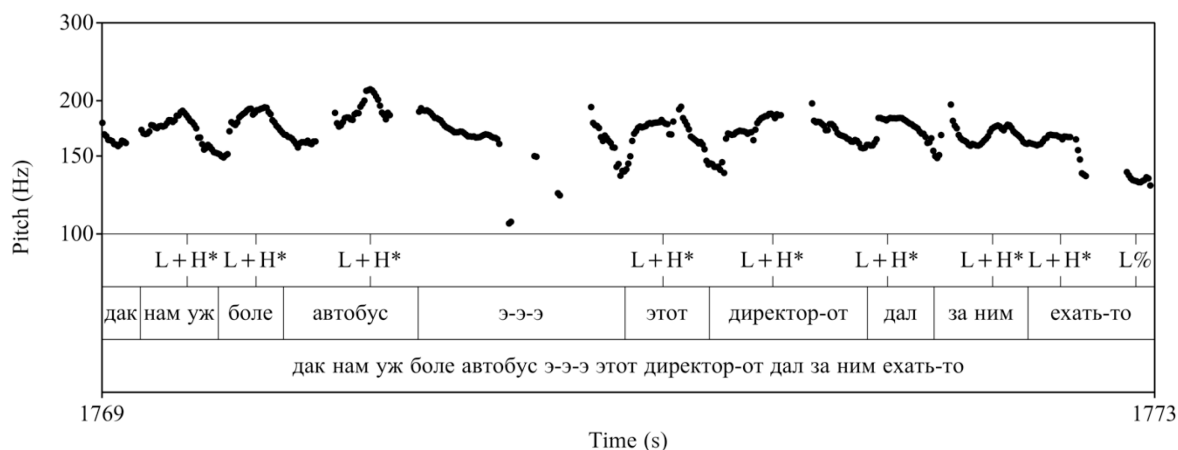


Рисунок 1: Кривая частоты основного тона фразы
Да́к на́м уж бо́ле авто́бус э-э-э дире́ктор-от да́л за ни́м е́хать-то
 (д. Вадюга Верхнетоемского р-на Архангельской области, диктор — М. П. Е.)

¹ Об использовании нами для описания просодических явлений нотации ToBI [21], [1] см. [13].

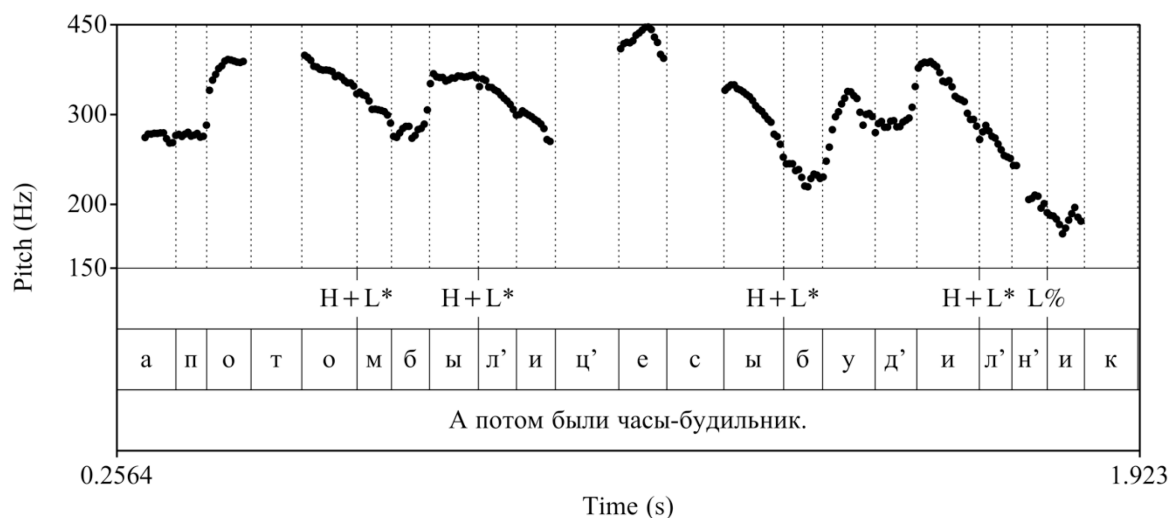


Рисунок 2: Кривая частоты основного тона фразы
А потом были часы-будильник
 (с. Церковное Плесецкого р-на Архангельской области, диктор — Ф. Т. Т.)

2 Цель исследования

Как видно из приведенных выше описаний «пословного мелодического контура», одни исследователи считают, что в нем тонально выделено каждое слово, другие – что только самостоятельные, третьи – что это явление существует лишь в виде тенденции и отдельным тональным движением оформляется «почти каждое слово». Целью настоящего исследования в связи с этим является анализ количественного соотношения тонально выделенных и невыделенных слов в (северно)русских говорах с «пословным тональным оформлением» и выделение минимальной просодической единицы² в этих говорах.

3 Материал исследования

Материалом исследования служили диалектные записи из

д. Вадюга Верхнетоемского р-на Архангельской области³ (тип 1, верхнепинежские говоры), сделанные в июле 1987 г. в ходе диалектологической экспедиции МГУ и ИРЯ РАН под руководством С. К. Пожарицкой (диктор Р. А. О., 67 лет).

д. Евсеевская Тарногского района Вологодской области (тип 2, вологодский говор), собранные Л. Ю. Зориной, И. И. Исаевым, О. В. Никитиным, С. В. Шильниковской в 2012 году (диктор К. З. Я., 83 года).

4 Процедура исследования

Процедура исследования была следующей:

1. Сначала из записей спонтанных диалогических текстов каждого говора было отобрано по 100 речевых отрезков из трех и более (до 15) слов, ограниченных либо реальными паузами, либо явной сменой базовых просодических параметров – интенсивности и частоты основного тона

² В современной интонологии, опирающейся преимущественно на автосегментно-метрическую фонологическую модель интонации [21; 2; 1], просодия описывается с точки зрения выделяемых в том или ином языке просодических единиц и ассоциированных с ними тональных явлений. Просодические системы языков мира могут — среди прочего — отличаться друг от друга тем, какая из просодических единиц является базовой для реализации тональных явлений: мора, слог (напр., китайский), стопа, лексическое слово (сербохорватский), фонетическое слово (новогреческий), акцентная группа (японский), синтагма (английский, немецкий, русский литературный), фраза [8: 434–435].

³ Этот говор среди прочих говоров верхней Пинеги и Выи по материалам экспедиций 1928 и 1929 гг. был описан в [15], по материалам экспедиции 1987 г. – в [14]. Подробное описание его просодической системы см. в [13], корпус звучащих текстов – [10].

(ЧОТ); ниже эти отрезки условно называются синтагмами. Смена просодического оформления в пинежском говоре очень хорошо видна на интонограммах. Основной тип мелодического контура в нем выглядит следующим образом: на начальных безударных слогах имеет место средний (базовый) начальный пограничный тон (%M), на первом ударном слоге наблюдается значительный подъем до высокого (или сверхвысокого), который сохраняется до первого заударного слога конечного в синтагме акцентированного слова с последующим падением к среднему (в случае смысловой незавершенности) или низкому (при завершенности) пограничному тону (M% или L%). Этот мелодический контур ('hat', 'шляпа') является одним самых распространенных в языках с фразовой просодией⁴. Ниже приведены примеры тонального оформления

- одной синтагмы с конечным средним пограничным тоном (рис. 3);
- одной синтагмы с конечным низким пограничным тоном (рис. 4);
- двух синтагм без физической паузы между ними (рис. 5; просодический шов, разделяющий их, находится между конечным M% и начальным %M пограничными тонами: *Испечёшь да и // помажешь масличком да и*);
- трех синтагм, разделенных паузами – на рис. 6;
- пяти синтагм на рис. 7 (после первой и второй имеют место физические паузы продолжительностью около 0,5 и 1 сек.; третья, четвертая и пятая разделены только сменой просодического оформления).

2. В дальнейшем на каждом отдельном слове было измерено изменение ЧОТ – либо на ударном слоге, либо (преимущественно в случае наличия глухого согласного в инициали ударного слога и явного различия в значении ЧОТ на предударном и ударном гласном) – на гласных ударного и предударного слогов. Отдельными считались такие слова, которые в данных говорах могут нести на себе тональный акцент. На этом основании, например, *вот, там, я, он* и т. п. являются отдельными словами; *на, от, ведь, дак, же* и т.п. – нет. Пример тонального акцента на слове *вот* см. на рис. 6 и 7. Общее число слов составило 573 в архангельском говоре (в среднем – 5,7 слов в синтагме) и 583 в вологодском (в среднем – 5,8 слов в синтагме).

3. Далее все полученные данные были распределены по двум основным группам: 1) слова, несущие на себе тональный акцент, и 2) слова, на которых тональный акцент отсутствует, что представляло собой не вполне тривиальную задачу. Так, на интонограмме фразы *Лони у меня было четыре дорожки* (5 фонетических слов, рис. 4) хорошо видно, что на словах *лони* и *четыре* тональный акцент явно присутствует (изменение тона на их ударных гласных составляет 85 и 39 Гц соответственно), в словах *у меня* и *дорожки* – отсутствует (7 и 5 Гц соответственно), а слово *было* представляет собой пограничный случай (17 Гц).

В современной интонологии принято считать, что надежной границей, позволяющей отличить наличие тонального акцента от его отсутствия, является изменение в 3 полутона (ST). J. 't Haart, исследовавший перцептивную значимость величины изменения ЧОТ, отмечал:

"Some listeners requiring only 1.5-2.0 semitones (ST), others apparently as much as 4.0 ST. The outcome suggests that only differences of more than 3 semitones play a part in communicative situations" [6: 811]; см. также [7: 29].

Однако другие исследования показывают, что и изменения в 1,5 полутона могут быть достаточно надежными показателями просодического выделения при условии, что они имеют место в одном и том же типе тонального контура:

"This finding warrants the conclusion that a difference of 1.5 ST is sufficient to create an impression of a difference in prominence, provided the contour within which the difference is embedded is kept constant" [23: 304].

На этом основании мы считали, что тональный акцент на том или ином слове в синтагме отсутствует, если изменение ЧОТ составляло менее 1,5 полутона; иными словами – менее 7% от исходного значения ЧОТ в Гц:

⁴ Ср.: «In Dutch, over 60 % of pitch contour tokens one encounters in everyday speech are realizations of a single basic intonation pattern, the so-called "hat pattern"» [18: 653].

“Our results also indicate that the prominence judgements made by our listeners showed better agreement with a Hertz scale than with a semitone scale” [23: 305]⁵.

4. Затем слова, несущие на себе тональный акцент, были распределены по группам в зависимости от

- величины изменения ЧОТ – от 1,5 до 3 ST, от 3 до 5 ST, более 5 ST (7-15%, 15-30% и более 30% от исходного значения ЧОТ в Гц соответственно)
- направления изменения ЧОТ (восходящее или нисходящее);
- типа тонального акцента (например, в говоре Вадюги сосуществуют восходящие акценты L+H* и L*+H, отличающиеся более ранним или более поздним таймингом [13]: в первом случае восходящее движение тона начинается в начале ударного гласного или на согласном слоговой инициали, во втором – в середине ударного гласного, см. рис. 8; в говоре Евсеевской имеют место нисходящие акценты H+L* (154 примера из 175) и H*+L (21 из 175), также отличающиеся таймингом: нисходящее движение начинается в начале ударного гласного или в его середине соответственно⁶, см. рис. 9);
- наличия в его составе ровного тона наряду с восходящим или нисходящим: акцент L*+H всегда имеет ровный тон в начале, L+H* – часто в конце (подробнее об этом см. [12]).

5. Наконец, на заключительном этапе из 100 синтагм архангельского диктора были выбраны 25, которые в качестве основания для сравнения были прочитаны носителем современного русского литературного языка (далее — СРЛЯ); в его произношении синтагм оказалось 27 (127 слов). Описанная выше процедура была повторена и на этом материале.

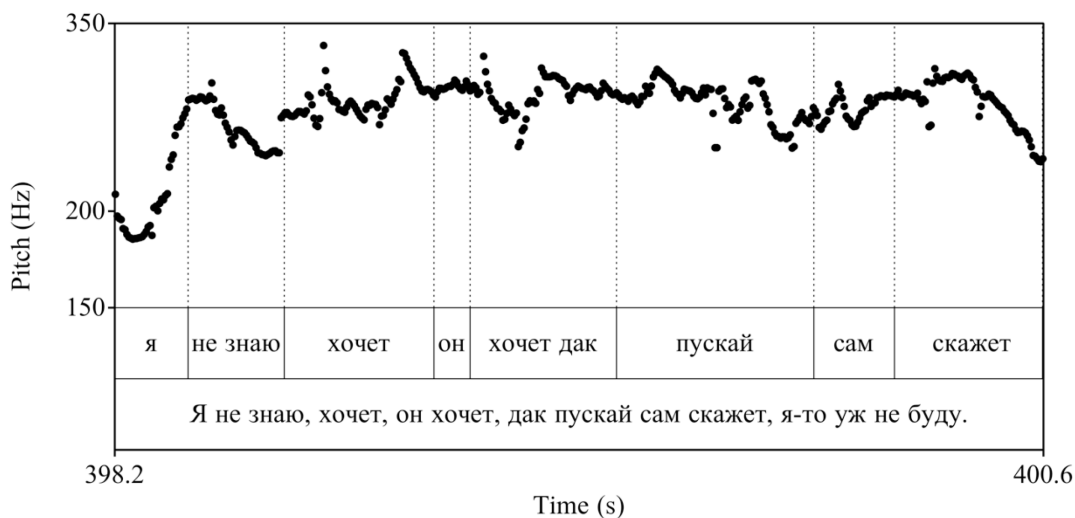


Рисунок 3: Кривая ЧОТ фразы

Я не знаю, хочет, он хочет, дак пускай сам скажет

(д. Вадюга Верхнетоемского р-на Архангельской области, диктор — Р. А. О.)

⁵ Представляется необходимым отметить в этой связи, что при такой методике выделения в число тональных акцентов не попадают акценты типа H* (ровный высокий тон на ударном гласном акцентоносителя [13]), в случае высокого тона на предупредном и глухом согласном между ним и ударным гласным; к счастью, количество таких случаев было в нашем материале пренебрежимо мало (все они были исключены из дальнейшего рассмотрения).

⁶ Эти акценты в дальнейшем рассматриваются совместно. Аналогичным образом, для целей настоящего исследования нам представлялось излишним выделять различные типы восходящих акцентов в Вадюге в зависимости от времени завершения повышения тона (в начале ударного гласного, в его середине или конце); этой проблематике посвящено отдельное исследование. Косвенным образом данная информация может быть получена из графы «наличие ровного тона» в Таблице 1 (см. ниже): ровный тон в составе акцента имеет место в первых двух случаях и отсутствует в третьем.

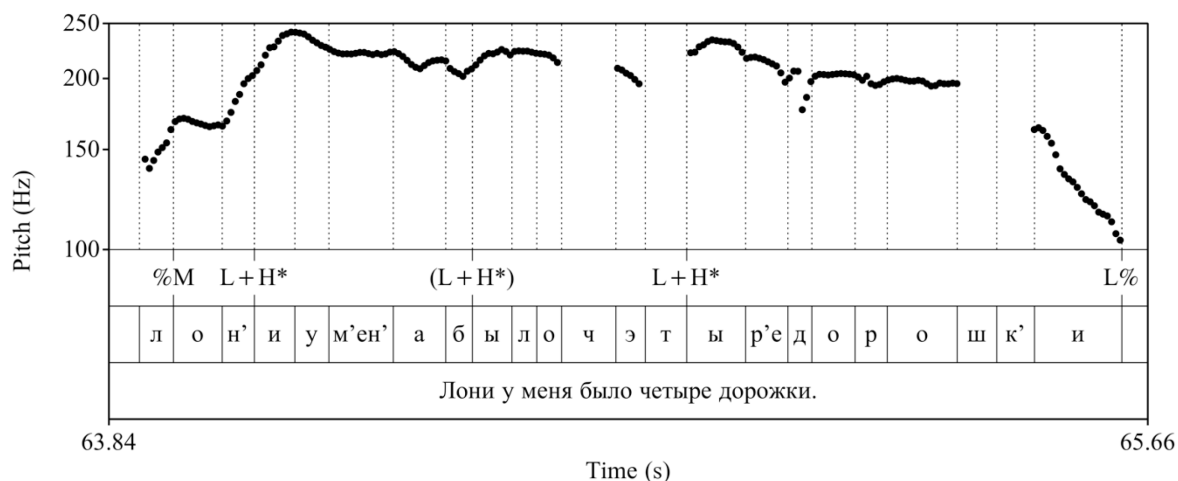


Рисунок 4: Кривая ЧОТ фразы *Лони у меня было четыре дорожки* (д. Ваймуша Пинежского р-на Архангельской области)

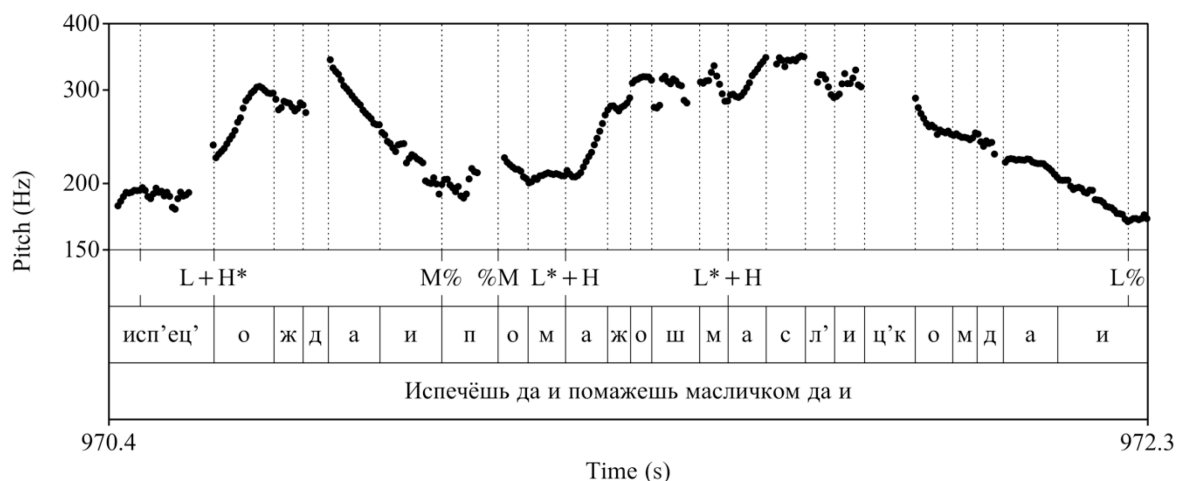


Рисунок 5: Кривая ЧОТ фразы *Испечёшь да и помажешь масличком да и* (д. Вадюга Верхнетоемского р-на Архангельской области, диктор — Р. А. О.)

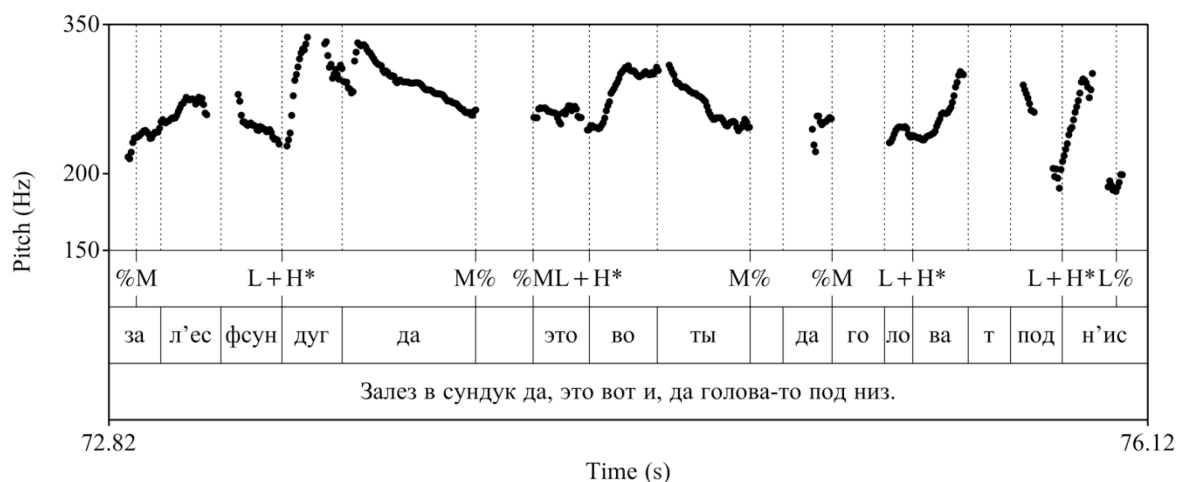


Рисунок 6: Кривая ЧОТ фразы *Залез в сундук да, это вот и, да голова-то под низ* (д. Вадюга Верхнетоемского р-на Архангельской области, диктор — Р. А. О.)

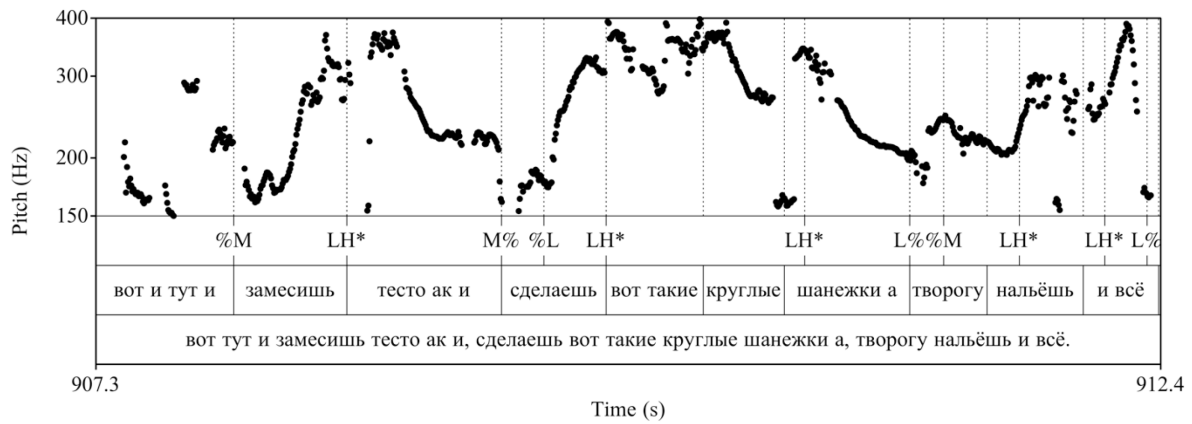
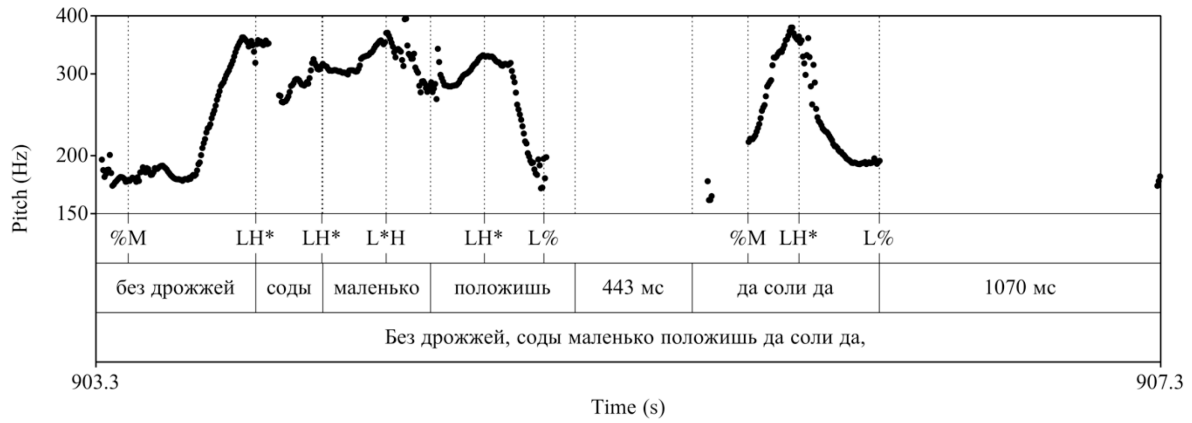


Рисунок 7: Кривая ЧОТ фразы *Без дрожжей, соды маленько положишь // да соли да, // вот тут и замесишь тесто ак и, // сделаешь вот такие круглые шанежки а, // творогу нальёшь и всё* (д. Вадюга Верхнетоемского р-на Архангельской области, диктор — Р. А. О.)

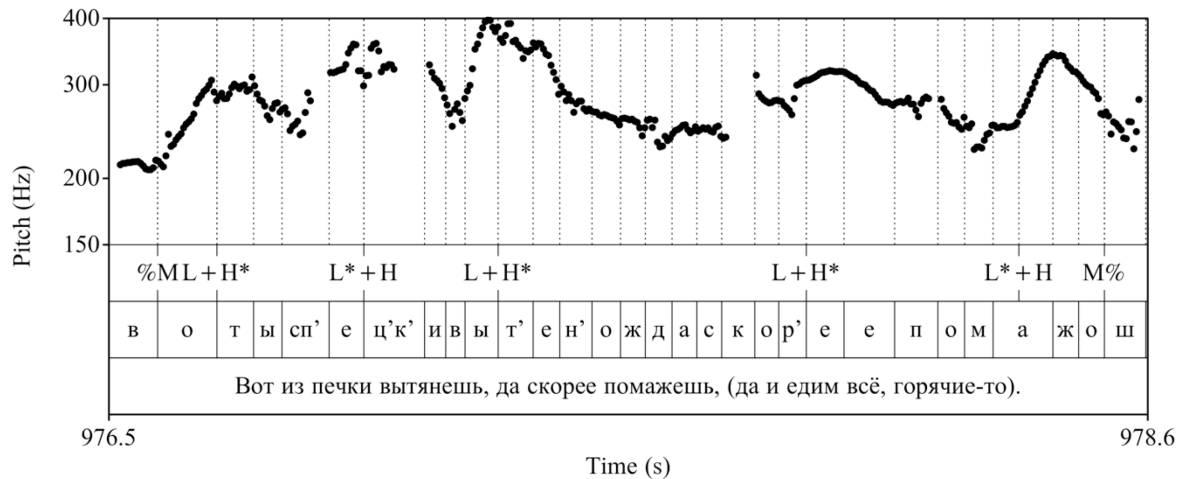


Рисунок 8: Кривая ЧОТ фразы *Вот из печки вытянешь, да скорее помажешь* (д. Вадюга Верхнетоемского р-на Архангельской области, диктор — Р. А. О.)

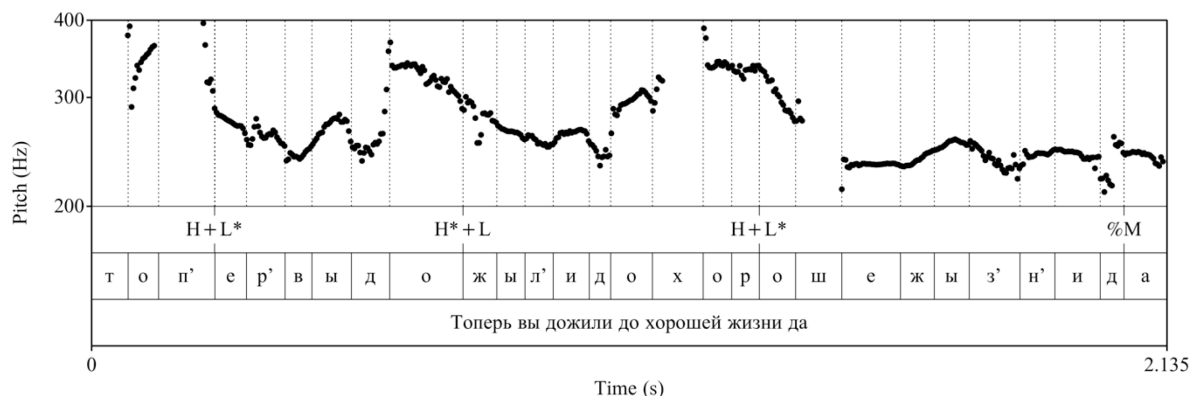


Рисунок 9: Кривая ЧОТ фразы
Теперь вы дожили до хорошей жизни да
 (д. Евсеевская Тарногского р-на Вологодской области, диктор — К. З. Я.)

5 Результаты исследования

Основные результаты исследования приведены ниже в Таблице 1 и на Рисунке 10.

	Арх.	%	Влг.	%	ЛЯ	%
синтагм	100		100		27	
слов	573		583		127	
акцентов	342		367		40	
акцентов в синтагме	3,4		3,7		1,5	
акцентов / слов	0,6		0,63		0,3	
1,5-3 ST	124	36%	118	32%	18	45%
3-5 ST	149	44%	153	42%	18	45%
5+ ST	69	20%	96	26%	4	10%
нисходящих	9	2,6%	206	56%	21	53%
восходящих	333	97,4%	161	44%	19	47%
наличие ровного тона	294	86%	121	33%	0	0
L+H*	262	77%	158	43%	19	47%
L*+H	71	21%	3	1%	0	0
H+L*	5	1%	175	47%	21	53%
H*+L	4	1%	31	9%	0	0
базовый тон	Н		М		М	
L%	90	90%	49	75%		
M%	10	10%	16	25%		
%M	214 Гц		222 Гц		220 Гц	
L%	198 Гц		192 Гц			
H%	265 Гц		249 Гц			
Н	255		264 Гц		260 Гц	
^Н	290		314 Гц		300 Гц	

Таблица 1: Количественные результаты: СРЛЯ, архангельский и вологодский говоры

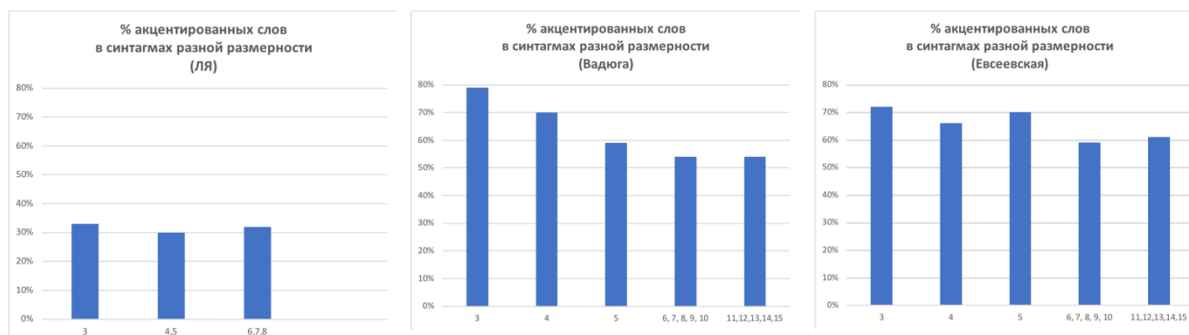


Рисунок 10: Процент акцентированных слов в синтагмах разной размерности в СРЛЯ, архангельском и вологодском говорах

6 Выводы

Полученные результаты дают основания для следующих заключений:

1. Верхнепинежский и тарногский говоры действительно отличаются от литературного языка большей частотностью тональных акцентов: если в СРЛЯ тонально выделены, в среднем, каждые три слова из десяти⁷, то в обследованных севернорусских говорах – шесть; тем самым, акцентированных слов в говорах с «пословным» мелодическим контуром вдвое больше, чем в СРЛЯ.

2. При этом тональные акценты в исследованных говорах фиксируются далеко не на каждом слове, способном нести акцент, а только на приблизительно 60% таких слов; таким образом, следует, по-видимому, говорить, что в этих диалектных системах не «почти каждое слово во фразе получает свое мелодическое оформление» [19: 64], а лишь почти каждое второе. Эти данные подтверждают тот факт, что в подобных говорах основной просодической единицей наименьшей размерности является не (фонетическое) слово, а акцентная группа [13]. Как в архангельском, так и в вологодском говорах – в отличие от СРЛЯ – имеет место тенденция к уменьшению отношения выделенных слов к невыделенным с увеличением количества слов в синтагме: чем длиннее синтагма, тем меньше в ней процент слов, маркированных изменением тона (см. рис. 10).

3. И вадюжский, и евсеевский говоры отличаются от литературного языка регулярным наличием ровного тона на выделенном гласном, что подтверждает вывод о том, что он в севернорусских говорах может являться частью тонального акцента [12]. При этом в верхнепинежском говоре ровный тон в составе тонального акцента значительно частотнее, чем в тарногском (86% и 33% соответственно). По всей вероятности, именно наличие ровного тона после восходящего (без немедленного падения) позволяет носителям СРЛЯ не воспринимать утвердительные предложения с восходящим тоном в пинежском говоре как вопрос.

4. Очередным отличием диалектных просодических систем от литературной является значительно большая частотность тональных акцентов с увеличенным интервалом (20% и 26% против 10% СРЛЯ).

5. Существенное различие между архангельским говором и вологодским заключается в соотношении восходящих и нисходящих тональных акцентов: если в говоре д. Вадюги нисходящие акценты крайне редки (2,6% всех случаев), то в говоре д. Евсеевской они составляют более половины всех случаев (56%), что – в этом отношении – сближает данный говор с литературным русским языком (53%). Таким образом, «пословный» мелодический контур действительно существует, как минимум, в двух разновидностях: с восходящим тональным движением и с нисходящим тоном в функции орнаментального акцента [11].

6. Ещё одним отличием между двумя обследованными просодическими системами является уровень базового тона, на котором происходят основные тональные изменения: в верхнепинежском говоре он высокий, в тарногском, как и в СРЛЯ, – средний. Можно предположить, что именно эта особенность говоров Верхней Пинеги провоцирует восприятие каждого слова в нем как тонально выделенного.

⁷ Эти данные совпадают с существующими в литературе сведениями о том, что общее число акцентных выделений в свободно подаваемых репликах составляет 37% в официальных и 25% в непринужденных высказываниях [5], [17].

7. В целом, интонационная система вологодского говора, при всех существенных отличиях, значительно ближе литературной русской, чем просодической системе архангельского говора.

References

- [1] Beckman M. E., Hirschberg J., Shattuck-Hufnagel S. The Original ToBI system and the evolution of the ToBI framework. *Prosodic typology. The phonology of intonation and phrasing*. Oxford, Oxford Univ. Press, 2005. P. 9–54.
- [2] Beckman M. E., Pierrehumbert J. Intonational structure in Japanese and English. *Phonology yearbook*. Cambridge University Press, Vol.3, 1986. P. 15–70.
- [3] Bryzgunova E. A. An analysis of Russian dialectal intonation [Analiz russkoy dialektnoy intonacii] In: *Experimental phonetic studies in the field of Russian dialectology [Eksperimentalno-foneticheskie issledovaniya v oblasti russkoi dialektologii]*. Moscow, Nauka, 1977. P. 231–262.
- [4] Evstigneeva M. Y. Intonation contour of phrases with the semantics of completeness in one Northern Russian dialect [Intonacionnoe oformlenie fraz s semantikoi zavershennosti v odnom severnorusskom govore]. In: *Actual problems of Russian dialectology. Materials of the International Conference on October 29-31, 2021 [Aktualnye problemy russkoi dialektologii. Materialy Mezhdunarodnoi konferencii 29–31 oktyabrya 2021 g.]*. Moscow, Vinogradov Russian Language Institute, 2021. P. 76–78.
- [5] Evchik N. S. Classification of accented syllables in spontaneous statements [Klassifikaciya vydelennykh slogov v spontannykh vyskazyvaniyakh]. In: *Romance linguistics [Romanskoe yazykoznanie]*. Minsk, No 1, 1978. P. 43–50.
- [6] Hart, J. Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America*. Vol. 69 (3), 1981. P. 811–821.
- [7] Hart, J., R. Collier and A. Cohen. *A Perceptual Study of Intonation. An Experimental-phonetic Approach to Speech Melody*. Cambridge University Press, 1990.
- [8] Jun S.-A. *Prosodic typology. The phonology of intonation and phrasing*. Oxford University Press, 2005. P. 430–458.
- [9] Kasatkina R. F. On the phonetic nature of word stress in Northern Russian dialects [O foneticheskoy prirode slovesnogo udareniya v severnorusskikh govorakh]. In: *Modern Russian dialects [Sovremennye russkie govory]*. Moscow, Nauka, 1991. P. 42–49.
- [10] Knyazev S. V. *Corpus of the Russian dialect spoken in the basins of Upper Pinega and Vyya rivers*. 2021. Moscow: Linguistic Convergence Laboratory, HSE University. (Available online at URL: <http://lingconlab.ru/vaduga/>, accessed on 15.01.2022.)
- [11] Knyazev S. V. Word-by-word melodic contour and one type of incompleteness marking in Tserkovnoe dialect, Arkhangelsk Oblast [Poslovnyi melodicheskii kontur i odin iz sposobov oformleniya nezavershennosti v govore s. Cerkovnoe Arxangelskoi obl.]. In: *Proceedings of the V. V. Vinogradov Russian Language Institute [Trudy Instituta russkogo yazyka imeni V. V. Vinogradova]*, Vol. 2 (27), 2021. P. 39–65.
- [12] Knyazev S. V. Structure of the pitch accent in Russian dialects with word-by-word melodic contour [O strukture tonalnogo akcenta v russkikh govorakh s poslovnym melodicheskim oformleniem]. In: *Russian Language and Linguistic Theory [Russkij yazyk v nauchnom osveshhenii]*. Vol. 43, №1, 2022.
- [13] Knyazev S. V. Sentence intonation in Russian dialects with word-by-word melodic contour [O frazovoy prosodii v russkikh govorakh s poslovnym melodicheskim oformleniem]. In: *Topics in the Study of Language [Voprosy yazykoznanija]*, 2022, № 1. P. 7–39.
- [14] Knyazev S. V., Levina A. N., Pozharitskaya S. K. On Upper Pinega and Vyya dialects [O govorax Verxney Pinegi i Vyi]. In: *Russian dialects: history and present state. Problems of Russian linguistics. VII [Russkie dialekty: istoriya i sovremennost']*. Problemy russkogo yazykoznanija]. Moscow, Moscow State University Publ., 1997. P. 198–220.
- [15] Kuznetsov P. S. On Upper Pinega and Upper Toyma dialects [O govorax Verxney Pinegi i Verxney Toymy]. In: *Materials and studies in Russian dialectology. Vol. 1 [Materialy i issledovaniya po russkoy dialektologii. Tom 1]*. Moscow – Leningrad, 1949. P. 5–44.
- [16] Nikolaeva T. M. *Phrase intonation in Slavic languages [Frazovaya intonatsiya slavyanskikh yazykov]*. Moscow, Nauka, 1977.

- [17] Nikolaeva T. M. Semantics of accentuation [Semantika akcentnogo vydeleniya]. Moscow, Nauka, 1982.
- [18] Nootboom S.G. The prosody of speech: melody and rhythm. In: The Handbook of Phonetic Sciences. Oxford, Basil Blackwell Limited, 1997. P. 640–673.
- [19] Paufoshima R. F. Phonetics of word and phrase in Northern Russian dialects [Fonetika slova i frazy v severnorusskix govorax]. Moscow, Nauka, 1983.
- [20] Paufoshima R. F. Traces of musical stress in modern Vologda dialect [Sledy muzykalnogo udareniya v sovremennom vologodskom govore]. In: Dialectography of Russian language [Dialektografiya russkogo yazyka]. Moscow, Nauka, 1985. P. 94–102.
- [21] Pierrehumbert J. B. The phonology and phonetics of English intonation. Ph.D. diss. Massachusetts Institute of Technology, 1980.
- [22] Post M. The Northern Russian pragmatic particle *dak* in the dialect of Varzuga (Kola Peninsula). An information structuring device in informal spontaneous speech. Doctoral diss., Universitetet i Tromsø, 2005.
- [23] Rietveld, A.C.M. and C. Gussenhoven. On the relation between pitch excursion size and prominence. Journal of Phonetics. Elsevier, Vol. 13, 1985. P. 299–308.

Knowledge Distillation of Russian Language Models with Reduction of Vocabulary

Alina Kolesnikova*¹

kolesnikova.af@phystech.edu

Vasily Kononov¹

vaskoncv@phystech.edu

Yuri Kuratov*^{1,2}

yurii.kuratov@phystech.edu

Mikhail Burtsev^{1,2}

burtcev.ms@mipt.ru

¹Neural Networks and Deep Learning Lab, MIPT, Dolgoprudny, Russia

²AIRI, Moscow, Russia

Abstract

Today, transformer language models serve as a core component for majority of natural language processing tasks. Industrial application of such models requires minimization of computation time and memory footprint. Knowledge distillation is one of approaches to address this goal. Existing methods in this field are mainly focused on reducing the number of layers or dimension of embeddings/hidden representations. Alternative option is to reduce the number of tokens in vocabulary and therefore the embeddings matrix of the student model. The main problem with vocabulary minimization is mismatch between input sequences and output class distributions of a teacher and a student models. As a result, it is impossible to directly apply KL-based knowledge distillation. We propose two simple yet effective alignment techniques to make knowledge distillation to the students with reduced vocabulary. Evaluation of distilled models on a number of common benchmarks for Russian such as Russian SuperGLUE, SberQuAD, RuSentiment, ParaPhaser, Collection-3 demonstrated that our techniques allow to achieve compression from $17\times$ to $49\times$, while maintaining quality of $1.7\times$ compressed student with the full-sized vocabulary, but reduced number of Transformer layers only. We make our code and distilled models available.

Keywords: language modeling, transformer, knowledge distillation, compact models, Russian

DOI: 10.28995/2075-7182-2022-21-295-310

Дистилляция знаний для русскоязычных моделей с уменьшением словаря

Алина Колесникова*¹

kolesnikova.af@phystech.edu

Василий Коновалов¹

vaskoncv@phystech.edu

Юрий Куратов*^{1,2}

yurii.kuratov@phystech.edu

Михаил Бурцев^{1,2}

burtcev.ms@mipt.ru

¹Лаборатория нейронных систем и глубокого обучения, МФТИ,

Долгопрудный, Россия

²Институт искусственного интеллекта AIRI, Москва, Россия

Аннотация

На текущий момент языковые модели типа Трансформер являются основным компонентом для большинства задач обработки естественного языка. Промышленное применение таких моделей требует минимизации времени вычислений и объема памяти. Дистилляция знаний - один из подходов к решению этой задачи. Существующие методы в этой области в основном ориентированы на уменьшение количества слоев или размерности эмбедингов/скрытых состояний. Другой способ - уменьшить количество токенов в словаре и, следовательно, матрицу эмбедингов модели-студента. Основной проблемой, которая возникает при уменьшении размера словаря, является несоответствие между входными последовательностями и предсказываемыми распределениями классов моделями учителя и студента. В результате невозможно напрямую применить дистилляцию знаний на основе KL. Мы предлагаем два простых и в тоже время эффективных метода выравнивания, чтобы применить дистилляцию знаний в студента с уменьшенным словарем. Оценка дистиллированных моделей на нескольких распространенных русскоязычных бенчмарках, таких как Russian SuperGLUE, SberQuAD, Rusementiment, ParaPhaser, Collection-3 показала, что предложенные методы позволяют сжать модель от 17 до 49 раз, сохраняя при этом качество модели-студента с полноразмерным словарем и уменьшенным количеством Трансформер-слоев, сжатой в 1.7 раз. Дистиллированные модели и код выложены в открытый доступ.

Ключевые слова: языковое моделирование, трансформер, дистилляция знаний, легковесные модели, русский язык

1 Introduction

Pre-trained Transformer language models have been found to be very successful across a wide range of NLP tasks. Most of the recent state-of-the-art models are based on variations of the original Transformer (Vaswani et al., 2017) and different self-supervised pre-training techniques like masked language modeling (Devlin et al., 2019). Such models became very large, starting from hundreds of millions of parameters (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019) to hundreds of billions (Brown et al., 2020; Smith et al., 2022; Rae et al., 2021; Lin et al., 2021). Large models require lots of computation, memory, and fast accelerators like TPUs/GPUs. It is challenging to use large models in practical applications where prediction time is critical and available disk/RAM is limited.

General approaches like pruning, quantization, and knowledge distillation (KD) were applied to Transformer language models to make them faster and smaller. Pruning (LeCun et al., 1989) removes some weights of the large models with negligible degradation of predictions. Quantization (Gong et al., 2014) reduces weights precision to float16, int8, int4, or even bits. Knowledge distillation (Buciluă et al., 2006; Ba and Caruana, 2014; Hinton et al., 2015) (KD) is used to train smaller student model to mimic behaviour of the larger teacher model.

However, in general, knowledge distillation relies on Kullback-Leibler (KL) divergence over teacher and student predictions. Language models are trained to predict tokens probability distribution in a vocabulary. It implies that teacher and student should share the same vocabulary. If a teacher and student models have different vocabularies, KL loss can not be directly applied as they have different sets of prediction classes. It makes KL-based knowledge distillation for models with mismatched vocabularies impossible. Another outcome of mismatched vocabularies is different tokenization for teacher and student models. It leads to different lengths of input and, therefore, output sequences, which also adds ambiguity to KL-based distillation in this case.

A ratio of embeddings parameters becomes larger as student models become smaller by reducing the number of Transformer layers and/or dimension of hidden representations. Embeddings can get over 50% of all parameters for small models as shown on Figure 1.

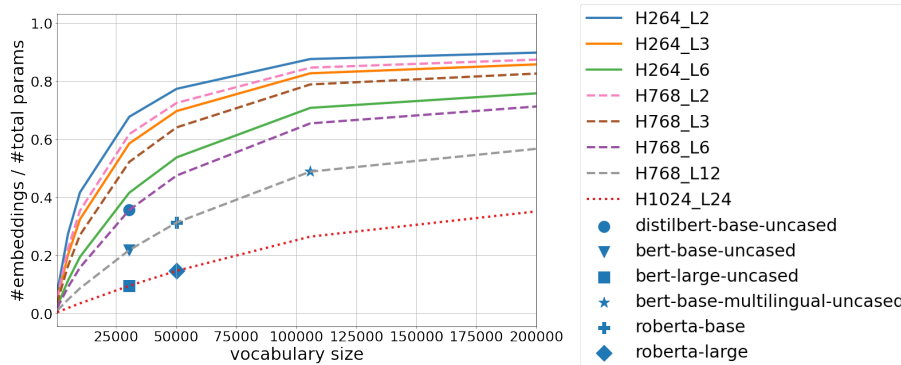


Figure 1: Ratio of number of parameters for embeddings to the full model. In smaller models embeddings have higher fraction of parameters compared to other models with the same vocabulary size. Selected models for English language are shown with the blue markers. The models are denoted by size of hidden representation (H) and number of layers (L).

One of the possible ways to reduce a fraction of embeddings parameters is to make the size of student vocabulary smaller. Moreover, changing student vocabulary could be reasonable for distilling to another domain or from multilingual to monolingual models. Changing student vocabulary leads to the problem of knowledge distillation with mismatched vocabularies.

This paper focuses on applying knowledge distillation to train student models with a smaller vocabulary than the teacher. We propose several strategies for output/intermediate representations alignment. The first one uses teacher and student representations corresponding to the tokens found in both vocabu-

laries (*match* strategy). The second aligns the sequences, produced by student tokenizer, with the teacher by aggregating representations corresponding to an alignment (*reduce* strategy).

We show that teacher’s knowledge can be effectively transferred to the student with mismatched vocabulary. We pre-train student models with proposed KD methods and evaluate them on a number of common benchmarks for the Russian language such as Russian SuperGLUE, SberQuAD, RuSentiment, ParaPhraser, and NER on Collection-3. Our students are from $17\times$ to $49\times$ and up to $104\times$ faster on GPU than the teacher while having competitive quality to the $1.7\times$ compressed student. We make our code¹ and pre-trained models² available online.

2 Related work

Knowledge distillation can be used to train task-specific fine-tuned and general pre-trained models. Task-specific distillation (Chia et al., 2018; Sun et al., 2019; Tang et al., 2019; Aguilar et al., 2020) takes two steps: fine-tuning a teacher model on a task and distilling it to a student model. The disadvantage of such approach is that it requires repeating both steps for each new task. Large teacher model fine-tuning could be too expensive.

Such models as DistilBERT (Sanh et al., 2019), TinyBERT (Jiao et al., 2020), MobileBERT (Sun et al., 2020b), MiniLM (Wang et al., 2020; Wang et al., 2021) use general pre-training distillation. Distillation could be performed only once to pre-train a general student model, and then student model fine-tunes on tasks, removing expensive teacher fine-tuning step. DistilBERT uses a triple loss: distillation loss between student and teacher output probabilities, student masked language modeling loss, and cosine loss for hidden representation of student and teacher models. TinyBERT adds trainable student-teacher projections for embeddings and Transformer layer output representations. These projections allow training student models with Transformer layer hiddens of arbitrary size. TinyBERT, MobileBERT, MiniLM use attention matrices as an additional source of knowledge for distillation. A student model trains to produce similar attention matrices to a teacher by additional loss term.

However, previously mentioned pre-training knowledge distillation approaches are not flexible enough. Student model vocabulary should be the same as a teacher model to compute the distillation loss. Different vocabularies also lead to different tokenization, hence different student and teacher sequence lengths. Therefore, student-teacher output probabilities, hiddens and attention matrices are not aligned to be used with losses mentioned above.

(Zhao et al., 2021) addresses these problems with mixed-vocabulary training. Authors propose first to pre-train student embedding matrix together with teacher model and then use it for regular student model MLM pre-training. Tokenization for each word in mixed-vocabulary training is performed by randomly selecting teacher or student vocabulary with corresponding embeddings matrix. This way, only the embeddings matrix is trained using teacher model knowledge. All other parameters of the smaller student model do not benefit from the teacher model. In our work, we propose methods that allow training student model with knowledge distillation in one stage and using teacher knowledge from all layers.

The problem with mismatched vocabulary is actually more general. Vocabulary tokens are essentially labels for a token classification task, that is, language modeling. In other words, the more general problem is knowledge distillation for teacher and student models with different sets of labels. Instead of predictions distillation, pre-classification layer outputs or other representations might be used for distillation (Tian et al., 2020; Sun et al., 2020a), making a connection to representation-based learning (Bromley et al., 1993; Chen et al., 2020).

Alternatively, the number of parameters in Transformer models can be reduced by parameters sharing (Lan et al., 2020), embeddings matrix factorization (Sun et al., 2020b; Hrinchuk et al., 2019; Lan et al., 2020), and pruning (Voita et al., 2019; Gordon et al., 2020). These approaches are complementary to knowledge distillation in general and to our methods as well.

¹github.com/ayeffkay/rubert-tiny

²See models with `distil-` prefix at huggingface.co/DeepPavlov

3 Distillation strategy

3.1 Background

One of the first attempts of pre-trained Transformer language model distillation is DistilBERT (Sanh et al., 2019). Authors introduce training objective which is a linear combination of the supervised masked language modeling (MLM) loss:

$$\mathcal{L}_{mlm} = - \sum_{i=1, i \in \text{masked_ids}}^{|X_t|} \sum_{j=1}^{|V_t|} y_{ij} \log p_{ij}^s, \quad (1)$$

distillation loss over the soft target probabilities of the teacher:

$$\mathcal{L}_{ce} = - \sum_{i=1, i \notin \text{masked_ids}}^{|X_t|} \sum_{j=1}^{|V_t|} p_{ij}^t \log p_{ij}^s, \quad (2)$$

and cosine distance loss for the student and teacher hidden representations:

$$\mathcal{L}_{cos} = \sum_{i=1, i \notin \text{masked_ids}}^{|X_t|} \text{cos_dist}(h_{n,i}^t, h_{m,i}^s), \quad \text{cos_dist}(h_{n,i}^t, h_{m,i}^s) = 1 - \frac{\langle h_{n,i}^t, h_{m,i}^s \rangle}{\|h_{n,i}^t\| \|h_{m,i}^s\|}, \quad (3)$$

here `masked_ids` is a set of subword indices, masked with some probability; $|X_t|$ is a subwords sequence length obtained after input sequence tokenization by teacher tokenizer; V_t is a teacher vocabulary with the size $|V_t|$; y_{ij} is a masked subword index in a vocabulary; p^s , p^t are subword probabilities produced by student and teacher models; h_m^s , h_n^t are student and teacher hidden states taken from m -th and n -th Transformer layers. Alternatively Kullback-Leibler divergence can be used instead of \mathcal{L}_{ce} .³

Usually, it is assumed that a teacher and a student use the same vocabulary, i.e. inputs for the teacher and the student will match after tokenization. But if a teacher and a student use different vocabularies, then tokenized inputs will be different and will not always have the same length. Further we represent the problem statement more formally and provide our solutions.

3.2 Problem statement

Given teacher with vocabulary V_t and student with vocabulary V_s , such that $|V_s| < |V_t|$, $V_s \cap V_t \neq \emptyset$.⁴ Then LM output probabilities shapes will be $(|X_t|, |V_t|)$, $(|X_s|, |V_s|)$ and hidden states shapes will be $(|X_t|, d_t)$, $(|X_s|, d_s)$ for teacher and student respectively, where X_t and X_s are inputs produced after tokenization by teacher and student tokenizers, d_s and d_t are hidden states dimension. As mentioned above, in general $X_t \neq X_s$ and $V_t \neq V_s$. The task is to define alignment $\mathcal{X} : X_s \rightarrow X_t$ for sequence length dimension to obtain $|X_s| = |X_t|$ and mapping $\mathcal{V} : V_s \rightarrow V_t$ between vocabularies.

For simplicity we will assume that $|X_t| \leq |X_s|$. The rationale behind this lies in the observation that because of reduced vocabulary size $|V_s|$ BPE tokenization algorithm will keep less amount of more frequent subwords, thus leading to longer student-generated outputs. Our observation confirmed when we compared sequence lengths produced by teacher and student pre-trained tokenizers with $\sim 1.2 \times 10^5$ and $\sim 3 \times 10^4$ vocabulary sizes respectively. On the corpus of $\sim 2.7 \times 10^7$ sequences only 0.2% of the student-tokenized sequences were shorter than the teacher.

³The difference will be in the term $p_{ij}^t \log(p_{ij}^t)$.

⁴We emphasize that *non-empty intersection condition between teacher and student vocabularies is necessary*, because the strategies below cannot be applied in the case of its complete mismatch.

3.3 Sequence length dimension alignment

We propose two strategies for sequence length dimension alignment: *match* strategy and *reduce* strategy. The first one can be applied *both* to the sequence length and vocabulary dimension, the second one for sequence length dimension only.

For the *match* strategy, after building student vocabularies of sizes 5×10^3 , 10^4 , 2×10^4 , 3×10^4 using BPE subword tokenization algorithm we found that $\sim 99\%$ student subwords are in the teacher vocabulary of $\sim 1.2 \times 10^5$ size. Therefore, we can take into account only matching subwords and mask all mismatched subwords (Figure 2)

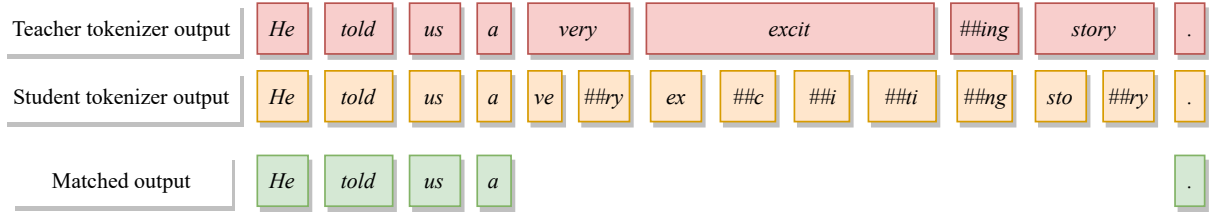


Figure 2: Match strategy for sequence length dimension alignment. The first sequence is produced by the teacher’s tokenizer, and the second by the student’s.

If n_{match} subwords in sequence and $|V_{match}|$ subwords in vocabulary match, then hidden states and output LM probabilities shapes are transformed as follows:

$$\begin{aligned} (|X_t|, d_t) &\mapsto (n_{match}, d_t), & (|X_t|, |V_t|) &\mapsto (n_{match}, |V_{match}|), \\ (|X_s|, d_s) &\mapsto (n_{match}, d_s), & (|X_s|, |V_s|) &\mapsto (n_{match}, |V_{match}|). \end{aligned} \quad (4)$$

This makes sequences equal by length and aligned for teacher and student models. LM output probabilities also have equal shapes. KL or CE losses can be used for distillation now with *match* strategy.

It can be seen that *match* strategy lowers overhead required to compute losses. The main drawback is that we lose from 75% (for 3×10^4 vocabulary size) to 96% (for 5×10^3 vocabulary size) subwords that can be used for distillation from the teacher. Another drawback is that embeddings corresponding to the matching subwords might occur in different contexts for teacher and student and thus might cover different meanings.

In general the task of finding correspondence between teacher- and student-tokenized sequences is ambiguous. For example in Figure 2 depending on the tokenizer we can obtain highly mismatched subword sequences:

- (5) (Teacher) *excit ##ing*
 (Student) *ex ##c ##i ##ti ##ng*

In *reduce* alignment strategy an auxiliary input for a student model receives teacher subwords greedily split into student subwords from left to right. Then student’s intermediate/output representations corresponding to the one teacher subword are aggregated by summation as shown on Figure 3. Assume that i -th teacher subword was splitted by student subwords with indices $k_1^i, k_2^i, \dots, k_l^i$. Then formally, aggregation procedure for hidden states can be written as follows:

$$h_i^t = \sum_{j \in \{k_1^i, \dots, k_l^i\}} h_j^s, \quad i = \overline{1, |X_t|} \quad (6)$$

Pre-softmax outputs aggregation procedure can be represented in a similar way.

This allows the student to learn mapping from the teacher’s vocabulary.

Reduce strategy leaves teacher representations shapes unchanged, and for the student we obtain sequence aligned to teacher sequence length:

$$(|X_s|, d_s) \mapsto (|X_t|, d_s), \quad (|X_s|, |V_s|) \mapsto (|X_t|, |V_s|) \quad (7)$$

This can be combined with the match strategy, if vocabulary alignment needed:

$$(|X_t|, |V_s|) \mapsto (|X_t|, |V_{match}|). \quad (8)$$

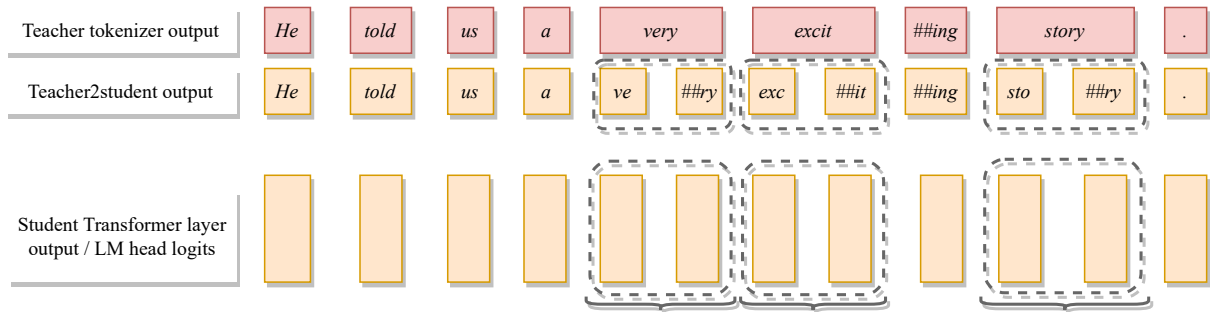


Figure 3: Reduce strategy. The first sequence is an output from the teacher’s tokenizer, the second is a greedy split result of the first sequence by subwords from the student’s vocabulary.

Reduce strategy combined with *match* over vocabulary introduces another way to use KL or CE losses for distillation. Compared to the match strategy only, we can use all teacher outputs and representations, so reducing the student sequence allows to extract knowledge for all tokens of the teacher vocabulary. But this approach still skips subwords from the student’s vocabulary which are not found in the teacher’s. This can be partially offset by passing two inputs to the student model. The first one is a teacher-to-student split with subsequent reduction to compute distillation losses, and the second output from the student’s tokenizer to compute supervised masked language modeling loss. The drawbacks of reduce compared to match strategy are higher overhead to compute losses and greedy split which might be not optimal.

4 Experiments

4.1 Pre-training

Corpus Teacher pre-training and distillation to the students was made on the same Russian Language data of $\sim 27M$ sentences collected from OpenSubtitles (Lison and Tiedemann, 2016), Dirty & Pikabu web resources, and Social Media segment of Taiga corpus (Shavrina and Shapovalova, 2017).

Models We used pre-trained rubert-base-cased-conversational (12-layer Russian BERT model)⁵ as a teacher with hidden states dimension of 768 and vocabulary size of 120K. It was the largest and the slowest model in our experiments.

Two students *distil-base*⁶ and *distil-small*⁷ have the same vocabulary and dimension of hidden states as the teacher, but a number of Transformer layers were reduced to 6 and 2. To train *distil-base* and *distil-small* we extended the distillation strategy proposed for DistilBERT (Sanh et al., 2019). Namely, we added MSE loss for averaged attention maps and cosine distance loss for averaged hidden states. To average teacher attention maps and hidden states, we grouped them by six Transformer layers for 2-layer *distil-tiny* and by two for 6-layer *distil-base* (because the teacher model has 12 Transformer layers).

Models *distil-tiny*(30|20|10|5) were students with 3 Transformer layers, hidden states dimension of 264 and reduced vocabulary sizes of 30k, 20k, 10k and 5k. We applied proposed alignment strategies to *distil-tiny** models.

We compare proposed distilled models to other available state-of-the-art distilled models for Russian rubert-tiny and rubert-tiny2⁸. Models rubert-tiny and rubert-tiny2 are 3-layer Trans-

⁵huggingface.co/DeepPavlov/rubert-base-cased-conversational

⁶huggingface.co/DeepPavlov/distilrubert-base-cased-conversational

⁷huggingface.co/DeepPavlov/distilrubert-tiny-cased-conversational

⁸huggingface.co/cointegrated/rubert-tiny, huggingface.co/cointegrated/rubert-tiny2, habr.com/ru/post/562064/

formers distilled from multiple teachers and combining MLM and Translation Ranking Modeling (TLM, (Feng et al., 2020)) losses.

All models that we trained and evaluated are listed in Appendix A Table 3 with corresponding inference speed and memory requirements.

Distillation with reduced vocabulary We distilled the teacher model into 3-layer student model `distil-tiny30` with 30k subwords in vocabulary. We tried different combinations of loss functions and alignment strategies. Combinations are summarized in Figure 4. In our experiments, we use MLM loss in summation with KL or MSE, or both of them. To compute KL loss teacher and student pre-softmax outputs should be aligned: 1. with the match strategy by sequence and vocabulary (`KL-match`); 2. with the reduce-match strategy, where reduction was made by sequence dimension and match-by vocabulary (`KL-reduce-match`). MSE loss for hidden states distillation was applied with match and reduce strategies. To match hidden sizes projection layers were used (see details in Appendix B.2).

To apply reduce strategy, we passed two inputs to the student: 1. a student-tokenized input for MLM loss; 2. teacher inputs tokenized by student for MSE and KL. Student representations corresponding to the student-tokenized input *were not aligned and were not used* to compute MSE or KL divergence.

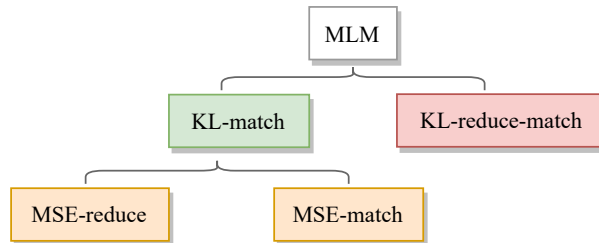


Figure 4: Combinations of loss functions (MLM, KL, MSE) and alignment strategies (reduce, match, reduce-match).

We initialized student models’ embeddings by re-using teacher embeddings (see details in Appendix B.1). Other training details could be found in Appendix B.

Ablation To check whether match and reduce strategies are effective for distilling knowledge from the teacher, we pre-trained `distil-tiny30` using only MLM loss and without any distillation losses. We also performed pre-training without KL divergence loss term to evaluate its contribution to `KL-match` & MLM & MSE combination.

To evaluate effect of reduced vocabulary on the distillation quality, we compared `distil-tiny*` models with reduced vocabulary to `distil-base` and `distil-small` with the same vocabulary as the teacher.

To determine how further vocabulary size reduction affects the distillation quality, we also distilled teacher into `distil-tiny` models with 5k, 10k and 20k vocabulary sizes (results are in Appendix C).

4.2 Fine-tuning

For evaluation we fine-tuned our models on ParaPhraser (Pivovarova et al., 2017), RuSentiment (Rogers et al., 2018), SberQuAD (Efimov et al., 2020), NER Collection-3 (Mozharova and Loukachevitch, 2016) and Russian SuperGLUE (Shavrina et al., 2020) datasets. Their description is given in Appendix D and Table 5.

Results for ParaPhraser, Collection-3, RuSentiment and SberQuAD are collected in the Table 1. From Table 1 we see expected result that pre-training with MLM is better than random initialization for further fine-tuning. Also, pre-training with distillation improves student models.

Results for the best distilled models on RussianSuperGLUE test sets are shown in the Table 2. We use the following naming conventions for `distil-tiny30` models for RussianSuperGLUE results:

- MLM & KL & MSE (RT) with reduce strategy and trainable hidden projections is `distil-tiny-1`;
- MLM & KL & MSE (MT) with the same losses combination and match strategy is `distil-tiny-2`;

- MLM & MSE (RF) with reduce strategy and frozen projections for hidden states is `distil-tiny-3`. We selected MLM & KL & MSE (MT) over MLM & KL & MSE (RF) despite the better average performance as this difference is caused by SberQuAD scores only. On the other datasets MLM & KL & MSE (MT) performs better or almost the same as MLM & KL & MSE (RF).

Logits distillation with KL divergence loss Proposed *match* and *reduce-match* strategies to align pre-softmax outputs of the teacher and student models improve results obtained by MLM pre-training only. Results from Table 1 show that *match* strategy performs better than *reduce-match*. Reduction of logits via summing might not result in the true probability of subword compounding of reduced subwords. The teacher model pre-training procedure does not guarantee that subword probability would be equal to multiplication of its compounding subwords probabilities.

Hidden states distillation Distilling hidden states with MSE loss can improve KL-match & MLM combination. On average, reduce strategy for hidden states alignment works better than match in combinations with KL divergence and without it. Distilling from hidden states allows extracting more knowledge from the teacher and its intermediate states. This observation holds for both the results in Table 1 and Russian SuperGLUE in Table 2.

Surprisingly, frozen projections, that is, non-trainable random projections, perform better for some of the configurations than trainable. For SberQuAD dataset, frozen projections steadily show higher F1 and EM scores, e.g., improving F1 for trainable projections from +1 to +26 F1 points. Though the result is not expected, it has also been previously observed that random projections could be very effective (Wieting and Kiela, 2019).

Model	Proj	Distillation Losses	ParaPhraser	RuSentiment	Collection-3	SberQuAD	
			F1	F1 (weighted)	Entity F1	F1	EM
teacher	-	MLM, NSP	86.30 \pm 0.96	76.00 \pm 0.53	97.01 \pm 0.13	83.82 \pm 0.15	65.60 \pm 0.12
distil-base	-	MLM, KL, MSE, Cos	82.86 \pm 0.47	75.82 \pm 0.98	96.40 \pm 0.20	80.05 \pm 0.43	60.96 \pm 0.51
distil-small	-		75.53 \pm 1.03	74.58 \pm 0.10	94.20 \pm 0.20	68.92 \pm 0.30	48.21 \pm 0.39
	-	-	72.48 \pm 0.32	69.27 \pm 0.35	75.61 \pm 0.41	17.54 \pm 0.09	4.46 \pm 0.14
	-	MLM	74.54 \pm 0.20	71.68 \pm 0.30	92.04 \pm 0.26	38.17 \pm 0.21	22.12 \pm 0.30
	M	MLM, KL	74.59 \pm 0.20	72.90 \pm 0.20	93.19 \pm 0.17	52.64 \pm 0.37	34.74 \pm 0.41
	RM		74.40 \pm 0.23	72.98 \pm 0.19	93.01 \pm 0.11	38.41 \pm 0.54	22.20 \pm 0.51
distil-tiny30	MF	MLM, KL, MSE	75.27 \pm 0.20	73.06 \pm 0.21	93.30 \pm 0.14	49.43 \pm 1.83	31.33 \pm 1.69
	MT		74.99 \pm 0.20	73.38 \pm 0.20	93.52 \pm 0.11	53.14 \pm 0.35	35.85 \pm 0.47
	RF		74.68 \pm 0.20	73.27 \pm 0.20	93.28 \pm 0.09	60.26 \pm 0.55	40.82 \pm 0.61
	RT		75.06 \pm 0.20	73.70 \pm 0.20	93.71 \pm 0.10	55.02 \pm 0.62	36.28 \pm 0.62
	MF	MLM, MSE	74.56 \pm 0.20	72.80 \pm 0.20	92.64 \pm 0.13	42.62 \pm 0.62	25.85 \pm 0.51
	MT		74.25 \pm 0.30	73.11 \pm 0.23	93.06 \pm 0.11	43.37 \pm 0.38	26.08 \pm 0.49
	RF		75.23 \pm 0.17	73.45 \pm 0.17	93.87 \pm 0.09	69.03 \pm 0.24	48.46 \pm 0.36
	RT		74.81 \pm 0.16	73.12 \pm 0.27	93.26 \pm 0.12	43.26 \pm 0.73	26.29 \pm 0.54
rubert-tiny	T	MLM, TLM, MSE	74.36 \pm 0.23	69.34 \pm 0.22	91.23 \pm 0.17	39.74 \pm 0.52	23.70 \pm 0.48
rubert-tiny2			78.72 \pm 0.15	71.84 \pm 0.24	93.72 \pm 0.11	67.80 \pm 0.22	47.64 \pm 0.32

Table 1: Fine-tuning results for ParaPhraser, RuSentiment, Colleciton-3 and SberQuAD. "Proj" column means type of alignment (first letter, match-M, reduce-R) and projection mode for hidden states (second letter, frozen-F, trainable-T). RM means reduce-match combination for KL loss. Empty "Losses" cell is to denote student without pre-training.

Distillation without KL divergence loss Surprisingly, the best of `distil-tiny30` students are MLM & MSE (RF) with reduce strategy and frozen projections did not use KL loss at all. MLM & MSE (RF) is very close by quality to `distil-small` and `rubert-tiny2`, requiring 10 \times (resp. 3 \times) less memory and being 2 \times (resp. 5 \times) faster on CPU. Moreover, for datasets from Table 1, except SberQuAD, losses combinations without KL divergence work very close to combinations with it. This result also holds on majority of Russian SuperGLUE tasks.

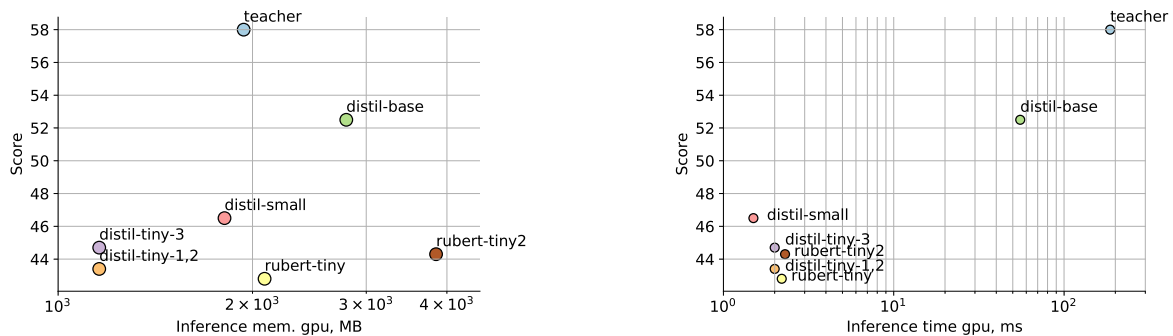
The low impact and inefficiency of KL-loss for distillation might be due to *match* a shift in matching subwords meanings in student and teacher vocabulary. But we do not have a solid proof for that and further investigation is needed.

Model	Score	LiDi	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DNQA	RuCoS
		Mcorr.	F1/Acc.	Acc.	F1a/EM	Acc.	Acc.	Acc.	Acc.	F1/EM
teacher	54.8	21.2	31.1/ 50.8	57.2	67.5/27.1	51.4	71.1	62.3	63	79/78.5
distil-base	49.84	8.5	33.0/47.1	61.0	51.1/6.3	49.5	63.5	63.0	65.5	69.0/68.6
distil-small	45.24	3.7	34.7/46.5	65.8	48.6/7.8	48.5	55.0	66.9	58.6	40.0/39.7
distil-tiny-1	42.63	4.2	28.8/48.9	49.0	40.4/6.6	53.7	55.1	63.6	60.4	35.5/35.2
distil-tiny-2	42.86	3.1	25.8/45.4	53.6	40.4/6.6	52.4	55.7	63.6	61.7	36.5/36.5
distil-tiny-3	44	4.6	35.0/50.1	52.7	43.3/7.4	52.8	56.5	66.9	61.7	33.0/32.7
rubert-tiny	42	-0.9	31.5/43.0	52.8	46.5/9.3	49.6	54.3	66.5	63.8	27.0/26.7
rubert-tiny2	45.19	17	36.7/43.7	57.1	44.5/ 9.8	50.4	59.5	65.9	58.7	31.0/30.5

Table 2: The model performance on the Russian SuperGLUE test sets. Matthews correlation for the LiDiRus task is scaled to $[-100, 100]$.

The results on Russian SuperGLUE partially meet the results on ParaPhraser, Collection-3, RuSentiment and SberQuAD. The teacher model significantly outperforms the rest. However, on PARus and RWSD *distil-small* achieves better results. This might be due to the limited size of the training data. All *distil-tiny** models achieve comparable results with the *distil-tiny-3*⁹ slightly ahead, so the contribution of the KL loss to the student performance is not clear. The *rubert-tiny2* model outperforms *rubert-tiny* confirming the previous results.

Dependence of model score on inference time and memory The dependence of models Russian SuperGLUE score on GPU¹⁰ inference time and memory is shown on Figure 5.



(a) The dependence of score on GPU inference memory

(b) The dependence of score on GPU inference time

Figure 5: The dependence of models Russian SuperGLUE average score on GPU inference time and memory. Random batches of size 16 and sequence length 512 were used.

From Figure 5a we can conclude that memory required for inference and model quality are not always correlated (e.g. *distil-base* and *rubert-tiny2* are worse by quality than the teacher but require more memory). It is caused by differences in the particular implementations of the Transformer architectures. Our *distil-tiny* models have the lowest memory consumption. At the same time from Figure 5b we can conclude that model score and inference time are highly correlated.

⁹We make it available at <https://huggingface.co/DeepPavlov/distilrubert-tiny-cased-conversational-v1>

¹⁰See hardware details in Appendix A

5 Conclusions and future work

We introduced two language model distillation strategies allowing to reduce student’s vocabulary. *Match* strategy uses only representations for the subwords which are common for a teacher and a student vocabularies. *Reduce* strategy aggregates a student’s subwords representations corresponding to particular teacher’s subwords. We performed experiments to show how vocabulary reduction affects the distillation process and how our strategies can be effectively applied for distillation based on teacher output and intermediate representations. We trained student models of different sizes which are from $1.3\times$ to $49\times$ smaller than the teacher while maintaining a good quality compared to the other SOTA models for Russian of similar size. We found that distillation without Kullback-Leibler divergence loss for models with reduced vocabularies performs the best. Our experiments showed that $17\times$ compressed student with reduced vocabulary can work very close to $1.3\times$ compressed student with the same vocabulary as the teacher. Additionally, we made the best of our models and code to train them publicly available.

As further improvements, we consider other ways of distilling intermediate representations based on contrastive and metric learning approaches as well as the more accurate mapping between mismatched subwords in vocabularies to transfer as much knowledge as possible during the distillation process.

Acknowledgments

This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138

Authors’ contributions

A.K. suggested some of the experiments, developed and optimized code for training and fine-tuning (except Russian SuperGLUE), carried out the experiments, performed most of the computations, and wrote the manuscript (except Sec. 1, 2). Y.K. suggested the original idea of experiments, helped with code optimization, helped with performing and designing experiments, participated actively in results discussion, wrote Sec. 1, 2 and edited the rest of the manuscript. V.K. performed fine-tuning on the Russian SuperGLUE benchmark, described the results, and suggested edits for the manuscript. M.B. supervised the team, discussed intermediate results and directions of the study, contributed to the manuscript’s final version. All authors discussed results and approved the final version of the paper.

References

- Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. 2020. Knowledge distillation from internal representations. // *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, P 7350–7357.
- Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? // Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a " siamese" time delay neural network. *Advances in neural information processing systems*, 6.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. // H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, *Advances in Neural Information Processing Systems*, volume 33, P 1877–1901. Curran Associates, Inc.
- Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. // *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, P 535–541.

- Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nikolay 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 Vikhrev, 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.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. // Hal Daumé III and Aarti Singh, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, P 1597–1607. PMLR, 13–18 Jul.
- Yew Ken Chia, Sam Witteveen, and Martin Andrews. 2018. Transformer to CNN: Label-scarce distillation for efficient text classification. // *NIPS Workshop CDNNRIA 2018*.
- 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.
- Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. Sberquad–russian reading comprehension dataset: Description and analysis. // *International Conference of the Cross-Language Evaluation Forum for European Languages*, P 3–15. Springer.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.
- Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. 2014. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*.
- Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. // *Proceedings of the 5th Workshop on Representation Learning for NLP*, P 143–155, Online, July. Association for Computational Linguistics.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. // *Proceedings of the IEEE international conference on computer vision*, P 1026–1034.
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. // *NIPS Deep Learning and Representation Learning Workshop*.
- Oleksii Hrinchuk, Valentin Khruikov, Leyla Mirvakhabova, Elena Orlova, and Ivan Oseledets. 2019. Tensorized embedding layers for efficient model compression. *arXiv preprint arXiv:1901.10787*.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. // *Findings of the Association for Computational Linguistics: EMNLP 2020*, P 4163–4174, Online, November. Association for Computational Linguistics.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. // *International Conference on Learning Representations*.
- Yann LeCun, John Denker, and Sara Solla. 1989. Optimal brain damage. *Advances in neural information processing systems*, 2.
- Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. 2021. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*.
- Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.
- 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*.

- Valerie Mozharova and Natalia Loukachevitch. 2016. Two-stage approach in russian named entity recognition. // *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, P 1–6. IEEE.
- Lidia Pivovarova, Ekaterina Pronoza, Elena Yagunova, and Anton Pronoza. 2017. Paraphraser: Russian paraphrase corpus and shared task. // *Conference on artificial intelligence and natural language*, P 211–225. Springer.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Anna Rogers, Alexey Romanov, Anna Rumshisky, Svitlana Volkova, Mikhail Gronas, and Alex Gribov. 2018. Rusentiment: An enriched sentiment analysis dataset for social media in russian. // *Proceedings of the 27th international conference on computational linguistics*, P 755–763.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: «taiga» syntax tree corpus and parser. *Proceedings of the “Corpora*, P 78–84.
- Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 4717–4726, Online, November. Association for Computational Linguistics.
- Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. // *EMNLP/IJCNLP (1)*, P 4322–4331.
- Siqi Sun, Zhe Gan, Yuwei Fang, Yu Cheng, Shuohang Wang, and Jingjing Liu. 2020a. Contrastive distillation on intermediate representations for language model compression. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 498–508, Online, November. Association for Computational Linguistics.
- Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020b. MobileBERT: a compact task-agnostic BERT for resource-limited devices. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 2158–2170, Online, July. Association for Computational Linguistics.
- Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*.
- Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. // *International Conference on Learning Representations*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. // I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- 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.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. // H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, *Advances in Neural Information Processing Systems*, volume 33, P 5776–5788. Curran Associates, Inc.
- Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 2140–2151, Online, August. Association for Computational Linguistics.
- John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. // *International Conference on Learning Representations*.
- Sanqiang Zhao, Raghav Gupta, Yang Song, and Denny Zhou. 2021. Extremely small BERT models from mixed-vocabulary training. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 2753–2759, Online, April. Association for Computational Linguistics.

A Models

We measured inference time and memory required for models from the Table 3 on NVIDIA GeForce GTX 1080 Ti and Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz using benchmark utils from Transformers library¹¹. For testing random sequences with `batch_size = 16` and `sequence_length = 512` were generated. We run each model 100 times to reduce an effect of possible external factors on time and memory values. Some of the distilled models require more memory for inference due to different implementations of DistilBERT and BERT architectures.

Model	# layers	# vocab, K	# hid	Params, M	Mem, MB	Inference time, ms		Inference mem, MB	
						cpu	gpu	cpu	gpu
teacher	12			177.9	679	5283.2	186.6	1550	1938
distil-base	6	119.5	768	135.5	517	2335.4	55.3	2177	2794
distil-small	2			107.1	409	802.4	1.5	1541	1810
distil-tiny30		30.5		10.4	41	374.7	2	714	1158
distil-tiny20	3	20	264	7.6	30	357.6	1.9	695	1148
distil-tiny10		10		5	19	356.5	1.8	679	1138
distil-tiny5		5		3.6	14	354.9	1.8	664	1126
rubert-tiny	3	29.6	312	11.8	45.5	942.9	2.2	1308	2088
rubert-tiny2		83.8		29.3	112	1786.6	2.3	3054	3848

Table 3: Teacher and student models characteristics. All models have 12 attention heads. "Mem" column is memory on disk required to store model, while "Inference time"/"Inference mem" is time/memory required for model to make inference on a given batch. Inference tests were made on batches of 16 random sequences with length 512. For distil-tiny* models, * corresponds to a vocabulary size in thousands.

Comparing to the teacher `distil-base` is $1.3\times$ lighter and $3.5\times$ faster on GPU. At the same time `distil-small` is $1.7\times$ lighter and $126\times$ faster on GPU. But the memory required for inference remains almost the same as for teacher.

As vocabulary size decreases, the students `distil-tiny` are getting lighter: from $17\times$ to $49\times$ for models from 30k to 5k vocabulary. Inference time and memory holds almost the same order. Models `distil-tiny` are up to $104\times$ faster on GPU; memory consumption is up to 1.7 times lower on GPU. But still `distil-small` is the fastest of all students because of the lowest number of Transformer layers.

Nevertheless, `rubert-tiny` is $15\times$ lighter (`rubert-tiny2` $6\times$) than our teacher. Both models are $85\times$ faster on GPU, but require even more memory for inference.

B Training details

Our code is based on DistilBERT open-source implementation¹². We trained students on 8 Tesla P100-SXM2-16Gb for 64 epochs with `batch_size = 4`, `gradient_accumulation_steps = 128` and AdamW optimizer (Loshchilov and Hutter, 2017). For learning rate we applied warmup from 0 to $5e^{-4}$ and when required number of warmup steps passed, learning rate was halved after three validation epochs, if validation loss was not improved. We used DeepPavlov library (Burtsev et al., 2018) for our fine-tuning experiments.

B.1 Weights initialization

We initialized student models with parameters from the teacher. To initialize student embeddings we made the following steps:

1. Subwords from teacher vocabulary were split by student subwords (see *reduce* in Sec. 3.3).
2. For each student subword we collected corresponding teacher subwords in which that subword occurred (according to the splits from previous step).

¹¹huggingface.co/docs/transformers/benchmarks

¹²github.com/huggingface/transformers/tree/master/examples/research_projects/distillation

- Student subword embeddings were initialized with averaged embeddings of the corresponding teacher subwords.

To initialize student layers, 12 Transformer layers of the teacher were grouped by 4 and averaged to match 3 student layers. Then we cut them to match student hidden states dimension.

B.2 Distilling teacher hidden states

The following steps were made:

- Student and teacher model have different number of Transformer layers. Therefore, for each input token we averaged outputs of all Transformer layers for this token.
- Match or reduce strategies were applied to align student sequence length dimension.
- Averaged and aligned student hidden states were projected by fully-connected layer to match the teacher hidden states dimension. We initialized projection layers randomly (He et al., 2015) and use them in two modes – *frozen* and *trainable*.
- MSE loss computed between aligned student and teacher hidden states.

C Experiments with different vocabulary sizes

As vocabulary size decreased, we expected more teacher knowledge would be lost, and students quality would decrease proportionally. Surprisingly we do not see this effect. For the same combination of losses KL-match & MLM we observe two groups of results in Table 4: 1. Scores on ParaPhraser and SberSQuAD increase as vocabulary size decreases. 2. Scores on RuSentiment and Collection-3 decrease as vocabulary become smaller.

Model	Proj	Distillation Losses	ParaPhraser	RuSentiment	Collection-3	SberQuAD	
			F1	F1 (weighted)	Entity F1	F1	EM
teacher	-	MLM, NSP	86.30 \pm 0.96	76.00 \pm 0.53	97.01 \pm 0.13	83.82 \pm 0.15	65.60 \pm 0.12
distil-base	-	MLM, KL, MSE, Cos	82.86 \pm 0.47	75.82 \pm 0.98	96.40 \pm 0.20	80.05 \pm 0.43	60.96 \pm 0.51
distil-small	-		75.53 \pm 1.03	74.58 \pm 0.10	94.20 \pm 0.20	68.92 \pm 0.30	48.21 \pm 0.39
distil-tiny30	M	MLM, KL	74.59 \pm 0.20	72.90 \pm 0.20	93.19 \pm 0.17	52.64 \pm 0.37	34.74 \pm 0.41
distil-tiny20			74.35 \pm 0.59	72.49 \pm 0.21	92.57 \pm 0.15	48.46 \pm 1.39	31.11 \pm 1.34
distil-tiny10			74.58 \pm 0.24	72.50 \pm 0.24	92.20 \pm 0.14	64.05 \pm 0.82	44.66 \pm 0.83
distil-tiny5			74.88 \pm 0.33	70.86 \pm 0.29	91.43 \pm 0.15	67.46 \pm 0.26	47.82 \pm 0.26

Table 4: Results for students with different vocabulary sizes. Teacher, distil-base, distil-small have 120k tokens in vocabulary.

D Fine-tuning datasets

ParaPhraser is a set of sentence pairs collected from news headlines and annotated as precise paraphrase, near paraphrase and non-paraphrase. The task we solve is binary classification – predict whether sentence pairs are paraphrases (precise or near paraphrases) or not. RuSentiment is a dataset for sentiment analysis of public posts on Russian social network VKontakte. Five categories were annotated "Neutral", "Negative", "Positive", "Speech Act", and "Skip". SberQuAD is a Russian QA dataset for a reading comprehension evaluation which contains paragraph-question-answer triples. Questions were constructed in such a way that answer is a some paragraph span. For NER task we used Collection-3: Persons-1000 collection¹³ which contains names of persons, additionally annotated with organizations and locations named entities.

RussianGLUE is an advanced Russian general language understanding evaluation benchmark that contains nine tasks, collected and organized similarly to the SuperGLUE (Wang et al., 2019) methodology. The benchmark can be divided into six groups including the general diagnostics of language models, common sense understanding, natural language inference, reasoning, machine reading and world knowledge.

¹³ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

Dataset	Type	Metric	Train	Validation	Test
ParaPhraser	Classification	F1	6702	500	1899
RuSentiment		F1 (weighted)	31030	3448	4961
SberQuAD	Span prediction	F1, EM	45328	5036	-
Collection-3	NER	Entity F1	9301	2153	1922
Russian SuperGLUE					
RUSSE	Common Sense	Acc	19845	8508	18892
PARus			500	100	400
TERRa	NLI	Acc	2616	307	3198
RCB		F1, Acc.	438	220	438
LiDiRus		MCC	0	0	1104
RWSD	Reasoning	Acc	606	204	154
MuSeRC	Machine Reading	F1, EM	500	100	322
RuCoS			72193	7577	7257
DaNetQA	World Knowledge	Acc	1749	821	805

Table 5: Summary of the common benchmark datasets for Russian with train/validation/test split sizes.

Эмоциональный анализ постов ВКонтакте: классификатор или регрессор?

Колмогорова А. В.

Сибирский федеральный
университет /Красноярск, Россия
nastiakol@mail.ru

Калинин А. А.

Сибирский федеральный
университет /Красноярск, Россия
verbalab@yandex.ru

Аннотация

В статье обсуждаются результаты решения двух задач машинного обучения: задачи классификации текстов социальных сетей на русском языке по критерию доминирующей эмоции и задачи регрессии, в рамках которой эмоции в тех же текстах социальных сетей предсказываются. В основе экспериментов – сформированный авторами датасет из 3879 текстов из пабликов ВКонтакте, размеченный 2000 ассессорами на краудсорсинговой платформе Толока. Аннотирование проводилось с использованием разработанного интерфейса для недискретной эмоциональной разметки текстов.

Ключевые слова: эмоциональный анализ текстов; классификатор; регрессор; тексты социальных сетей
DOI: 10.28995/2075-7182-2022-21-311-322

Emotional Analysis of posts in VKontakte: classification or regression?

Kolmogorova A. V.

Siberian Federal University /
Krasnoyarsk, Russia
nastiakol@mail.ru

Kalinin A. A.

Siberian Federal University /
Krasnoyarsk, Russia
verbalab@yandex.ru

Abstract

The article summarizes the results of two tasks in machine learning paradigm: the task of classification according to the criterion of dominating emotion on the data of social networks posts in Russian and the regression task using the same data. The experiments are conducted on the data set collected from VKontakte social network and consisted of 3879 posts assessed by 2000 informants on Toloka crowd sourcing platform. The annotation procedure was based on the original interface for non-discrete emotion assessment elaborated by researchers.

Keywords: emotional text analysis; classifier; regressor; texts of social networks

1 Введение

Эмоциональный анализ текстов – интенсивно развивающееся направление, продолжающее традиции уже сложившейся методологии сентимент-анализа [1; 6]. Если для последнего характерно внимание к двум или трем тональностям (позитивная/негативная/нейтральная), то эмоциональный анализ сфокусирован на детектировании конкретных эмоций: страха, гнева, радости и т.д.

Количество выделяемых классов и характер используемых языков в подобного рода проектах варьирует: например, 4 – для испанского [8], 6 – для русского и английского языков [16; 2]. В нашей работе мы использовали в качестве основы восьмичастную модель эмоций шведского нейрофизиолога Г. Лёвхема [11] – так называемый Куб Лёвхейма.

Для большинства проектов в данной области своеобразным камнем преткновения является организация процедуры разметки (ассессмента) текстов на предмет содержащейся в них эмоции. Отобрав посты из трех эмоционально насыщенных пабликов ВКонтакте в качестве материала, мы разработали авторский интерфейс для недискретной оценки эмоций в тексте, т. е. разметки,

предусматривающей не приписывание тексту эмоционального лейбла, а помещение текста в определенную точку эмоционального континуума. Размеченная таким образом коллекция из почти 4000 текстов стала основой для дальнейшего обучения моделей.

В фокусе данной публикации – 2 эксперимента. Один связан с решением задачи классификации текстов по критерию ведущей эмоции, другой – с решением задачи построения регрессионной модели, предсказывающей многомерные эмоциональные характеристики текста.

2 Задачи классификации и регрессии

Регрессия – контролируемый алгоритм машинного обучения, используемый для прогнозирования непрерывных значений выходных данных на основе входных данных. Задача регрессии требует предсказания количества. Иначе говоря, прогнозирующее регрессионное моделирование – это задача приближения функции отображения (f) от входных переменных (X) к непрерывной выходной переменной (y). Регрессия может иметь действительные или дискретные входные переменные. Применительно к решению задачи эмоционального анализа текста, регрессионные модели предсказывают «количество» той или иной эмоции в тексте, подаваемом на вход для анализа – например, в нем представлены в таких-то количествах Грусть, Радость, Удивление и Отвращение.

Классификация представляет собой контролируемый алгоритм машинного обучения, который предсказывает определенные дискретные значения (категории или классы), к которым принадлежат входные данные. При решении задачи классификации от модели МО требуется заключение о принадлежности документа к определенному классу на основе ранее изученных примеров каждого из возможных классов. Иными словами, классификационное прогнозирующее моделирование – это задача приближения функции отображения (f) от входных переменных (X) к дискретным выходным переменным (y). В контексте нашего исследования это означает, что модель МО предсказывает эмоциональный класс, к которому принадлежит поданный на вход для анализа текст – он, например, грустный.

Хотя модели, основанные на классификации, широко распространены в практике эмоционального анализа, за последние пять лет количество работ, использующих регрессионные модели для обработки текстовых данных, содержащих эмоции, также постоянно растет. Это связано с тем, что понимание эмоции как дискретной системы накладывает определенные ограничения на точность и нюансированность анализа: естественное переживание некоторого эмоционального состояния редко бывает окрашено только одной эмоцией – даже если ведущим аффектом является, например, Радость, к ней могут примешиваться, хотя и с меньшей выраженностью, Воодушевление или Удивление, а также нотки Грусти. Переход от дискретной парадигмы, когда эмоция, вербализованная в тексте, рассматривается как класс, к использованию непрерывных ее измерений открывает более широкие перспективы для получения результатов максимально приближенных к естественным процессам интерпретации эмоции в тексте читающим его человеком.

Отметим, что в пока в исследовательской практике наблюдаются противоречивые результаты и полярные мнения относительно преимуществ регрессионных моделей над классификационными в контексте анализа эмоций в текстовых данных.

Например, в [12], взяв в качестве исходных данных короткие тексты о погоде в Твиттере, исследователи сосредоточились на проверке гипотезы о том, что регрессионный анализ в сочетании с оценкой достоверности классов работает лучше, чем традиционные методы классификации, такие как SVM и K-Nearest Neighbor, основанные на дискретных целевых переменных. Разработчики пришли к выводу, что в целом регрессионный подход не показал явных преимуществ: по видимому, он более оправдывает себя в задачах, требующих более детализированного анализа, предполагающего вычисление множества разнородных признаков.

В исследовании [4] регрессионный подход был применен для анализа текстовых данных из различных открытых датасетов, размеченных на основе многомерной модели эмоций, известной как VAD-модель [13]. Модель построена на количественном измерении трех параметров: уровня возбуждения, уровня доминирования и значений полярности эмоции (негативная / позитивная). Комбинация значений данных параметров, в конечном счете, определяет степень интенсивности шести базовых эмоций. В итоге исследователями было достигнуто приемлемое качество предсказания, а сама методология оценена ими как перспективная.

Очевидно, что не последнюю роль в достижении показателей эффективности модели играют данные, на которых она будет работать: обучаться и тестироваться. Мы предположили, что наш датасет, который был получен путем применения процедуры недискретной разметки текстов из ВКонтакте на основе многомерной модели эмоций Куб Лёвхейма, хорошо совместим с регрессионными моделями.

3 Специфика интерфейса для недискретной оценки эмоции в тексте

В эмоциональном анализе текстов принято два основных подхода к разметке: дискретный (категориальный) и многомерный [5]. Как мы указывали выше, первый хорошо сочетается с задачами по классификации, основу которой, как правило, составляет модель шести базовых эмоций П. Экмана [7]. Второй используется реже, поскольку моделей, представляющих эмоции как континуум, в рамках некоторого пространства, заданного тремя параметрами, немного. Куб Лёвхейма – одна из таких моделей.

Шведский нейрофизиолог Гуго Лёвхейм в [11] предложил оригинальную концепцию эмоций, которую визуализировал, подобно Дж. Расселу и А. Мерабиану, авторам VAD-модели, в виде куба. Суть идеи Лёвхейма сводится к тому, что характер эмоции, субъективно испытываемой человеком, определяется уровнем трех нейротрансмиттеров в амигдале головного мозга. Это норадреналин (вертикальная ось Куба Лёвхейма), серотонин (горизонтальная ось) и дофамин (глубина Куба Лёвхейма) (рис.1). В восьми вершинах Куба расположены восемь базовых эмоций. Каждая из них имеет двучленную номинацию, где первое слово обозначает самую слабую степень интенсивности эмоции, а второе – самую сильную. В дальнейшем нам пришлось упростить номинации, оставив только по одному слову. Для того, чтобы найти адекватные эквивалентные номинации на русском языке, мы предварительно провели психолингвистический эксперимент, в котором 30 информантам, носителям русского языка, было предложено соотнести эмоциональные тексты из ВКонтакте (например, под хештегами #Подслушано_успех или #Подслушано_фууу и др.) с названиями эмоций из списка, в котором фигурировали разные возможные варианты перевода английских слов, в оригинале использованных Г. Лёвхеймом. Так например, для лексемы *excitement* нами были предложены варианты: Возбуждение, Воодушевление, Энтузиазм, Оживление. Наибольший рейтинг по итогам эксперимента с текстами получила номинация Воодушевление. Для лексемы *disgust* нами были предложены варианты: Отвращение, Омерзение, Брезгливость, Раздражение. Первое место в «рейтинге» информантов получила номинация Отвращение.

Итак, по вертикальной оси, указывающей на уровень норадреналина, такие эмоции как Стыд, Страх, Радость и Отвращение располагаются внизу (что указывает на низкий уровень норадреналина), а Грусть, Гнев, Воодушевление и Удивление – наверху (указывая на высокий уровень этого нейротрансмиттера).

Следующая ось – горизонтальная, она соответствует уровню серотонина. Стыд, Страх, Грусть, Гнев находятся на «левой стороне» (низкий уровень). Отвращение, Радость, Удивление и Воодушевление находятся на «правой» стороне, что соответствует высокому уровню данного нейротрансмиттера.

И, наконец, последняя ось соответствует уровню дофамина. На ней располагаются Грусть, Стыд, Отвращение и Удивление как низко-дофаминовые эмоции, и Страх, Гнев, Воодушевление и Радость как высоко-дофаминовые эмоции.

Мы использовали данную модель в качестве концепт-идеи, поскольку данное пространство внутри Куба является своеобразным эмоциональным континуумом, в котором мы можем локализовать каждый текст, оцененный информантами при аннотировании. Проблема состояла в том, какой интерфейс можно предложить аннотаторам, чтобы они могли не просто приписать тексту определенную метку, но и показать количество той или иной эмоции, присутствующей в нем.

Ввиду трехмерности и недискретности эмоциональной модели Лёвхейма, процедура аннотирования должна учитывать следующую специфику целевых переменных:

- аннотирование должно представлять собой недискретную, в идеале – непрерывную оценку на определенном интервале, а не простую атрибуцию к определенному классу;
- аннотирование должно содержать оценки для всех трех измерений Куба Лёвхейма.

Поскольку эмоция в рамках подхода Лёвхейма представляет собой точку в трехмерном пространстве, то в идеале пользователь должен выбрать эту самую точку в пределах Куба (рис. 1). Но здесь мы сталкиваемся с проблемой удобства интерфейса. Для такого рода задачи требуется интерфейс, позволяющий использовать 3D-пространство со множеством сопутствующих операций: вращение, приближение, движение в двух плоскостях. Такой инструментарий крайне дорог в разработке, но, что самое важное – он был бы крайне неудобен в использовании, т. к. прежде чем поставить эту самую точку, соответствующую эмоциональному положению текста, ассессору необходимо было бы совершить немало предварительных шагов, предварительно ознакомившись с информацией о нейротрансмиттерах. Это непростая задача.

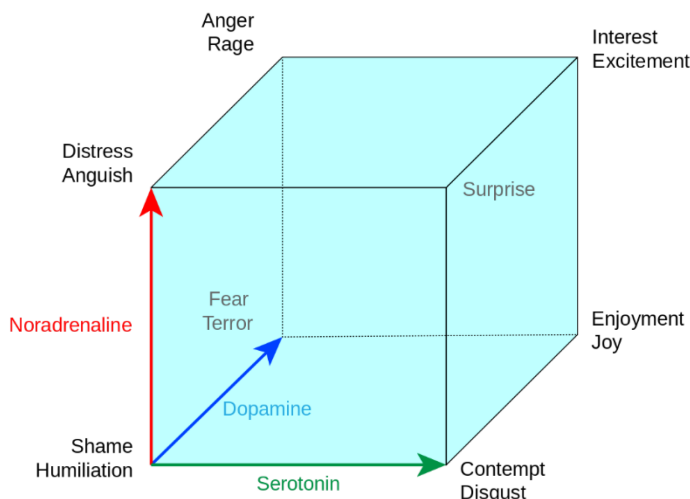


Рис. 1: Визуализация модели Куба Лёвхейма [Lövheim, 2012]

Приняв вышеперечисленные ограничения, мы пришли к выводу, что процедуру разметки необходимо упрощать. Немного потеряв в точности, мы все же сможем получить большое количество данных за счет привлечения ассессоров, которым можно будет поставить задачу, выполнимую с минимальным уровнем подготовки и с использованием знакомых и понятных UI-примитивов в интерфейсе разметки. Для реализации такого подхода мы обратили внимание на одно интересное свойство кубов, а именно – главные диагонали, соединяющие противоположные вершины Куба.

Эти вершины Куба (соответствующие восьми крайним эмоциям) образуют четыре оппозиции, которые накладываются на диагонали внутри него (см. рис. 2). Оппозиции основаны на принципе «качелей нейротрансмиттеров» (табл.1): у эмоций, образующих вершины диагоналей, противоположные значения уровня нейротрансмиттеров. Если, например, у Отвращения максимальный уровень серотонина, и минимальные – норадреналина (гормона стресса, вызывающего активацию всех физиологических проявлений возбуждения [14]) и дофамина (гормона, вызывающего прилив мотивации, обострение восприятия времени и пространства [Ibid]), то у Гнева – все с точностью наоборот, поэтому в определенном смысле он является агрессивным антонимом Отвращения.

оппозиция эмоций	серотонин	норадреналин	дофамин
Грусть	–	+	–
Радость	+	–	+
Гнев	–	+	+
Отвращение	+	–	–
Стыд	–	–	–
Воодушевление	+	+	+
Страх	–	–	+
Удивление	+	+	–

Таблица 1: Сочетания максимального / минимального уровня нейротрансмиттеров у эмоций, образующих вершины диагоналей Куба Лёвхейма

Подобная дихотомия, опирающаяся на диагонали, позволяет нам условно представить «спаренные диагональю» эмоции как антиподы, а поскольку понятие «противоположность», или «антоним», более близко специально неподготовленному информанту (ассессору), чем, например, понятия «моноамин» или «нейротрансмиттер», то мы можем заменить задачу указания конкретной моноаминовой конфигурации на задачу поиска баланса между четырьмя противоположностями, существенно упростив таким образом задачу информантам.

Указание значения баланса на уровне интерфейса, предназначенного для ассессора, может быть реализовано через ползунок (slider-bar). Перемещая ползунок по шкале, ассессор может гибко указывать на «степень выраженности» двух противоположных эмоций.

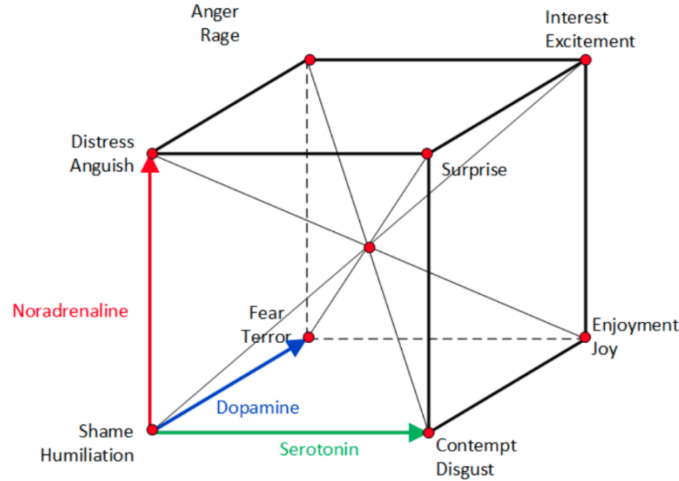


Рис. 2: Диагонали внутри Куба Лёвхейма

Так, например, с использованием такого инструмента можно разметить текст как «умеренно гневный», «немного стыдный», «не содержащий ни грусти, ни радости» и «слишком страшный», используя все четыре шкалы.

Дизайн интерфейса разметки представлен на рисунке 3. Крайние точки каждой шкалы обозначают максимально выраженную эмоцию. Точка посередине – это нейтральное значение. Чем меньше, по мнению ассессора, выражена эмоция, тем ближе к середине он перемещает слайдер. Перед оценкой текста, на экране ассессору предлагалась следующая инструкция: *Внимательно прочитайте текст. Какие эмоции автор выражает в тексте? На каждой из шкал поставьте отметку ближе к той эмоции, которая сильнее выражена в тексте. Поставьте отметку настолько близко, насколько очевидна и сильна эта эмоция в тексте. Например, 1 шаг от центра – если оттенок эмоции присутствует, но выражен слабо; 3 шага – если эмоция явно присутствует; 5 шагов – если эмоция, без сомнения, доминирует. Если в тексте нет эмоций, обозначенных на шкале, оставьте отметку в среднем положении.*

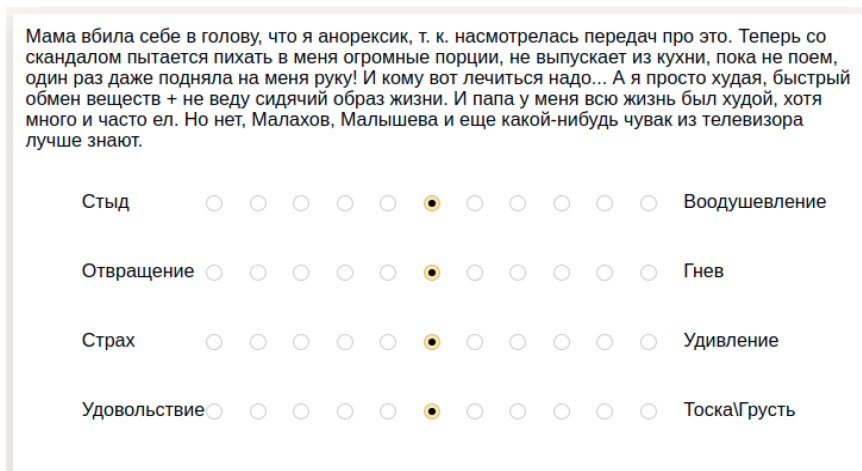


Рис. 3: Пример экрана с недискретной разметкой с опорой на диагонали Куба Лёвхейма

Основными преимуществами вышеупомянутого инструмента оценки являются:

- 1) ассессор может указать не только на первичную эмоцию, но и на вторичную, если, например, образец текста в основном касается чувства страха, но аннотатор хочет упомянуть легкое воздействие гнева;
- 2) в течение разметки ассессоры могут изменять степень выражения эмоции, перемещая ползунок относительно средней точки;
- 3) по результатам разметки мы имеем возможность рассматривать каждую оппозицию как период $(-1,1)$ и сопоставлять положение ползунка с вектором, угол которого совпадает с соответствующей диагональю, и величину этих векторов, которая равна значению, выбранному для данной эмоциональной оппозиции.

После снятия данных мы можем аппроксимировать результирующую точку в виде агрегации векторов. Разумеется, мы теряем возможность более точного указания точки, поскольку указание координат идет только на отрезках диагоналей, и финальное агрегированное значение получается приблизительным, аппроксимированным, но одновременно мы делаем интерфейс удобным, а целевые метрики – максимально приближенными, и можем соблюсти баланс точности данных и размера выборки.

С использованием данного интерфейса было размечено 3789 текстов из пабликов «Подслушано», «Палата №6» и «Карамель» ВКонтакте. В разметке приняли участие 2000 ассессоров, зарегистрированных на платформе Толока. Минимальным условием было двойное покрытие выборки, т.е. минимально каждый текст из выборки оценивался двумя информантами, но для большинства текстов этот критерий перевыполнялся.

В качестве метрики для оценки согласованности ответов аннотаторов мы использовали коэффициент внутриклассовой корреляции, предложенный в [3]. Согласно метрике, коэффициент 0 означает, что между значениями в классе нет никакой корреляции, 1 – что они идентичны. В нашей выборке коэффициент межклассовой корреляции составил 0,6325, что показывает, что степень согласованности аннотаторов в целом удовлетворительная. Явные аутлайнеры, составившие в общей сложности не более 2 % от выборки, исключались, поскольку как было показано в одном из наших экспериментов [10], маргинальные оценки чаще всего давались ассессорами с высоким уровнем индивидуальной эмпатии.

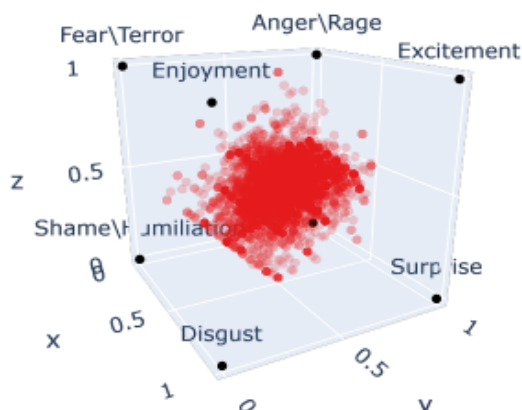


Рис. 4: Визуализация выборки оцененных текстов в трехмерном пространстве Куба

Полученная нами выборка визуализирована на рис. 4. Каждый единожды оцененный текст – это точка в пространстве Куба.

4 Решение задачи классификации

При разработке системы автоматической классификации эмоций нами были проведены эксперименты, комбинирующие ранее описанные модели векторного представления текста и алгоритмы машинного обучения. Первая задача, которая была поставлена, сводилась к апробации моделей дискретной классификации, поскольку данные модели имеют широкую практику применения.

Для решения этой задачи четыре вектора, которые получал при аннотировании каждый текст нашей выборки, агрегировались в один усредняющий вектор, а ближайшая к нему вершина принималась за метку эмоционального класса текстов: если ближайшей вершиной оказывалась Радость, то текст относился к соответствующей категории «радостных» текстов.

В каждом из экспериментов использовались идентичные обучающая и тестовая выборки. Количество фрагментов на каждый из восьми классов эмоций в обучающей выборке было следующим: Стыд – 757, Грусть – 551, Воодушевление – 397, Радость – 390, Гнев – 336, Отвращение – 307, Удивление – 248, Страх – 159.

Текстовые данные в тестовой выборке разбиты подобным образом (соотношение обучающей и тестовой – 80% / 20%): текстов, размеченных как Стыд – 189, Грусть – 138, Воодушевление – 100, Радость – 98, Гнев – 83, Отвращение – 76, Удивление – 62, Страх – 40.

Векторизация текстов была выполнена с помощью трех наиболее популярных моделей векторного представления:

- Bag-of-words + TF-IDF (Sparse vectors);
- fastText (Common Crawl RU 300);
- ELMo (Deep Pavlov, Russian Wikipedia 1024).

В качестве методов машинного обучения, проводящих непосредственную классификацию, нами использовались:

1. GaussianNB
2. KNeighbors Classifier
3. LogisticRegression
4. Linear Discriminant Analysis
5. XGBClassifier (xgboost)
6. MultiLayerPerceptron — 5 Layers + SGD + Cross-Entropy Loss
7. bidirectional LSTM
8. Conv1D

Для работы с векторами fastText и ELMo в случаях, где алгоритм ожидает одного вектора, а не цепочки (1-6), вектора токенов, входящих в тексты, были усреднены, чтобы классификатор мог работать с одним вектором, представляющим весь текст. Для «нейронных» классификаторов (6-8) применялся SGD оптимизатор и перекрестная энтропия в качестве лосс-функции.

Для оценки качества работы классификатора мы использовали комбинацию двух наиболее важных характеристик классификатора – точности и полноты.

Такая оценка позволяет учитывать количество истинно-положительных, истинно-отрицательных, ложно-положительных и ложно-отрицательных атрибуций по отношению к полноте выполнения задачи и дает сбалансированную характеристику модели. В дальнейшем под точностью/эффективностью работы классификатора мы будем понимать взвешенное среднее арифметическое значений метрики F1, рассчитанное для каждой из эмоций. Результаты значения метрики F1 при комбинации рассмотренных ранее алгоритмов и моделей векторизации представлены в Таблице 2. Алгоритм линейного дискриминантного анализа с признаками репрезентаций TF-IDF обучить не удалось ввиду недостаточно высокой вычислительной мощности. При вычислении TF-IDF весов модель обрабатывает обширные массивы данных, устанавливая большое количество линейных зависимостей на мелких списках признаков.

Модели векторизации/ алгоритмы МО	Наивный Байес	к-ближ. соседей	Логист. регрессия	ЛДА	xgboost	многоп. перцеп.	LSTM	Conv1D
TF-IDF	43,26	45,40	58,64	–	47,97	54,92	55,92	55,43
fastText	40,43	41,53	44,06	60,80	52,97	52,66	53,65	53,40
ELMo	40,34	44,87	61,62	54,89	53,97	53,03	55,20	55,86

Таблица 2: Результаты работы классификатора по мере F1, %

Среди описанных выше классификаторов наилучшую точность показывает алгоритм логистической регрессии (табл.2). Лучшее качество предсказания этот алгоритм получает при использовании признаков векторного представления ELMo, достигая среднего значения метрики F1 в 61,62 %. Худшее качество классификатор, основывающийся на логистической регрессии, показывает при использовании векторных представлений fastText. Интересно, что fastText на всех остальных алгоритмах дает достаточно высокую точность.

Помимо логистической регрессии, близкое к максимальному значению точности значение показывает алгоритм линейного дискриминантного анализа, работающий в совокупности с векторным представлением fastText – 60,8 %. Для всех остальных алгоритмов и векторных представлений текста максимальная точность классификации по мере F1 варьирует от 55 до 38 %.

Худшее качество при предсказании и обучении демонстрирует наивный байесовский классификатор. Такие результаты можно объяснить тем, что данный алгоритм считает все признаки независимыми друг от друга, в то время как в векторных репрезентациях все признаки, на самом деле, зависимы.

Для более подробного анализа эффективности алгоритма логистической регрессии в совокупности с векторным представлением ELMo рассмотрим матрицу ошибок (рис. 5).

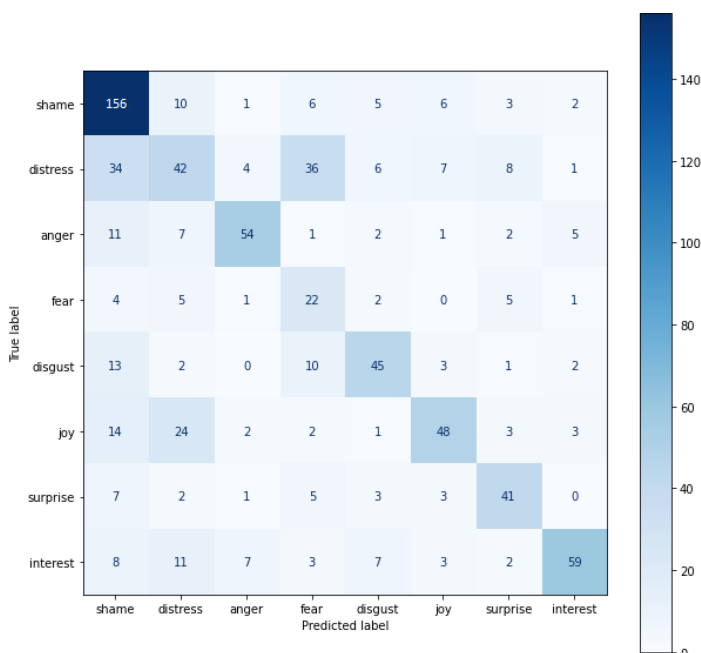


Рис. 5: Матрица ошибок алгоритма логистической регрессии с векторами ELMo

Как видно на изображении (рис. 5), большинство текстов классификатор соотносит корректно, диагональные элементы матрицы явно выражены. Каждая из эмоций в большинстве случаев определяется правильно. Эмоция Грусти иногда ошибочно распознается как Стыд или Страх, а эмоция Радости – как Стыд или как Грусть. Во всех остальных случаях ошибки незначительны.

В таблице 3 представлена числовая оценка эффективности работы классификатора логистической регрессии с векторным представлением ELMo, рассчитанная по метрикам точности, полноты и F1 для каждого из классов эмоций.

Эмоциональный класс	Точность	Полнота	F1
Стыд	82,54	63,16	71,56
Грусть	30,43	40,78	34,85
Гнев	65,06	77,14	70,59
Страх	55	25,88	35,2
Отвращение	59,21	63,38	61,22
Радость	49,48	67,60	57,14
Удивление	66,13	63,08	64,57
Воодушевление	59	80,82	68,21

Таблица 3: Эффективность алгоритма логистической регрессии с векторами ELMo по классам эмоций, %

Табличные данные, как и матрица ошибок, наглядно показывают, что Грусть является наименее распознаваемой эмоцией: к данному классу корректно относится только каждый третий текст. С наибольшей точностью система детектирует Стыд: 8 из 10 предсказаний совпадают с фактической оценкой ассессоров. Также достаточно хорошо определяются эмоции Удивление и Гнев. Отдельного внимания заслуживает класс «Страх», который определяется с относительно приемлемой точностью, но выделяется низкой полнотой. Как мы ранее упоминали, грустные тексты достаточно часто ошибочно относят к категории страшных, что и вызывает соответствующее снижение полноты классификации.

К сожалению, в силу ряда причин, в основном связанных с несовпадением технического стека, в данном эксперименте по классификации были упущены современные архитектуры – «трансформеры», которые можно использовать в качестве входных векторов с последующей тренировкой классификатора. Однако данные модели нашли свое применение во втором эксперименте с регрессией.

5 Решение задачи регрессии

Задача регрессии в контексте анализа эмоций выглядит, на наш взгляд, более интересной так как позволяет предсказывать числовую переменную, а не дискретную, как в случае с классификацией. Такой подход делает анализ более нюансированным и точным. Нами уже были предприняты некоторые эксперименты по применению регрессионных подходов в [9]. Однако в данной публикации мы хотели бы обсудить результаты последующих экспериментов.

При помощи классических алгоритмов (DecisionTreeRegressor, ExtraTreeRegressor, RandomForestRegressor, MLPRegressor – строки 1–4 в табл. 4) мы не получили результатов приемлемого качества, поэтому в данной работе мы решили обратиться к архитектурам типа «трансформер», базовая схема которых изложена в [15].

В настоящее время существует значительное количество уже подготовленных языковых моделей, таких как BERT, RoBERTa, GPT-2, XLNet и других эмбедингов на основе трансформер-архитектуры. Эти модели хорошо себя показывают для разных текстовых задач, но у них есть ряд серьезных ограничений:

- Недостаточные возможности для получения агрегированного вектора текста из векторов представлений отдельных токенов, входящих в текст. Эмбединги на базе трансформеров – это плотные векторы, сопоставленные с токенами из текста, поэтому, чтобы использовать их для описания всего текста, мы должны их агрегировать. Простое усреднение – не лучший подход, поскольку теряется много данных. Токен CLS также не является хорошим кандидатом, поскольку он был обучен для задачи классификации – мы стремимся к регрессии.
- Плохая поддержка русского языка. Несмотря на то, что BERT имеет многоязычную версию и есть специально обученный RuBERT, мы не можем их использовать, поскольку исходный набор данных для многоязычной поддержки представляет собой дампы Википедии. Таким образом, он полностью отличается от того источника данных, с которым мы работаем (личные эмоционально насыщенные тексты из социальных сетей)

- Тяжеловесность. Это главный недостаток. Даже самая младшая модель трансформеров занимает около 100 МБ. При таком «весе» ее использование неэффективно на этапе проверки концепции, когда основная цель – собрать наиболее жизнеспособное и легкое решение, чтобы при необходимости доработать его.

Несмотря на обозначенные выше ограничения, было решено опробовать данные подходы в связи с общей высокой эффективностью их архитектур.

Перед нами стоял ряд вопросов:

- Какую базовую модель выбрать?
- Какие функции для оценки ошибки выбрать?
- Как организовать обучение?

Как уже было отмечено выше, несмотря на распространенность архитектуры, все популярные и доступные модели ориентированы на английский язык. Среди инициатив по тренировке моделей можно отметить проект DeepPavlov, в рамках которого тренируются модели, адаптированные под русский язык. Среди серии таких предобученных моделей была выбрана модель “distilrubert-tiny-cased-conversational”. Она отличается относительно небольшим размером 100М параметров, обучена на смешанном датасете из русских текстов Википедии и русскоязычных новостей, а затем сжата из RuBERT методом дистилляции.

Мы реализовали следующую схему (рис.6):

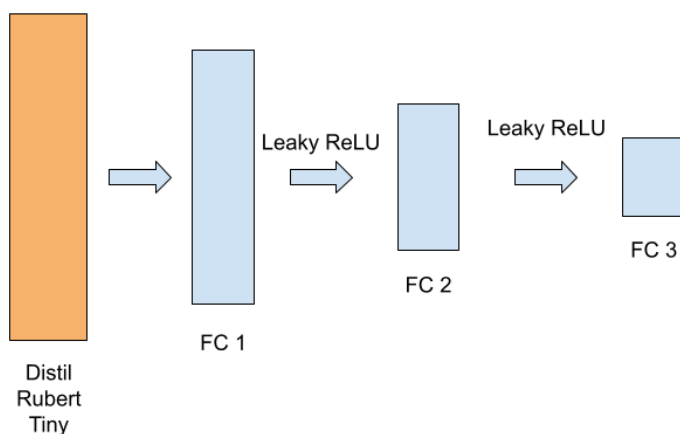


Рис. 6: Схема модели для тренировки регрессора на основе языковой модели

Получив из языковой модели эмбединги для цепочки токенов, мы усредняли их, чтобы получить вектор фиксированной размерности, описывающий весь текст. Усредненный вектор текста проходил через цепочку полносвязных слоев (Fully Connected – FC) и в результате преобразовывался в четырехмерный тензор, соответствующей целевой переменной – значениям диагоналей Куба. В качестве функции активации для перехода использовалась Leaky ReLU. В качестве функции потерь использовалась MSE, в качестве валидационной метрики – MAE.

В данном подходе мы «замораживали» веса языковой модели, и тренировали исключительно блок полносвязных слоев. Такой подход (Frozen BERT) позволяет быстрее тренировать модель за счет меньшего количества тренируемых параметров.

Удовлетворительных результатов добиться не получилось, так при использовании MSE для нескольких целевых переменных, усреднение выходных тензоров поощряло значение 0 для всех целевых переменных. После этого мы решили тренировать также и параметры языковой модели (Fine-tuned BERT). Это заняло существенно больше времени, но и результаты существенно улучшились (см. Табл. 4).

№ п/п	Модель	MSE	MAE
1	DecisionTreeRegressor	10,78	2,40
2	ExtraTreeRegressor	10,32	2,33
3	RandomForestRegressor	6,05	1,82
4	MLPRegressor	6,35	1,95
5	Frozen BERT	6,34	2,57
6	Fine-tuned BERT	3,59	1,36

Таблица 4.: Результаты тренировки на базе регрессионных моделей

Как можно видеть при использовании трансформер-архитектур наблюдается рост метрик сравнительно с использованием классических методов машинного обучения, но данный рост стал возможен только при включении всех параметров языковой модели при дообучении. Исходный код доступен в Colab ноутбуке [https://colab.research.google.com/drive/14PZHxmzAN2SXz_jHzMmW-PW6uu_vxNxj?usp=sharing].

6 Заключение

Проведенные эксперименты в целом демонстрируют, что на данном, размеченном недискретным образом, датасете задача классификации решается с «неровными» результатами – Гнев и Воодушевление показывают достаточно высокие, хотя и недостаточные, значения F1, а Страх и Грусть дают чрезвычайно низкие значения данной метрики.

В то же время регрессионная модель дает неплохие значения метрик MSE и MAE, что можно интерпретировать как признак экологичности процедуры в случае, когда данные недискретной разметки используются для предсказания эмоциональных признаков текста.

Именно регрессионная модель будет использоваться в дальнейших экспериментах и станет основой для разработки приложения для эмоционального анализа русскоязычных текстов социальных сетей

References

- [1] Alexeeva Svetlana. V., Kolcov Sergey N., Koltsova Olesia Yu. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016” Moscow, June 1–4, 2016. pp. 277–288.
- [2] Alm Cecilia O., Roth Dan, Sproat Richard. Emotions from Text: Machine Learning for Text-based Emotion Prediction // Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Vancouver, 2005. pp. 579–586.
- [3] Bland J. Martin, Altman Douglas G. Measurement Error and Correlation Coefficients // British Medical Journal, 1996, Vol. 313 (7048), pp. 41–24.
- [4] Buechel Sven & Hahn Udo. Emotion Analysis as a Regression Problem-Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation // ECAI, 2016, pp. 1114–1122.
- [5] Calvo Rafael A. & Mac Kim Sunghwan. Emotions in text: Dimensional and categorical models // Computational Intelligence, 2013, Vol. 29(3), pp. 527–543.
- [6] Chaturvedi Deepshikha, Chopra Shalu. Customers Sentiment on Banks // International Journal of Computer Applications, 2014, Vol. 98(13), pp. 8–13.
- [7] Ekman Paul. An argument for basic emotions // Cognition & Emotion, 1992, Vol. 6(3-4), pp. 169–200.
- [8] Heras-Pedrosa De Las Carlos, Sánchez-Núñez Pablo, Peláez José. I. Sentiment Analysis and Emotion Understanding during the COVID-19 Pandemic in Spain and Its Impact on Digital Ecosystems // International Journal of Environmental Research and Public Health, 2020, Vol. 17(15), pp. 5542, July 2020.

- [9] Kolmogorova Anastasia, Kalinin Alexander, Malikova Alina. Lövheim Cube-Backed Emotion Analysis: From Classification to Regression. In: Digital Transformation and Global Society. DTGS 2021. Communications in Computer and Information Science, 2022, Vol 1503. Springer, Cham. https://doi.org/10.1007/978-3-030-93715-7_7
- [10] Kolmogorova Anastasia, Kalinin Alexander, Malikova Alina. Semiotic Function of Empathy in Text Emotion Assessment // Biosemiotics, 2021, №14 (3). Cham: Springer Nature. DOI: 10.1007/s12304-021-09434-y
- [11] Lövheim Hugo. A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters // Medical Hypotheses, 2012, Vol. 78, pp. 341–348.
- [12] Önal, I., Ertuğrul, A.M.: Effect of using regression in sentiment analysis. In: 22nd Signal Processing and Communications Applications Conference (SIU), pp. 1822–1825. IEEE, Trabzon (2014). <https://doi.org/10.1109/SIU.2014.6830606>
- [13] Russell James A. & Mehrabian Albert. Evidence for a three-factor theory of emotions // Journal of Research in Personality, 1977, Vol.11(3), pp. 273– 294.
- [14] Talanov Max & Leukhin Alexey & Lövheim Hugo, Viverdú Jordi, Toshev Alexander, Gafarov Fail. Modeling psycho-emotional states via neurosimulation of monoamine neurotransmitters. In: Viverdú, J., Müller, V. (eds.) Blended cognition. Springer Series in Cognitive and Neural Systems, 2019, Vol. 12, pp. 127–157. Springer, Cham (2019). doi: 10.1007/978-3-030-03104-6_6
- [15] Vaswani Ashish & Shazeer Noam & Parmar Niki & Uszkoreit Jakob & Jones Llion & Gomez Aidan & Kaiser Lukasz & Polosukhin Illia. Attention is all you need // Computing Research Repository. — 2017. — Vol. arXiv:1706.03762. —Access mode: <https://doi.org/10.48550/arXiv.1706.03762>.
- [16] Volkova Svetlana, Chetviorkin Ilya, Arendt Dustana, Van Durme Bedjamin. Contrasting public opinion dynamics and emotional response during crisis // Social Informatics, Cham, Switzerland:Springer, 2016. pp. 312–329.

Speaker-agnostic mouth blendshape prediction from speech

Vladislav Korzun^{1,3}

Moscow, Russia

korzun@phystech.edu

Vladimir Berzin^{1,3}

Moscow, Russia

berzin@phystech.edu

Gadecky Dmitry^{1,3}

Moscow, Russia

gadetskiy.dv@phystech.edu

Arkady Ilin^{2,3}

Moscow, Russia

arkady.ilin@skoltech.ru

¹Moscow Institute of Physics and Technology

²Skolkovo Institute of Science and Technology

³Tinkoff

Abstract

This paper describes a simple end-to-end deep learning approach for automated 3D lip animation from audio. Our solution is speaker-independent, which means that once trained on one voice, the model can be applied to any voice without need for retraining. This solution only requires a small amount of data, which can be easily obtained with a modern iPhone. Along with that we also propose a new combined approach for evaluating blendshape prediction models.

Keywords: 3D, lip sync, neural networks, facial animation

DOI: 10.28995/2075-7182-2022-21-323-332

Спикер-независимое предсказание блендшейпов области рта по речи

Владислав Корзун^{1,3}

Москва, Россия

korzun@phystech.edu

Владимир Берзин^{1,3}

Москва, Россия

berzin@phystech.edu

Дмитрий Гадецкий^{1,3}

Москва, Россия

gadetskiy.dv@phystech.edu

Аркадий Ильин^{2,3}

Москва, Россия

arkady.ilin@skoltech.ru

¹Московский физико-технический институт

²Сколковский институт науки и технологий

³Тинькофф

Аннотация

В данной статье описывается простой подход глубокого обучения для автоматизированной 3D-анимации губ из аудио. Наше решение является спикер-независимым, что означает, что после обучения на одном голосе модель может быть применена к любому голосу без необходимости переобучения. Представленный алгоритм требует лишь небольшого количества данных, которые можно легко получить с помощью современного iPhone. Наряду с этим мы также предлагаем новый комбинированный подход для оценки моделей предсказания блендшейпов.

Keywords: 3D, синхронизация губ, нейронные сети, лицевая анимация

1 Introduction

3D lip-sync estimates lip motion corresponding to the audio recording of a person's speech. It is a core problem of avatar head animation extensively studied for decades, as the realistic lip motion heavily affects the perception of liveliness and decreases the uncanny valley effect (Mori et al., 2012).

Several approaches to 3D face animation have been proposed in recent years. The majority of them as in (Karras et al., 2017) and (Cudeiro et al., 2019) aim to create a sequence of entire face meshes based on a speech recording. These models are quite large, and they require high-quality training data, which is difficult to obtain without specialized equipment.

In this paper, we present a simple approach for speaker-agnostic 3D lip-syncs through blendshapes generation. We use an affordable motion capture approach and a small amount of data to train our network. Our system could be trained on one person's speech and used on a different voice as it mostly relies on speaker-independent audio feature encoding and light audio feature to blendshape decoder. By evaluating several audio encoders, we show that good audio features are more useful than a complex model in this task, and there is no need to consider the long context of the phrase.

To sum up, our contributions are as follows:

1. Lightweight speaker agnostic audio-to-blendshape model
2. A new combined approach to evaluating synthesized blendshapes

Our paper is organized in the following way: Section 2 describes related work, section 3 presents our method, section 4 describes experiments and section 5 contains the conclusion.

2 Related Work

The creation of a three-dimensional facial animation based on speech could be done in several ways. First: direct face mesh generation. In this method, a 3D mesh is represented as a set of 3D vertices $\vec{p} \in R^{3N}$ with fixed topology. In paper (Karras et al., 2017) the coordinates of the facial grid vertices are generated directly from the audio feature input window with an additional emotional state. Thus, the entire animation is generated from the sliding window frame by frame. This system has been trained on 3-5 minutes of high-quality 3D scanning sequences of a particular person. This system could be also used on a different voice, but the face 3D model itself is fixed. At the same time, obtaining 3D scans for other faces is a separate challenge and imposes certain limitations on the application of this approach.

To overcome the aforementioned problem with different faces, the parametric face models could be used, such as 3DMM (Banz and Vetter, 1999), FLAME (Li et al., 2017) or FaceWarehouse (Cao et al., 2013). Parametric face models could be represented as a function $\mathcal{M}(\vec{\theta}) = \vec{p} \in R^{3N}$, where $\vec{\theta}$ - set of parameters. The division of parameters may be different for different models. The FLAME model, for example, separates the parameters into three groups: one represents the face form, another its position, and the third its expression. This division may be useful, for example, for transferring animation from one face to another, changing only face shape parameters.

There are several facial animation models based on parametric face models. For example, VOCA (Cudeiro et al., 2019) uses FLAME by generating 3D meshes in the same topology. As a result, it can be utilized to adjust numerous factors during inference, such as facial shape. Authors also provided VOCASET, a large dataset of 4D scans. Although this method can be used to create facial motion for a variety of faces, the model itself is computationally demanding as it generates the entire face mesh.

Another preferred representation to encode facial animations in CGI production are blendshapes (Lewis et al., 2014). Blendshapes could be used to represent any facial expression as a weighted combination of basis vectors. As a result, any face mesh might be computed as $\vec{p} = \mathcal{B}w$, where $\mathcal{B} \in R^{3N \times m}$ - basis, $w \in R^m$ - vector of coefficients. The term "basis" usually refers to basic facial expressions such as an open mouth or a closed eye that are chosen to be as independent as possible. It allows w to be meaningful. For example, if we want a facial expression with a half-opened mouth and a closed left eye we take w with a coefficient of 0.5 for the opened mouth and a coefficient of 1 for the left eye. All the other coefficients are set to zero. It's also worth noting that basis vectors usually reflect the margin from idle faces, rather than the mesh itself, i.e. $\vec{p} = \vec{p}_0 + \mathcal{B}w$. As a result, the facial movement could be

represented as a series of w_i , each of which has a dimension fewer than the number of vertices in the 3D face model.

It is worth noting that blendshapes have already been used as an intermediate representation for neural face puppetry. In (Thies et al., 2020) authors create a latent expression vector from audio. Audio expressions could then be interpreted as blendshapes coefficients of a person-specific generic 3D face model. The mapping from audio-expressions to blendshapes coefficients is trained individually for each person by minimizing the vertex-to-vertex distance between obtained and visually tracked coefficients. Finally, the face mesh obtained from blendshapes is used to generate the final image using neural render.

3 Method

3.1 Problem

In this section, we describe our method for generating facial animations based on blendshape values prediction. First and foremost, let us define the problem. Given a digital audio signal, $w(t), t \in [1, T]$ generate a sequence of corresponding blendshape vectors $\{w_j\}_{j \in 1, M}, w_j \in [0, 1]^m$. Audio signal could be represented as a sequence of audio features $\{a_i\}_{i \in 1, N}, a_i \in R^m$. Here N is not equal to M , as the audio features can have different frame rate from target animation.

3.2 Data

First of all, we require appropriate data to train our models. The most challenging part is to acquire blendshapes. A 3D mesh and a basis B are required to obtain the blendshapes weights vector w . Both are difficult to collect and necessitate the use of expensive equipment to scan a 3D mesh from a human face as well as manual adjustments to fine-tune the basis. However, there is an affordable solution to collect blendshapes directly. Using the depth sensor from the most recent iPhones, LiveLinkFace can extract blendshapes from the frontal camera as well as RGB video with audio. We recorded a small dataset with our team members using this application. It is worth noting that we initially attempted to record data during zoom calls. It has the potential to make the collected facial motions more natural and expressive, but we were unable to achieve a significant improvement in results using this source. Then we focused on the method used to collect the GRID dataset (Alghamdi et al., 2018), in which participants were asked to record 100 short phrases. Following this concept, we created a corpus of short phrases found on the internet. Our corpus contains 2500 uttered phrases of two different speakers (2000 and 500 phrases respectfully) ranging in length from three to ten words. We also use additional 20 long phrases for different voices including synthesized ones to test our models.

The input data was captured at 60 frames per second and included 61 blendshapes. We only use 33 of them, picking only those responsible for the lower half of the face. As a result, we do not cover blinking and eye movements as according to recent works (Chen et al., 2020), movements in these areas cannot be unambiguously determined from speech. We also downsampled our data to 30 frames per second.

3.3 Audio processing

For our model to be used on a different voice, the audio encoder should be able to produce speaker-agnostic audio features. We discovered two major approaches to acquiring such features.

For starters, we could use pretrained embeddings, which are commonly used in speech recognition. Wav2vec 2.0 (Baevski et al., 2020) - a framework trained to produce context representations of audio features - is one of them. This model consists of two major components: a multi-layer convolutional feature encoder and a transformer that builds contextual representations. These representations are used to train our model as audio features.

Wav2vec model showed state-of-the-art results on speech recognition, but its audio embeddings may have some drawbacks in our task. To demonstrate the possible issue we use two subsets of our data. Both have the same phrases recorded by two different speakers. Then, over these subsets, we construct 2D t-SNE (Van der Maaten and Hinton, 2008) projections for audio features and corresponding blendshapes. Finally, we highlight points corresponding to the same phoneme of the one phrase for both speakers with different colors. We found that audio embeddings are projected to almost the same points for both

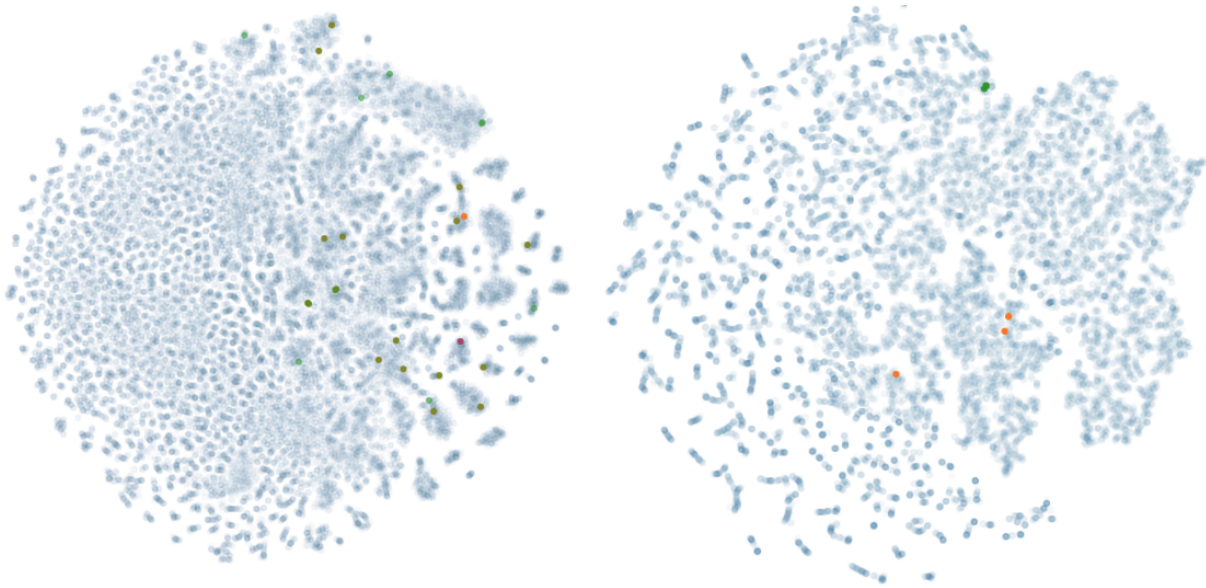


Figure 1: t-SNE projections for wav2vec features (left) and blendshapes (right). Green (first speaker) and Orange (second speaker) colors represent small windows around the same phoneme for two speakers. Dark Green - the intersection of windows for two speakers (orange+green), and Purple - intersection phonemes embeddings itself

speakers while corresponding blendshape projections are far from each other (Figure 1). This difference in the value of the target variable could affect the convergence of our models when we use data from more than one speaker or the same speaker but in different conditions.

To address this issue, we should employ features that preserve the *timbre* of the voice while removing the speaker’s identity. Voice conversion is one of the more straightforward approaches that could be used: given an input speech recording, convert it to sound like a different speaker. Furthermore, the same approach was previously applied to 2D face animation. The authors of MakeItTalk (Zhou et al., 2020) used AutoVC (Qian et al., 2019) to convert input speech to the fixed speaker. Following the same procedure, we were able to train our model on one speaker and then apply it to different voices without having to retrain the entire pipeline.

AutoVC is a tricky intermediate space autoencoder. In this model, the encoder encodes the input Mel-spectrogram and passes it along with the speaker embedding to the decoder. The main idea is to find a dimension of intermediate space that allows only the content of speech to be encoded while retaining the speaker style information from embeddings. We take a pre-trained AutoVC model from MakeItTalk and convert all input Mel-spectrograms, including training data, to a single voice.

3.4 Blendshapes vector prediction

We suggest that our problem is similar to the generation of gestures for 3D skeletons. Both problems can be thought of as sequence-to-sequence problems where sequence audio features are considered as input and a sequence of real-value vectors as output. For gesture generation, these output vectors reflect joint rotation angles, and in our task, these vectors will contain blendshape values. As a result, we can facilitate the approach that produced good results in gesture generation.

First, like in (Kucherenko et al., 2019) we attempted to generate a blendshapes vector from a fixed-length window of audio features. We encode a window of audio features using a recurrent neural network, then pass this encoding through a simple perceptron to obtain a blendshapes vector for a single frame. Here we do not use additional encodings of blendshapes, such as those from autoencoder. We also use the Savitzky–Golay filter (Savitzky and Golay, 1964) during inference to smooth predicted animation.

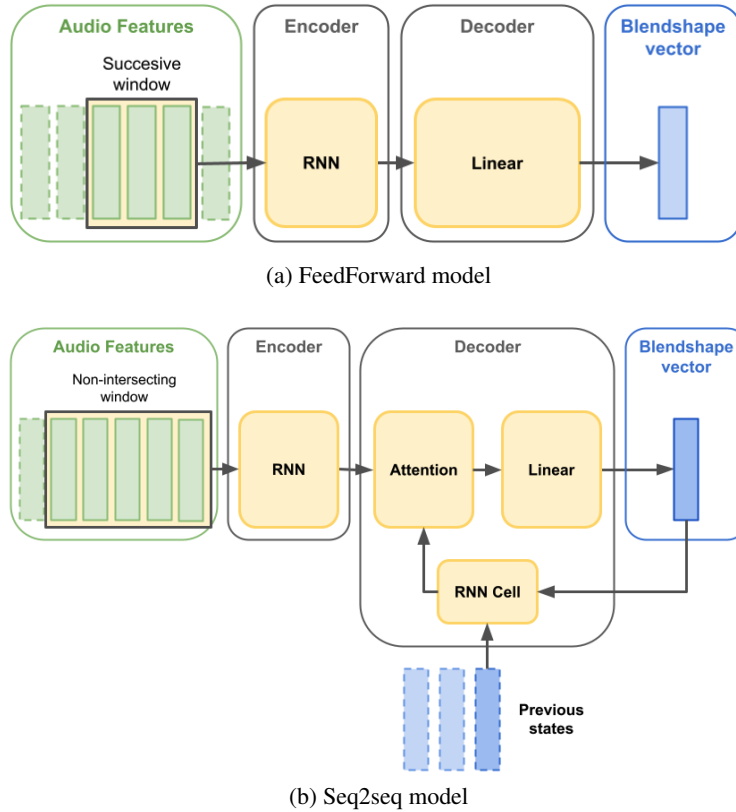


Figure 2: Architectures of FeedForward and Seq2seq models

The full architecture of the FeedForward model is visualized in Fig. 2 (a)

The second model is similar to the one proposed in (Korzun et al., 2021). Following this work, we employ a modified variant of sequence to sequence model. It is made up of two primary parts. First is a recurrent audio encoder which produces contextual features. Second - a recurrent decoder with dot-product encoder-decoder attention uses these features to predict final blendshapes vectors. As in the aforementioned paper, we initialize decoder states with the outputs of the previous state. Figure 2 (b) depicts the entire architecture of the Seq2seq model.

For training, we implement simple MSE loss with no additional continuous and variance losses as in mentioned work for several reasons. To begin with, lip motion does not need to be as varied as body motion. As a result, we were able to give up on variance loss. Second, our experiments revealed that predicted motion is already smooth enough. Although, there are jerks between sequences. To eliminate them, we use the Savitzky–Golay filter.

3.5 Training

To train our models we use the PytorchLightning framework. We use every 10th sample of the first speaker data as validation data. And also every 10th sample of the second speaker as test data. Thus, we have 1800 and 200 samples of first data as train and validation respectively along with 450 samples of the second speaker for additional train data. The remaining 50 samples are used as a test subset. All our models were trained with Adam optimizer with a learning rate set to 1e-3. We also use the early stopping technique with a focus on validation loss.

4 Experiments

In this section we first present our experimental setup, then we evaluate our models and discuss the efficacy of our approaches.

We train our models on several subsets of data to find the dependency of generated animation quality on data quantity. Let us define designations for our models and datasets. There are feedforward (ff) and sequence-to-sequence (s2s) models trained on audio features from Wav2vec2 (w2v) and transferred Mel spectrograms via AutoVC (avc). There are 4 different models in total: ff_avc, ff_w2v, s2s_avc and s2s_w2v. We also train our models on different breakdowns of the training dataset: subsets 1-3 consist of various tracks of the first speaker with 450, 900, and 1800 samples respectively. It is worth mentioning that subsets were recorded separately from each other in several takes with a significant time gap between them. Subset 4 contains all the training data for the first speaker (like subset 3), along with 450 samples of the second speaker’s tracks. Therefore, there are 16 experiments in total.

Evaluation metrics For quantifying the performance of our method, we compute the following three metrics.

- **L_2 distance:** We calculate simple L_2 distance between each generated blendshapes vector and corresponding target one;
- **Mean landmark distance (LMD):** Following the method first described in (Chen et al., 2018) and utilizing the dlib model (King, 2009), we detect the sequences of landmark positions from the 3D face model animation based on real blendshapes, collected from a real speaker, and separately detect landmark positions from animation based on the blendshapes, predicted from corresponding speakers speech. As in (Zhou et al., 2020) we also normalize landmarks over mouth width. Then we calculate the distance between the resulting sequences.
- **SyncNet confidence and minimum distance:** Following (Chung and Zisserman, 2016) we use SyncNet model to obtain audio and video embeddings and measure synchronization between audio sequence and facial movements. To evaluate the change in behavior of this metric on the animated 3D model as opposed to the real face we also take additional measurements of this metric on permuted audio sequences and animations, where the audio is intentionally taken from one track and the facial movements are from a completely different track. See section 4.2 for more details.

4.1 LMD and L_2

Here we show tables with results for different experiments. Firstly, standard L_2 norm and landmark distance metrics can be calculated between real blendshapes and predicted ones. The results are showed in tables 1 and 2. To understand the trustworthiness of LMD we also calculated the distance for permuted corresponding samples. This test showed an increase in the measured distance when applied to asynchronous data. Here LMD for permuted data was equal to 0.75.

model	subset ₁	subset ₂	subset ₃	subset ₄
ff_avc	2.79	2.37	2.62	1.50
ff_w2v	2.58	2.14	2.82	0.77
s2s_avc	2.17	2.06	2.24	0.95
s2s_w2v	2.28	1.90	2.54	0.77

Table 1: L_2 distance, 1e-3

subset ₁	subset ₂	subset ₃	subset ₄
0.590	0.549	0.543	0.467
0.453	0.450	0.460	0.356
0.546	0.555	0.506	0.468
0.462	0.463	0.500	0.377

Table 2: LMD

Here we can see the correlation between L_2 and LMD. It’s also worth noting that as the amount of data for one speaker increases, the metrics do not improve. Adding extra data, especially for wav2vec features, will only degrade the output. Only by including data from a second speaker (subset 4) are we able to enhance metrics significantly. This impact could be attributable to two factors: the test dataset and subset 4 share the same speaker, or the model’s generalization ability has improved. To prove this hypothesis, we test our models on different voices in section 4.3. It is also worth noticing that wav2vec features give better results and the difference between seq2seq and feedforward models is not noticeable. That could mean that lip motion depends only on a short context.

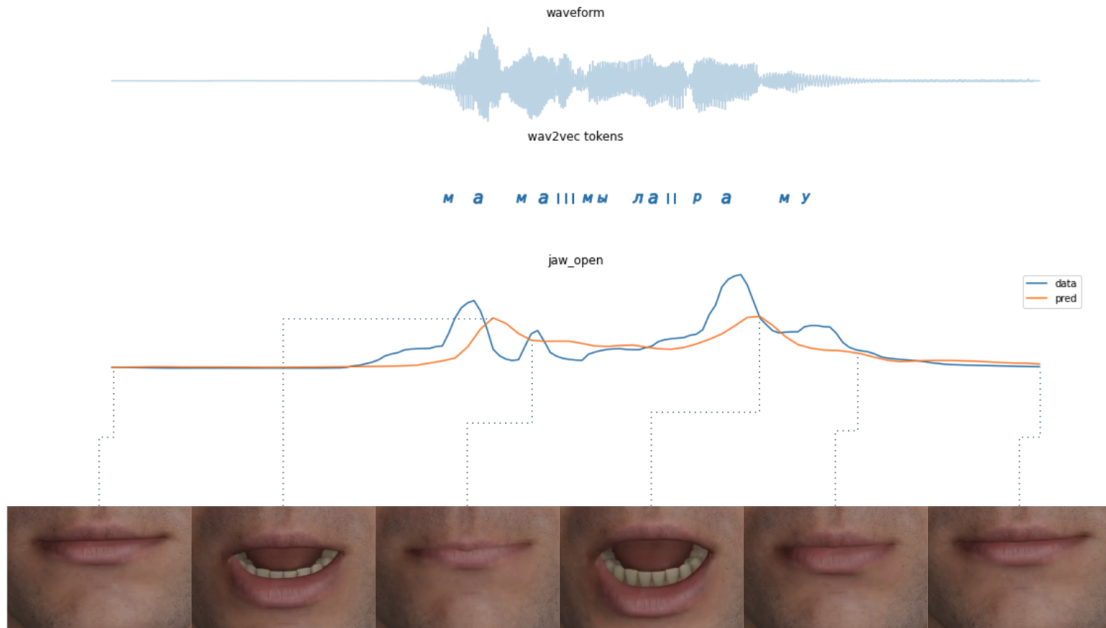


Figure 3: Sample of our model

In figure 3, we demonstrate how one of our models operates on a test sample. For this example, we use a feedforward model with a wav2vec audio encoder (ff_w2v) trained only on a single speaker (subset 3). There are no such phrases nor this speaker in the training dataset.

Figure 3 contains the input raw audio, corresponding tokens acquired by wav2vec, values of *jaw_open* blendshape values for real data (blue) and predicted result value (orange) along with several relevant 3D model renders. Despite modest differences in blendshape value lines, the model was able to reproduce the majority of phonemes.

Even though it has lost the second peak corresponding to the second vowel, this difference did not strike the eye during manual viewing of the produced video, but was plainly visible upon closer study.

4.2 Evaluation of SyncNet metric

Next we consider SyncNet confidence and distance collected during this experiment. The results can be seen in section tables 3 and 4.

model	subset ₁	subset ₂	subset ₃	subset ₄
ff_avc	10.21	9.89	9.61	9.24
ff_w2v	9.17	9.25	8.89	8.39
s2s_avc	10.36	10.61	9.69	9.12
s2s_w2v	9.64	9.54	9.91	8.54

Table 3: SyncNet distance

subset ₁	subset ₂	subset ₃	subset ₄
3.71	3.94	4.15	4.76
5.05	4.85	5.31	5.65
3.37	3.23	4.25	4.54
4.43	4.23	4.21	5.37

Table 4: SyncNet confidence

The SyncNet confidences also show growth if we add the second speaker. It also proves the assumption that the wav2vec features are better and the seq2seq model does not outperform the more simple feedforward model.

To further justify the application of the SyncNet metric to evaluate our model, we conducted the following experiment on the test dataset. First, we used all 61 blendshapes in the model to assess confidence and distance over the original test tracks with no alterations (Table 5, *All BS* column). Following that, we

Metric	All BS	Mouth BS	Mouth BS / wrong track
Distance	8.55	8.58	11.95
Confidence	5.69	5.66	2.34

Table 5: SyncNet distance and confidence on consistent and inconsistent tracks

used the same tracks but limited the number of blendshapes to those corresponding to the mouth region motions, i.e., to 33. (Table 5, *Mouth BS* column). Finally, after animating the same 33 blendshapes, we swapped their respective audio tracks at random. (Table 5, *Mouth BS / wrong track* column). Finally, within each group, the confidence and distance values obtained during each experiment are averaged.

As a result of these findings, limiting the number of blendshapes used in our model’s training and inference has no significant impact on the metric’s behavior. Furthermore, the metric unambiguously identifies cases of audio recording substitution or divergence.

4.3 Synthetic data

In addition to the recordings of a real speech, artificially synthesized speech can also be fed into our model. As an example, we dubbed 20 longer tracks with the voices of Alena and Dorofeev from Tinkoff’s VoiceKit speech synthesis service, and then added the same lines taken from a human speaker for comparison. The SyncNet scores can be seen in the tables 6 and 7. For each model, we use the subset with the best score.

model	Real	Alena	Dorofeev
ff_avc (subset ₁)	9.50	11.44	10.85
ff_w2v (subset ₁)	9.11	11.17	11.22
s2s_avc (subset ₃)	9.60	12.66	12.22
s2s_w2v (subset ₄)	8.88	10.96	11.24

Table 6: SyncNet distance for real and synthetic recordings

Real	Alena	Dorofeev
3.30	3.14	4.26
3.82	3.61	4.13
3.06	1.83	2.67
3.82	3.28	3.54

Table 7: SyncNet confidence for real and synthetic recordings

When compared to samples from the test dataset, this result shows some metric degradation on these tracks. This includes the voice of the real person. The main difference, we believe, lies in the length of the tracks. The test dataset is primarily comprised of short phrases, whereas these tracks are at least twice as long. Visual inspection does not reveal a significant reduction in motion quality.

Despite this, synthetic voices can sometimes outperform natural voices in terms of SyncNet scores. It could be explained by the synthetic voice’s speech smoothness. Women’s synthetic voices, on the other hand, have a significant drop in scores, possibly due to the different timbre. It’s also worth noting that the best results came from different subsets. It could lead to the conclusion that the best results on subset 4 were obtained in section 4.1 due to the presence of the same speaker, and that these models could be trained on just one speaker.

We also try to find out an explanation of SyncNet behavior. We manually examine the phrase where models have been scored differently. In the prediction with a lower score, the mouth did not close completely, while in the other it did. Nevertheless, even the first model was relatively correct in reproducing phonemes visually.

4.4 User study

We also provided short blind comparison between some of the models. We record a quite long phrase with the length of 22 seconds for synthetic Dorofeev’s synthetic voice. Then we render videos for all models and subjectively pick the best four of them. To compare models left we ask users to choose the best from them in a following way. First, we pick a first pair of videos randomly and show them to the

user simultaneously to choose the best of two. Then, we do the same with a second pair left. Finally, we ask user to pick the best video between the winners of a previous stage. The results of the user study are shown in table 8.

In addition to the model’s victory in the second stage when compared to the winner of another pair (first column - Abs Winner), we collect statistics on how many times the model passed to the next stage regardless of the model’s victory in the second stage (second column - Pairwise winner).

model	Abs Winner	Pairwise winner
ff_avc (subset ₁)	13	39
ff_w2v (subset ₄)	8	28
ff_w2v (subset ₁)	14	36
s2s_w2v (subset ₄)	8	26

Table 8: User study

More simple models performed better in a user study on a synthetic phrase. This result is highly correlated with SyncNet scores on the same voice. This observation could provide yet another reason to use SyncNet as an objective metric for facial animation generation.

5 Conclusion

We propose a simple approach for automatic lip-sync for 3D face models. We show that good audio features are more useful than a complex model in this task and that no lengthy context is required. We also make some observations about the audio encoders that are used and data dependency. Our method has its flaws, and generated motions aren’t perfect, but they’re good enough to avoid catching the eye and annoying.

We are also looking for ways to make predicted motion even better. First, we are thinking about conducting a more in-depth user study of our models. Second, we are looking for a way to incorporate more data from videos with expressive facial movements.

We believe that our work will aid in the creation of automatic character animation and will serve as a solid foundation for future research.

6 Acknowledgements

The reported study was funded by RFBR according to the research project № 20-31-90051

References

- Najwa Alghamdi, Steve Maddock, Ricard Marxer, Jon Barker, and Guy J Brown. 2018. A corpus of audio-visual lombard speech with frontal and profile views. *The Journal of the Acoustical Society of America*, 143(6):EL523–EL529.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.
- Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3d faces. // *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, P 187–194.
- Chen Cao, Yanlin Weng, Shun Zhou, Yiyong Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425.
- Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip movements generation at a glance. // *Proceedings of the European Conference on Computer Vision (ECCV)*, P 520–535.
- Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. 2020. Talking-head generation with rhythmic head motion. // *ECCV*.

- Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. // *Asian conference on computer vision*, P 251–263. Springer.
- Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. 2019. Capture, learning, and synthesis of 3d speaking styles. // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, P 10101–10111.
- Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12.
- Davis E. King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758.
- Vladislav Korzun, Dimov Ilya, and Zharkov Andrew. 2021. Audio and text-driven approach for conversational gestures generation. *Dialogue*, 20:425–432.
- Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. // *Proceedings of the ACM International Conference on Intelligent Virtual Agents, IVA '19*, P 97–104.
- John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. 2014. Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports)*, 1(8):2.
- Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1.
- Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, 19(2):98–100.
- Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. // *International Conference on Machine Learning*, P 5210–5219. PMLR.
- Abraham Savitzky and Marcel JE Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.
- Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. // *European conference on computer vision*, P 716–731. Springer.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15.

RuArg-2022: Argument Mining Evaluation

Evgeny Kotelnikov¹, Natalia Loukachevitch², Irina Nikishina³, and
Alexander Panchenko³,

¹Vyatka State University, ²Lomonosov Moscow State University,

³Skolkovo Institute of Science and Technology

kotelnikov.ev@gmail.com, louk_nat@mail.ru,

{irina.nikishina, a.panchenko}@skoltech.ru

Abstract

Argumentation analysis is a field of computational linguistics that studies methods for extracting arguments from texts and the relationships between them, as well as building argumentation structure of texts. This paper is a report of the organizers on the first competition of argumentation analysis systems dealing with Russian language texts within the framework of the Dialogue conference. During the competition, the participants were offered two tasks: stance detection and argument classification. A corpus containing 9,550 sentences (comments on social media posts) on three topics related to the COVID-19 pandemic (vaccination, quarantine, and wearing masks) was prepared, annotated, and used for training and testing. The system that won the first place in both tasks used the NLI (Natural Language Inference) variant of the BERT architecture, automatic translation into English to apply a specialized BERT model, retrained on Twitter posts discussing COVID-19, as well as additional masking of target entities. This system showed the following results: for the stance detection task an F1-score of 0.6968, for the argument classification task an F1-score of 0.7404. We hope that the prepared dataset and baselines will help to foster further research on argument mining for the Russian language.

Keywords: Argumentation Mining, Stance Detection, Premise Classification, COVID-19

DOI: 10.28995/2075-7182-2022-21-333-348

RuArg-2022: соревнование по анализу аргументации

Евгений Котельников¹, Наталья Лукашевич², Ирина Никишина³,
и Александр Панченко¹

¹Вятский государственный университет,

²Московский государственный университет им. М. В. Ломоносова,

³Сколковский институт науки и технологий

kotelnikov.ev@gmail.com, louk_nat@mail.ru,

{irina.nikishina, a.panchenko}@skoltech.ru

Аннотация

Анализ аргументации – это область компьютерной лингвистики, в которой исследуются методы извлечения из текстов аргументов и связей между ними, а также построения аргументационной структуры. Настоящая статья представляет собой отчет организаторов о первом соревновании русскоязычных систем анализа аргументации в рамках конференции «Диалог». В ходе соревнования участникам были предложены две задачи: определение позиции автора по заданной теме и классификация доводов. Для обучения и тестирования систем был подготовлен и размечен корпус, содержащий 9,550 предложений (комментариев к постам в социальных медиа) по трем тематикам, связанным с пандемией COVID-19: вакцинация, карантин и ношение масок. Система, занявшая первое место по обеим задачам, использовала NLI вариант (Natural Language Inference – вывод по тексту) применения архитектуры BERT, автоматический перевод на английский язык для использования специализированной модели BERT, дообученной на постах Твиттера, обсуждающих ковид, а также дополнительное маскирование целевых сущностей. Эта система показала следующие результаты: для задачи определения позиции F1-score=0.6968, для задачи классификации доводов F1-score=0.7404. Мы надеемся, что подготовленные наборы данных и методы помогут стимулировать дальнейшие исследования по анализу аргументации для русского языка.

Ключевые слова: анализ аргументации, определение позиции автора текста, классификация доводов, COVID-19

1 Introduction

People have been constantly arguing at all social levels and the *Argumentation Theory* was developed to study and control the process of coming to a conclusion from premises through logical reasoning. According to this theory, an argument must include a *claim* containing a *stance* towards some topic or object, and at least one *premise* (“favor” or “against”) of this stance. Often a “premise” is called an “argument” when it is clear from the context which claim it is being referred to.

With the development of intellectual systems and neural networks, arguments can now be both produced and studied automatically. Therefore, the *Computational Argumentation* task arose to address the problem of computational analysis and synthesis of natural language argumentation. In this paper we focus on its branch — *Argument Mining* (or *Argumentation Mining*) — which explores methods for extracting arguments and their relationships from texts, as well as constructing an argumentative structure.

There is a large number of works devoted to this field which are thoroughly reviewed by (Stede and Jodi, 2018; Lawrence and Reed, 2020; Stede, 2020; Vecchi et al., 2021; Schaefer and Stede, 2021). Special attention has been paid to the stance detection as a sub-task of Argument Mining (Küçük and Can, 2020; ALDayel and Magdy, 2021; Küçük and Can, 2021) where the authors describe the proposed approaches so far, descriptions of the relevant datasets and tools, and some other related issues.

The growing interest in the task is justified by the application of the Argument Mining algorithms for argument search, fact checking, automated decision making, argument summarization, writing support and intelligent person assistants. For instance, Args.me¹, ArgumenText² and CAM (comparative argumentative machine)³ (Schildwächter et al., 2019) are well known systems widely used for searching arguments.

The main research forum for the task is the Argument Mining workshop series. Since 2014, eight workshops on the analysis of arguments have already been held⁴ addressing burning issues like multi-task learning (Tran and Litman, 2021; Putra et al., 2021) and Argumentation Mining in different areas (science (Lauscher et al., 2018; Fergadis et al., 2021), news articles (Bauwelinck and Lefever, 2020) and cross-lingual research (Rocha et al., 2018)). Moreover, there are several shared tasks on the topic adjacent to the Argument Mining: Shared Task on Argumentation Mining in Newspaper Editorials (Kiesel et al., 2015), SemEval-2016 (Stance Detection) (Mohammad et al., 2016), Touché (Argument Retrieval) in 2020 (Bondarenko et al., 2020) and 2021 (Bondarenko et al., 2021).

In this paper, we present RuArg-2022 — the first shared task on Argument Mining for the Russian language. It consists of two sub-tasks: stance detection and premise classification. The first task aims to determine the point of view (stance) of the text’s author in relation to the given claim. The second task is devoted to classification of texts according to premises (“for” or “against”) to a given claim.

To highlight the differences between the two tasks, consider the following example: *Я против масок, но приходится их носить: мне проще так, чем с кем-то что-то обсуждать и кому-то что-то доказывать* (*I am against masks, but I have to wear them: it’s easier for me than to discuss something with someone and prove something to someone*). In this sentence there is an explicit stance against masks but it gives a premise for wearing masks.

The contribution of the current paper is three-fold. First, we prepare a gold standard dataset for stance detection and premise classification. Second, we develop and release a baseline for the argument mining tasks that uses a multi-task multi-label BERT architecture. Third, we compare and analyse the results of the participants of the shared task and propose steps for further improvement of both sub-tasks. All the materials and data could be found on GitHub⁵ and CodaLab⁶ competition pages.

Thus, our work is the first, to the best of our knowledge, dealing with argument mining task for the Russian language. While our setup is a simple text categorization task, we argue that it may be an

¹<https://www.args.me/index.html>

²https://www.informatik.tu-darmstadt.de/ukp/research.ukp/ukp_research_projects/ukp_project_argumenttext/index.en.jsp

³<http://ltdemos.informatik.uni-hamburg.de/cam/>

⁴<https://2021.argmining.org>

⁵<https://github.com/dialogue-evaluation/RuArg>

⁶<https://codalab.lisn.upsaclay.fr/competitions/786>

important building block of larger argument mining pipelines featuring retrieval of arguments from large text collections (Bondarenko et al., 2020).

2 Previous Work

The Argument Mining task (Palau and Moens, 2009) involves the automatic identification of argumentative structures in free text. According to (Cabrio and Villata, 2020), “researchers have investigated argument mining on various registers including legal texts, scientific papers, product reviews, news editorials, Wikipedia articles, persuasive essays, political debates, tweets, and online discussions”. A detailed overview of all argument mining related papers is out of the scope of the current work. We refer the reader to the recent surveys on this topic: (Lawrence and Reed, 2020) and (Schaefer and Stede, 2021).

The topic of COVID-19 is nowadays popular not only in the biomedical field, but also in social science and, especially, NLP research (Verspoor et al., 2020). There already exist several datasets on COVID-19 for stance detection (Wührl and Klinger, 2021), argument mining and fact extraction/verification. For instance, in (Beck et al., 2021) the authors collect a dataset from German Twitter on people’s attitude towards the government measures. First, they identify relevant tweets for governmental measures and if relevant, detect what stance is expressed. (Menin et al., 2022) create a linked data version of the COVID-19 data set and enriched it via entity linking and argument mining.

Another dataset collected lately (Reddy et al., 2021) comprises the following topics related to COVID-19: origin of the virus, transmission of the virus, cure for the virus and protection from the virus. The authors present a pipeline for detecting claim boundaries and detecting stance. Unlike most stance detection datasets (Hanselowski et al., 2019; Allaway and McKeown, 2020) it involves identifying the claimer’s stance within a claim sentence and not the stance for target–context pairs. (Li et al., 2022) follows (Reddy et al., 2021) and identify the stance from the perspective of each claimer, namely whether the claimer affirms or refutes a claim. They finetune a Bart-large model (Lewis et al., 2020) to automatically identify the stance. One more dataset related to the COVID-19 pandemic is collected by seven science teachers through three scenarios (Atabey, 2021). This dataset contains not only stances about vaccination, curfew and distance education, but also arguments and supporting reasons that might construct an argument mining dataset.

Most of the above mentioned publications on COVID-19 stance detection refer to the FEVER-like dataset COVID-Fact (Saakyan et al., 2021) of 4,086 claims concerning the COVID-19 pandemic. It could be also applied for the argument mining needs. Another dataset for COVID-19 fact checking is presented in (Liu et al., 2020) which also could be reformatted for the argument mining tasks.

As regards argumentation mining for the Russian language, there are not so many studies and datasets on the topic. (Fishcheva and Kotelnikov, 2019) translated into Russian and researched the English language Argumentative Microtext Corpus (ArgMicro) (Peldszus and Stede, 2015; Skeppstedt et al., 2018). In (Fishcheva et al., 2021) this corpus was expanded with machine translation of the Persuasive Essays Corpus (PersEssays) (Stab and Gurevych, 2014). XGBoost and BERT were applied to classify “for”/“against” premises.

Salomatina et al. (Salomatina et al., 2021) propose an approach to the partial extraction of the argumentative structure of a text by using patterns of argumentation indicators. They also try to recognize the relations between extracted arguments. (Irina et al., 2021) develop a web resource for analysis of argumentation in popular science discourse. The annotation model is based on the ontology of argumentation and D. Walton’s argumentation schemes (Walton et al., 2008). A scenario of argument annotation of texts allows constructing an argumentative graph based on the typical reasoning schemes.

To the best of our knowledge, there are no manually labelled publicly available datasets in Russian. In this work we present such dataset for the first time.

3 Dataset

The dataset is based on VKontakte users’ comments discussing COVID-2019 news texts (Chkhartishvili et al., 2021). We choose the COVID-19 pandemic (and anti-epidemic measures in general) as the topic

of the dataset because we assume that the analysis of arguments on social measures against COVID-19 is still relevant for the modern society and especially for the understanding of the current public sentiment. From the gathered comment collection, sentences discussing masks, vaccines and quarantine were extracted using keywords (Nugamanov et al., 2021).

The annotation process included two stages: labelling by stance and labelling by premises. At both stages sentences were labelled in relation to the following claims:

1. “Vaccination is beneficial for society.”
2. “The introduction and observance of quarantine is beneficial for society.”
3. “Wearing masks is beneficial for society.”

In the following subsections we describe the annotation process of the dataset for both sub-tasks: Stance Detection 3.1 and Premise Classification 3.2.

Stance	Premise	Numerical label
for	for	2
other (neutral/contradictory/unclear)	no argument	1
against	against	0
irrelevant	irrelevant	-1

Table 1: System of categories used to label the dataset.

3.1 Stance Annotation

The current dataset has been already annotated in (Nugamanov et al., 2021). In the current work the dataset was additionally checked and synchronized with premise annotation from the second step.

At the first stage of stance annotation each sentence was labelled by several experts (three on average). An annotator should indicate the stance it expresses towards each of the above-mentioned aspects (or indicate that the sentence is not relevant to the aspect). The annotators’ group included professional linguists and psychologists. We consider four stance labels, namely:

- **for:** positive stance, which means that the speaker expresses his support for the topic;
- **against:** negative stance — the topic of discussion is not endorsed by the speaker;
- **other:** neutral stance (this label is used for factual sentences without any visible attitudes from the author); contradictory stance (for such a label, evident positive and negative attitudes should be seen in a message); unclear stance (the presence of a stance is seen, but the context of sentence does not give possibility to determine it);
- **irrelevant:** text does not contain stance on the topic.

The coding scheme for the stance annotation is presented in Table 1.

A sentence is considered to be relevant to an aspect, if at least two annotators considered it relevant. Sentences collected using keywords also can be irrelevant, for example a sentence mentioning Elon Musk (“Mask” in Russian spelling) is not relevant to the mask aspect.

3.2 Premise Annotation

At the second stage of annotation, the dataset was also annotated by premises for all three claims. The following four classes (labels) were used:

- **for:** the stance is supported with argument in favor of the topic;
- **against:** the argument explains the author’s negative outlook on the topic;
- **no argument:** no explanation is given for supporting/criticism of the topic;
- **irrelevant:** text does not contain stance and, consequently, premise on the topic.

The annotated sentences from the previous step were divided into three subsets: training, validation, and test (see Subsection 3.4). The labelling of each sentence by premises from the training and validation datasets was carried out by three annotators; the test sentences were labelled by four annotators. The final

labels for training and validation datasets were assigned with the agreement of at least two annotators, for test dataset – with the agreement of at least three annotators.

A sentence was considered as a premise if the annotator could use it to convince an opponent about the given claim, such as “*Masks help prevent the spread of disease.*” Detailed instructions for annotators are available in the competition repository⁷.

The task of premise annotation should be separated from stance detection and sentiment analysis tasks. For example, the following statement does not contain a premise in relation to masks, although there is an author’s stance “for”: *It is high time to involve the city of “brides” in the production of protective masks.* It is also necessary to distinguish between sentiment polarity (positive and/or negative) and argumentation. In the following sentence there is a negative polarity towards quarantine, a positive polarity towards Trump, but no rational premises “for” or “against” quarantine are given: *And the fact that Trump did not introduce a suffocating quarantine is well done!*

The difference between the two tasks is illustrated in Table 2. For example, the first sentence possesses both stance and premise: the speaker expresses his negative attitude towards the vaccine by the reason of its short-term effectiveness. The second sentence, on the contrary, definitely supports vaccination without giving any specific arguments to support his/her opinion.

Text	Masks		Quarantine		Vaccines	
	Stance	Premise	Stance	Premise	Stance	Premise
И какой смысл в вакцине если антитела только 3 месяца? (<i>And what’s the point of a vaccine if the antibodies work only for 3 months?</i>)	—	—	—	—	against	against
Должна быть вакцина которую, будут прививать с детства!!! (<i>There must be a vaccine that will be vaccinated from childhood!!!</i>)	—	—	—	—	for	no argument
Вот только там на момент, когда была 1000 выявленных, уже неделю карантин действовал. (<i>At the time when there were 1000 identified, quarantine had been in effect for a week.</i>)	—	—	other	against	—	—
Развитие ситуации: если соблюдать карантин месяц, то вирус будет остановлен. (<i>The development of the situation: if the quarantine is observed for a month, the virus will be stopped.</i>)	—	—	for	for	—	—
Вопрос к властям :почему из гос резерва не получили люди масок когда их не хватало или и резерва уже нет (<i>Question to the authorities : why didn’t people get masks from the state reserve when there were not enough of them or there is no reserve anymore</i>)	for	no argument	—	—	—	—
Любители масок не ужели вы думаете что эта косметическая тряпочка поможет от вируса?! (<i>Mask lovers don’t you really think that this cosmetic rag will help against the virus?!</i>)	against	no argument	—	—	—	—

Table 2: Examples for each topic — masks, vaccines, and quarantine (we keep the original spelling and punctuation). Note that for each topic, annotation of stances and premises was performed. Refer to Table 1 for the classification schema used to label the data. “Irrelevant” class is denoted as “—”.

3.3 Dataset Verification

After completion of stance and premise annotation procedures, we verified the labels of the dataset. To this end, we looked through the contingency tables of stances and premises, and checked the following issues:

⁷<https://github.com/dialogue-evaluation/RuArg/tree/main/annotation>.

1. the sentence with an irrelevant label for one of the sub-tasks cannot be relevant for another sub-task;
2. the sentence with contradictory stance and premise (e.g., positive stance but premise “against”) should be examined more carefully.

As a result, annotations for 289 sentences (3.0% from the whole dataset containing such issues) were revised and improved. Generally, if a sentence contains both a stance and a premise, then their polarity coincides (both “for” or both “against”). However, in 12 sentences the polarity is opposite. This is due to the fact that the sentence simultaneously contains the author’s point of view and indicates the opponents’ premises, for example: “*This is exactly why everyone should wear masks, but the main channels broadcast that masks are not needed and useless for healthy people.*”

3.4 Dataset Statistics

Each sentence has 6 labels: for each of the two sub-tasks (stance detection and premise classification) there is a label for each of the three aspects (masks, vaccines and quarantine).

The inter-annotator agreement was calculated by Krippendorff’s alpha and it turned out quite high – 0.84. Dataset statistics are presented in Table 3. As one may observe the dataset is skewed (imbalanced). There are various schemes in the literature to perform evaluation of this kind of data (Rosenberg, 2012). We resort to a scheme by simply excluding the largest “irrelevant” class as it is done in Sentiment Analysis. For instance, at the SemEval-2016 Task 14 (Nakov et al., 2016) the organizers exclude the “NEUTRAL” class from the evaluation as the largest one.

The distribution of labels by class is shown in Figure 2

Dataset	Total	Stance			Premise			Irrelevant
		For	Other	Against	For	No argument	Against	
<i>Masks</i>								
train	6,717	704	1,832	594	339	2,451	340	3,587
val	1,431	148	388	126	62	542	58	769
test	1,402	147	401	123	63	523	85	731
all	9,550	999	2,621	843	464	3,516	483	5,087
<i>Quarantine</i>								
train	6,717	587	1 341	172	217	1,756	127	4,617
val	1,431	125	290	39	46	369	39	977
test	1,402	116	274	40	50	358	22	972
all	9,550	828	1,905	251	313	2,483	188	6,566
<i>Vaccines</i>								
train	6,717	374	866	418	149	1,238	271	5,059
val	1,431	78	183	92	24	282	47	1,078
test	1,402	75	181	81	21	262	54	1,065
all	9,550	527	1,230	591	194	1,782	372	7,202

Table 3: Statistics of the constructed dataset used in RuArg-2022 shared task.

4 Evaluation

The main performance metric in each of the two sub-tasks are $F1_{stance}$ and $F1_{premise}$ scores, which are calculated according to the following formula:

$$F1 = \frac{1}{n} \sum_{c \in C} F1_{rel_c}, \quad (1)$$

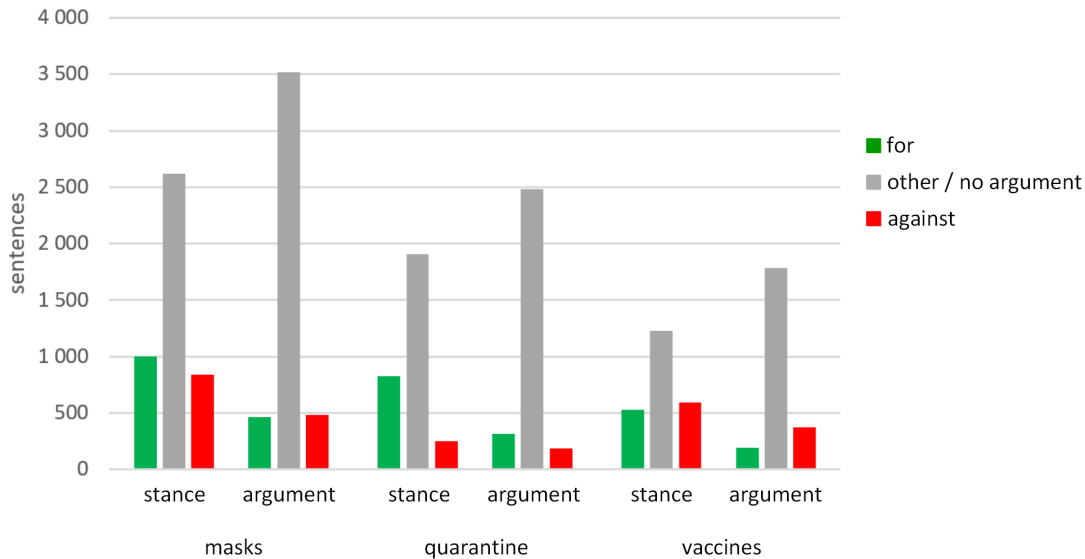


Figure 1: Distribution of the labels.

where $C = \{“masks”, “vaccines”, “quarantine”\}$, n is the size of C and $F1_{rel}$ -score is macro $F1$ -score averaged over first three relevance classes (the class “irrelevant” is excluded). Namely, the following procedure is used:

1. $F1$ -scores are averaged over three out of four classes (the “irrelevant” class is excluded) – macro $F1_{rel}$ -score is obtained for a given claim;
2. macro $F1_{rel}$ -scores for all three claims are averaged – we get macro $F1$ -score relative to the task (stance detection or premise classification);
3. For each of the three claims, $F1$ -score is calculated for each class (label) separately.

As a result, two main macro $F1_{rel}$ -scores are calculated – one for each sub-task. Participants’ systems are ranked by these metrics (two separate lists). The $F1_{rel}$ -score for claims and $F1$ -score for individual classes (labels) will be also discussed in Section 7.

5 Baseline

We implement a simple baseline that finetunes the pre-trained ruBERT model (Devlin et al., 2019; Kuratov and Arkipov, 2019) on the provided dataset. We chose “DeepPavlov/rubert-base-cased” model from Hugging Face⁸. We experiment with training a single model that predicts all the required labels. However, it did not performed well, so we finetune three pre-trained BERT models separately for three topics: “masks”, “vaccines”, and “quarantine”. Each model comprises the following layers:

1. the pre-trained BERT layer with the unfrozen weights;
2. a dense layer for stance detection;
3. a dense layer for argument classification.

Then we applied categorical cross-entropy loss to train on both stance and argument labels simultaneously. The results are presented in Section 7.

6 Participating Systems

RuArg-2022 shared task attracted 16 participants, 13 of them participated in the final phase. We provide descriptions of the top 7 solutions which outperformed the baseline for at least one sub-task. We denote

⁸<https://huggingface.co/DeepPavlov/rubert-base-cased>

each team either with its team name (if any) or with their CodaLab user names. In cases of multiple submissions from one team, we report only the best result. The scores of the teams are shown in Table 4.

camalibi (msu) First, this team used RuBERT-classifier⁹ to determine the relevance of the texts using NLI-method: to form an input example, a second sentence with the aspect (“masks”, “quarantine”, or “vaccination”) was added to each original sentence from the dataset. The output 1 was for the “Relevant” result and 0 for “Irrelevant”.

For the stance classification task the texts were pre-processed and then translated into English using pretrained seq2seq-model¹⁰. Then, each text was processed according to the rule: $keyword \rightarrow @ * ASPECT * keyword @$, where *ASPECT* is the aspect for which a given text is relevant and *keyword* is the word from a list of words corresponding to each aspect.

Then for both RuArg sub-tasks the domain-specific BERT-classifier¹¹ was trained using NLI-method: for each text and each aspect for which a given text is relevant, six input examples were constructed (three for each stance label and three for each premise label). Final input examples looked like:

- “Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.”, “**Against quarantine**”,
- “Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.”, “**None-stance quarantine**”,
- “Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.”, “**In-favor quarantine**”,
- “Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.”, “**Negative to quarantine**”,
- “Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.”, “**Neutral to quarantine**”,
- “Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.”, “**Positive to quarantine**”.

The stance or premise label was chosen as the one where the corresponding input example had the maximum softmax output.

sevastyannm (vyatsu) This participant utilized pre-trained ruRoberta-large language model¹² which was trained on additional data obtained from “PersEssays_Russian” and “ArgMicro_Russian” datasets (Fishcheva et al., 2021) with similar annotation schemes. Both datasets were united by argumentative discourse units and used to train model to solve 4-class classification problem.

First, ruRoberta-large was fine-tuned on the united “PersEssays_Russian” and “ArgMicro_Russian” dataset with 8,780 units. Then, the model was fine-tuned separately on each of 6 tasks from the competition dataset (*'masks_stance'*, *'masks_argument'*, *'quarantine_stance'*, *'quarantine_argument'*, *'vaccines_stance'*, *'vaccines_argument'*). For final class prediction the participant used token averaging and 2-layer linear neural network classifier. All models trained with $learning_rate = 10^{-5}$ and $weight_decay = 0.01$.

For the model trained on the additional dataset the participant used the following hyperparameters: $input_size = 70$, $num_epochs = 3$, $batch_size = 16$. For model trained on the RuArg dataset hyperparameters were as follows: $input_size = 100$, $batch_size = 32$, $num_epochs = \{2, 2, 4, 4, 4, 7\}$ for *'masks_stance'*, *'masks_argument'*, *'quarantine_stance'*, *'quarantine_argument'*, *'vaccines_stance'*, *'vaccines_argument'* respectively.

iamdenay (IICT) This team used the pre-trained Crosslingual RoBERTa-large model and fine-tuned on the augmented data. They mostly augmented the data containing stances and arguments about “*quarantine*”. For the augmentation the participants used mT5 model to paraphrase sentences in order to increase the size of the text set. To increase accuracy of the proposed method they used six different

⁹<https://huggingface.co/DeepPavlov/rubert-base-cased-conversational>

¹⁰<https://huggingface.co/Helsinki-NLP/opus-mt-ru-en>

¹¹<https://huggingface.co/digitalepidemiologylab/covid-twitter-bert-v2>

¹²<https://huggingface.co/sberbank-ai/ruRoberta-large>

models, one per task $\{ 'masks_stance', 'masks_argument', 'quarantine_stance', 'quarantine_argument', 'vaccines_stance', 'vaccines_argument' \}$.

ursdth This team proposed a pipeline-based framework for the classification of texts with or without recognizable rhetorical structure. The first stage involved fine-tuning sequential model on the classification dataset including texts of different lengths and complexity. In the second stage, they frozen the base model and then trained a discourse-aware neural module on top of it for the classification of texts with discourse structure.

They used pre-trained Conversational RuBERT for the discourse unit classification. For texts with automatically recognizable discourse structure, they proposed a relation-aware Tree-LSTM over the discourse units' class predictions. Stance and premise labels were predicted jointly.

Both development and test datasets were treated as unseen, and the official development dataset was not used for the parameters adjustment. The predictions were obtained by averaging outputs from five models trained on cross-validation during experiments over labeled data. This is similar to an ensemble, where each model is trained using 80% of the train data.

sopilnyak (auteam) This team started with training a classifier to detect irrelevant sentences for each sub-task. They applied binary Logistic Regression classifier trained on TF-IDF features, calculated from BPE tokens.

Then they excluded irrelevant sentences and further trained the models (for each sub-task separately) as a blend of:

1. fine-tuned ruRoberta-large from Sber AI with a two-layer classification head on top. They unfroze 30 top layers and used very low learning rate ($5 \cdot 10^{-6}$) to prevent model from over-fitting on a small dataset. Also they utilized weighted cross-entropy loss so that the results on unbalanced dataset would be more accurate.
2. Logistic Regression classifier on TF-IDF features calculated on BPE tokens.

kazzand This participant applied Transformer-based deep text feature extraction and hierarchical classification. Firstly, they trained simple TF-IDF + Logistic Regression pipeline for each text type (masks, quarantine, vaccines). Secondly, they trained 6 separate models for each task $\{ 'masks_stance', 'masks_argument', 'quarantine_stance', 'quarantine_argument', 'vaccines_stance', 'vaccines_argument' \}$ using Sentence-BERT for embeddings computation served as input to the Logistic Regression or KNN model.

invincible The first step for the team was a preprocessing: they removed punctuation symbols, converted text to lowercase, and removed special symbols including the “[USER]” substring. They further used the DistilRuBERT model¹³ to vectorize the text into a vector of numbers and saved as a row of the new matrix. This feature matrix was used as an input for the classification models.

Overall, there were nine models, three for each topic {“masks”, “vaccines”, “quarantine”}. The initial data were separated into three subsets corresponding to each topic. Then the following algorithm was applied: first, SVM model (with sigmoid kernel and balanced target) detected irrelevant sentences for each topic and classified them as “irrelevant” for both stance and argument types. Then for positive-classified sentences, two neural network models were applied. They consisted of Flatten layer, Dense layer with ReLU activation function, Dropout layer and final Dense layer with Sigmoid activation function. They used the following hyperparameters: *optimizer* = “adam”, *loss* = “sparse_categorical_crossentropy”, *num_epochs* = 5.

Importantly, since the classes are highly unbalanced (labels “for” and “against” are highly underrepresented), a random oversampling was applied to all dataframes before fitting the models. Accurate class-balancing allowed improving both scores significantly.

7 Results and Discussion

Table 4 presents respectively the results for “stance detection” and “premise classification” tracks.

¹³<https://huggingface.co/DeepPavlov/distilrubert-tiny-cased-conversational>

#	Participant	Base Transformer model	Additional data	Stance F1-score	#	Premise F1-score	#
1	camalibi	covid-twitter-bert-v2	Yes	0.6968	1	0.7404	1
2	sevastyannm	RuRoBERTa-large	Yes	0.6815	2	0.7235	2
3	iamdenay	RuRoBERTa-large	Yes	0.6676	3	0.6555	4
4	ursdth	RuBERT Conversational	No	0.6573	4	0.7064	3
5	sopilnyak	RuRoBERTa-large	No	0.5603	5	0.4438	10
6	kazzand	Sentence-BERT	No	0.5552	6	0.5603	6
7	morty	n/a	n/a	0.5353	7	0.5453	7
8	invincible	RuBERT Conversational	Yes	0.5286	8	0.5428	8
9	dr	n/a	n/a	0.4750	9	0.6036	5
10	baseline	ruBERT	No	0.4180	10	0.4355	9

Table 4: Competition results of the participant systems.
The places of participants for each sub-task are indicated in the brackets.

All the results are quite stable for both sub-tasks, only **sopilnyak** did not manage to overcome the premise baseline, demonstrating high results (top-5) at the stance detection sub-task. The range of the models used to solve the task is not wide: the participants choose between (ru)BERT, (crosslingual)RoBERTa(-large) and old good Logistic Regression model.

In comparison to the baseline, all the participants trained classification models separately for each sub-task. Evidently, multitask classification is more challenging than training classification models separately.

Interestingly, several best results were obtained with the help of the additional datasets or/and data augmentation (**camalibi** – top-1, **sevastyannm** – top-2, **iamdenay** – top-3 for stance detection and top-4 for premise classification, and also **invincible** – top-8). Top-1 **camalibi** used the special version of BERT model in which domain-oriented dataset was actually integrated; top-2 **sevastyannm** utilized additional dataset, top-3/top-4 **iamdenay** applied mT5 for paraphrase generation, top-8 **invincible** used random oversampling. From these observations we can assume that any kind of additional data is beneficial for these tasks, however, the more diverse the data is, the better.

The most different and outstanding approach in comparison to other participants was presented by the winner system of **camalibi**. This participant applied NLI method which performed best for both sub-tasks. Moreover, model trained on the English language was applied, therefore, **camalibi** did translate the whole dataset for the task.

We also compared the detailed results for the top 5 systems and the baseline. The scores are presented in Appendix A. From Figure 2 we can see that F1-scores for Premise Classification are slightly higher than for Stance Detection. The task of Stance Detection is equally hard for all three topics, whereas we can see that the scores for Masks Premise and Quarantine Premise are higher than Vaccines Premise results. It can be seen that the difference between the top 3 participants are not very much different from each other. As for the baseline results, we can see that the results for vaccine and quarantine are two times lower than the results of the (at least) top 4 participants. At the same time, Masks Stance and Premise results are higher than for vaccines and quarantine and not significantly different from the top results. To sum up, we can conclude that the algorithms for the top 3 results demonstrate similar results across different subsets.

8 Conclusion

We present the results of the first shared task on Argument Mining for Russian. For this shared task, we created a new dataset on the vital COVID-19 topic. We introduce and rely on the following claims:

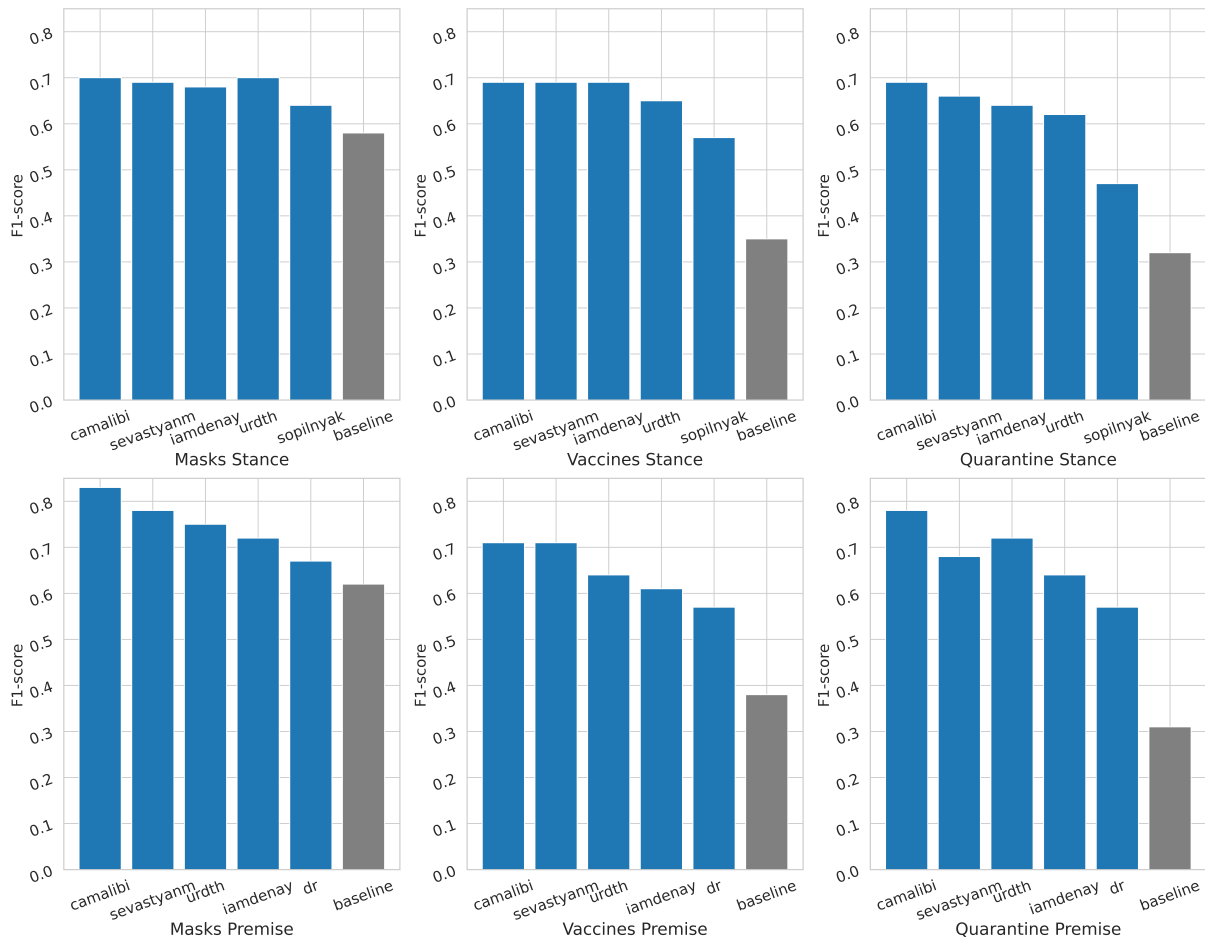


Figure 2: Results for the first top 5 participants for each sub-task.

“Vaccination is beneficial for society”, “The introduction and observance of quarantine is beneficial for society”, and “Wearing masks is beneficial for society”.

Overall, 13 teams participated in the shared task, and more than half of them outperformed the baseline model. The winning system in both sub-tasks used the NLI (Natural Language Inference) variant of the BERT architecture, automatic translation into English to apply a specialized BERT model, pretrained on Twitter posts discussing COVID-19, and additional masking of target entities. This system showed for stance detection F1-score of 0.6968, for premise extraction F1-score of 0.7404 which considerably outperforms the proposed BERT-based baseline (F1-scores of 0.4180 and 0.4355, respectively).

According to the provided results, we see that the argument mining is a feasible task, especially on the COVID-19 dataset. All the data and codes are available online.¹⁴ We hope that these materials will help to foster further research and developments in the area of argument mining for the Russian language.

As future work, we see it promising to explore more complex argument mining setups such as sequence tagging (Chernodub et al., 2019) or information retrieval (Bondarenko et al., 2020).

Acknowledgements

The work of Natalia Loukachevitch in selection of users’ comments and stance annotation is supported by Russian Foundation for Basic Research (project N 20-04-60296). The work of Evgeny Kotelnikov on premise annotation is supported by Russian Science Foundation (project N 22-21-00885¹⁵).

¹⁴<https://github.com/dialogue-evaluation/RuArg>

¹⁵<https://rscf.ru/en/project/22-21-00885/>

References

- Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.
- Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8913–8931, Online, November. Association for Computational Linguistics.
- Nejla Atabey. 2021. Science teachers' argument types and supporting reasons on socioscientific issues: Covid-19 pandemic. *International Journal of Psychology and Educational Studies*, 8(2):214–231.
- Nina Bauwelinck and Els Lefever. 2020. Annotating topics, stance, argumentativeness and claims in Dutch social media comments: A pilot study. // *Proceedings of the 7th Workshop on Argument Mining*, P 8–18, Online, December. Association for Computational Linguistics.
- Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 2021. Investigating label suggestions for opinion mining in german covid-19 social media. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P (to appear), virtual conference, August. Association for Computational Linguistics.
- Alexander Bondarenko, Matthias Hagen, Martin Potthast, Henning Wachsmuth, Meriem Beloucif, Chris Biemann, Alexander Panchenko, and Benno Stein. 2020. Touché: First shared task on argument retrieval. // *Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020)*, P 517—523.
- Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of Touché 2021: Argument Retrieval. // K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021)*, volume 12880 of *Lecture Notes in Computer Science*, P 450–467, Berlin Heidelberg New York, September. Springer.
- Elena Cabrio and Serena Villata. 2020. *Proceedings of the 7th Workshop on Argument Mining*, Online, December. Association for Computational Linguistics.
- Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural argument mining at your fingertips. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, P 195–200, Florence, Italy, July. Association for Computational Linguistics.
- Alexander Chkhartishvili, Dmitry Gubanov, and Ivan Kozitsin. 2021. Covid-19 information consumption and dissemination: A study of online social network vkontakte. // *2021 14th International Conference Management of large-scale system development (MLSD)*, P 1–5. IEEE.
- 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.
- Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. // *Proceedings of the 8th Workshop on Argument Mining*, P 100–111, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Irina Fishcheva and Evgeny Kotelnikov. 2019. Cross-Lingual Argumentation Mining for Russian Texts. // *Proceedings of the 8th International Conference "Analysis of Images, Social networks and Texts" (AIST 2019)*, *Lecture Notes in Computer Science*, P 134–144.
- Irina Fishcheva, Valeriya Goloviznina, and Evgeny Kotelnikov. 2021. Traditional machine learning and deep learning models for argumentation mining in russian texts. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog-2021"*, P 246–258.
- Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. // *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, P 493–503, Hong Kong, China, November. Association for Computational Linguistics.

- Daria Ilina, Irina Kononenko, and Elena Sidorova. 2021. On Developing a Web Resource to Study Argumentation in Popular Science Discourse. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog-2021”*, P 318–327.
- Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. 2015. A shared task on argumentation mining in newspaper editorials. // *Proceedings of the 2nd Workshop on Argumentation Mining*, P 35–38, Denver, CO, June. Association for Computational Linguistics.
- Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Survey*, 53(1):1–37.
- Dilek Küçük and Fazli Can, 2021. *Stance Detection: Concepts, Approaches, Resources, and Outstanding Issues*, P 2673–2676. Association for Computing Machinery, New York, NY, USA.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *ArXiv*, abs/1905.07213.
- Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018. ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. // *Proceedings of the 5th Workshop on Argument Mining*, P 22–28, Brussels, Belgium, November. Association for Computational Linguistics.
- John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.
- 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.
- Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-Shyuan Chiang, Tuan M Lai, Pengfei Yu, Zixuan Zhang, and Heng Ji. 2022. Covid-19 claim radar: A structured claim extraction and tracking system. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) Demo Track*.
- Zhenghao Liu, Chenyan Xiong, Zhuyun Dai, Si Sun, Maosong Sun, and Zhiyuan Liu. 2020. Adapting open domain fact extraction and verification to COVID-FACT through in-domain language modeling. // *Findings of the Association for Computational Linguistics: EMNLP 2020*, P 2395–2400, Online, November. Association for Computational Linguistics.
- Aline Menin, Franck Michel, Fabien Gandon, Raphaël Gazzotti, Elena Cabrio, Olivier Corby, Alain Giboin, Santiago Marro, Tobias Mayer, Serena Villata, and Marco Winckler. 2022. Covid-on-the-Web: Exploring the COVID-19 scientific literature through visualization of linked data from entity and argument mining. *Quantitative Science Studies*, 2(4):1301–1323, 02.
- Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. // *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2016)*, P 31–41.
- 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, San Diego, California, June. Association for Computational Linguistics.
- 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.
- Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. // *Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL ’09*, P 98–107, New York, NY, USA. Association for Computing Machinery.
- Andreas Peldszus and Manfred Stede. 2015. An Annotated Corpus of Argumentative Microtexts. // *Proceedings of the 1st European Conference on Argumentation*, P 801–815.
- Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. 2021. Multi-task and multi-corpora training strategies to enhance argumentative sentence linking performance. // *Proceedings of the 8th Workshop on Argument Mining*, P 12–23, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with background knowledge. *arXiv preprint arXiv:2112.08544*.

- Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. 2018. Cross-lingual argumentative relation identification: from English to Portuguese. // *Proceedings of the 5th Workshop on Argument Mining*, P 144–154, Brussels, Belgium, November. Association for Computational Linguistics.
- Andrew Rosenberg. 2012. Classifying skewed data: Importance weighting to optimize average recall. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 3:2239–2242, 01.
- Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. // *ACL/IJCNLP*.
- Natalia Salomatina, Irina Kononenko, Elena Sidorova, and Ivan Pimenov. 2021. Identification of connected arguments based on reasoning schemes “from expert opinion”. *Journal of Physics: Conference Series*, 1715.
- Robin Schaefer and Manfred Stede. 2021. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.
- Matthias Schildwächter, Alexander Bondarenko, Julian Zenker, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Answering comparative questions: Better than ten-blue-links? // *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19*, P 361–365, New York, NY, USA. Association for Computing Machinery.
- Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. // *Proceedings of the 5th Workshop on Argument Mining*, P 155–163.
- Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. // *Proceedings of the International Conference on Computational Linguistics*, P 1501–1510.
- Manfred Stede and Schneider Jodi. 2018. *Argumentation Mining*. Morgan & Claypool.
- Manfred Stede. 2020. Automatic argumentation mining and the role of stance and sentiment. *Journal of Argumentation in Context*, 9(1):19–41.
- Nhat Tran and Diane Litman. 2021. Multi-task learning in argument mining for persuasive online discussions. // *Proceedings of the 8th Workshop on Argument Mining*, P 148–153, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, P 1338–1352.
- Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace. 2020. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.
- Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.
- Amelie Wüthrl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. // *Proceedings of the 20th Workshop on Biomedical Language Processing*, P 131–142, Online, June. Association for Computational Linguistics.

A All results

From the detailed results for each label we can see that the top 1 result is not always the best approach: many models from the top list perform equally well or even outperform the winner in many cases. For instance, from Tables 5, 7 and 8 we can see that **sevastyannm** and **urdth** are very competitive approaches. The top 1 ranking is achieved by demonstrating stable results across different subsets and good (normally best) precision scores. Interestingly, the best recall scores on quarantine subsets and the vaccines premise subset is achieved by the baseline (more than 0.95 points).

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	1.00	1.00	1.00	0.68	0.66	0.67	0.80	0.83	0.81	0.65	0.61	0.63	0.71	0.70	0.70
sevastyannm	1.00	1.00	1.00	0.65	0.63	0.64	0.80	0.82	0.81	0.62	0.62	0.62	0.69	0.69	0.69
iamdenay	1.00	0.99	1.00	0.67	0.65	0.66	0.77	0.81	0.79	0.61	0.56	0.58	0.68	0.67	0.68
urdth	1.00	1.00	1.00	0.62	0.65	0.63	0.82	0.80	0.81	0.64	0.65	0.65	0.69	0.70	0.70
sopilnyak	0.98	1.00	0.99	0.64	0.50	0.56	0.77	0.86	0.81	0.63	0.50	0.56	0.68	0.62	0.64
baseline	0.99	0.99	0.99	0.46	0.49	0.47	0.77	0.77	0.77	0.50	0.48	0.49	0.58	0.58	0.58

Table 5: Results for masks stance for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	1.00	1.00	1.00	0.78	0.73	0.76	0.94	0.94	0.94	0.73	0.84	0.78	0.82	0.84	0.83
sevastyannm	1.00	1.00	1.00	0.85	0.61	0.71	0.93	0.96	0.94	0.65	0.70	0.67	0.81	0.76	0.78
urdth	1.00	1.00	1.00	0.67	0.73	0.70	0.93	0.93	0.93	0.68	0.60	0.64	0.76	0.75	0.75
iamdenay	1.00	1.00	1.00	0.85	0.47	0.61	0.89	0.96	0.93	0.64	0.62	0.63	0.79	0.68	0.72
dr	1.00	0.99	1.00	0.77	0.47	0.58	0.88	0.96	0.92	0.59	0.46	0.52	0.75	0.63	0.67
baseline	1.00	0.99	0.99	0.56	0.41	0.48	0.89	0.92	0.91	0.44	0.51	0.47	0.63	0.61	0.62

Table 6: Results for masks premise for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	1.00	1.00	1.00	0.78	0.58	0.67	0.72	0.86	0.79	0.71	0.56	0.63	0.74	0.67	0.69
sevastyannm	1.00	1.00	1.00	0.71	0.74	0.73	0.75	0.75	0.75	0.63	0.59	0.61	0.70	0.69	0.69
iamdenay	1.00	1.00	1.00	0.79	0.60	0.69	0.71	0.86	0.78	0.70	0.52	0.60	0.73	0.66	0.69
urdth	1.00	1.00	1.00	0.67	0.57	0.61	0.70	0.82	0.76	0.68	0.52	0.59	0.68	0.64	0.65
sopilnyak	0.99	1.00	1.00	0.61	0.43	0.51	0.67	0.79	0.72	0.54	0.41	0.47	0.61	0.55	0.57
baseline	0.99	1.00	0.99	0.43	0.15	0.22	0.56	0.85	0.67	0.38	0.11	0.17	0.46	0.37	0.35

Table 7: Results for vaccines stance for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	1.00	1.00	1.00	0.79	0.57	0.67	0.89	0.94	0.92	0.55	0.52	0.54	0.75	0.68	0.71
sevastyanm	1.00	1.00	1.00	0.63	0.59	0.61	0.90	0.92	0.91	0.67	0.57	0.62	0.73	0.69	0.71
urdth	1.00	1.00	1.00	0.56	0.59	0.58	0.89	0.88	0.88	0.45	0.48	0.47	0.64	0.65	0.64
iamdenay	1.00	1.00	1.00	0.60	0.50	0.55	0.86	0.85	0.86	0.34	0.52	0.42	0.60	0.63	0.61
dr	1.00	1.00	1.00	0.55	0.39	0.46	0.85	0.91	0.88	0.39	0.33	0.36	0.60	0.54	0.57
baseline	1.00	0.99	0.99	0.43	0.11	0.18	0.80	0.95	0.87	0.33	0.05	0.08	0.52	0.37	0.38

Table 8: Results for vaccines premise for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	0.99	1.00	1.00	0.88	0.35	0.50	0.84	0.83	0.83	0.70	0.81	0.75	0.80	0.66	0.69
sevastyanm	0.99	1.00	1.00	0.57	0.40	0.47	0.85	0.78	0.82	0.62	0.79	0.70	0.68	0.66	0.66
iamdenay	0.99	1.00	0.99	0.67	0.25	0.36	0.84	0.85	0.84	0.70	0.75	0.72	0.73	0.62	0.64
urdth	0.99	1.00	1.00	0.56	0.35	0.43	0.84	0.71	0.77	0.56	0.82	0.67	0.65	0.63	0.62
sopilnyak	0.98	1.00	0.99	0.00	0.00	0.00	0.77	0.80	0.78	0.59	0.67	0.63	0.45	0.49	0.47
baseline	1.00	0.99	1.00	0.00	0.00	0.00	0.65	0.97	0.77	0.62	0.11	0.19	0.42	0.36	0.32

Table 9: Results for quarantine stance for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	0.99	1.00	1.00	0.60	0.27	0.37	0.91	0.96	0.93	0.85	0.68	0.76	0.84	0.73	0.77
sevastyanm	0.99	0.99	0.99	0.44	0.50	0.47	0.91	0.92	0.92	0.72	0.62	0.67	0.69	0.68	0.68
urdth	0.99	1.00	1.00	0.47	0.41	0.44	0.95	0.92	0.93	0.73	0.88	0.80	0.72	0.74	0.72
iamdenay	0.99	1.00	1.00	0.33	0.41	0.37	0.91	0.90	0.91	0.70	0.60	0.65	0.65	0.64	0.64
dr	0.99	1.00	1.00	0.42	0.23	0.29	0.88	0.94	0.91	0.66	0.42	0.51	0.65	0.53	0.57
baseline	1.00	0.99	1.00	0.00	0.00	0.00	0.82	0.99	0.90	0.50	0.02	0.04	0.44	0.34	0.31

Table 10: Results for quarantine premise for top-5 participants.

Grammatical variation: methods for quantitative corpus analysis (the German preposition *wegen*)

E.B. Krotova

Institute of Linguistics, RAS
Moscow, Russia
elena_krotova@inbox.ru

E.N. Tsvetaeva

Moscow State Linguistic University
Moscow, Russia
elenatsvetaeva@yandex.ru

A.V. Sharandin

Russian State University
for the Humanities
Moscow, Russia
sharandin@gmail.com

D.O. Dobrovol'skij

Russian Language Institute, RAS, Institute
of Linguistics, RAS, Moscow, Russia
Stockholm University, Stockholm, Sweden
dobrovol'skij@gmail.com

Abstract

The corpus study of grammatical variation is not considered a complex task anymore. However, the corpus annotations used to solve the task are based on specific theoretical assumptions which do not necessarily cover all possible variants found in empirical material. On the example of the German preposition *wegen* and its variation in case government this article shows why available tools in corpus managers can be insufficient. An experiment is presented in which data obtained from annotated corpora are automatically labeled using a developed algorithm. An analysis of the corpus material showed that the preposition *wegen* primarily governs the genitive (about 60% of all uses). The dative, which is recognized by modern linguistics as a conventional variant, is relatively rare (5%). Additionally, the analysis also revealed that for 35% of nouns following *wegen* case cannot be determined, a phenomenon which is not mentioned in the corresponding dictionary entries of all lexicographic sources known to us. Thus it is not possible to identify the proportion of such underdetermined instances of case government using only the existing annotation of the analyzed corpora, since it did not include tags for nouns with no clear case assignment.

Keywords: Variation, grammatical norm, corpus analysis, German

DOI: 10.28995/2075-7182-2022-21-349-357

Вариативность грамматической нормы: возможности корпусного и количественного анализа (на материале немецкого предлога *wegen*)

Е.Б. Кротова

ИЯЗ РАН
Москва, Россия
elena_krotova@inbox.ru

А.В. Шарандин

РГГУ
Москва, Россия
sharandin@gmail.com

Е.Н. Цветаева

МГЛУ
Москва, Россия
elenatsvetaeva@yandex.ru

Д.О. Добровольский

ИРЯ РАН, ИЯЗ РАН, Москва, Россия
Стокгольмский университет,
Стокгольм, Швеция
dobrovol'skij@gmail.com

Аннотация

Изучение вариативности грамматической нормы с использованием корпусных данных в современных условиях не является сложной задачей, однако разметка корпуса опирается на определенные лингвистические теории и не всегда в должной мере учитывает эмпирический материал.

В статье на примере немецкого предлога *wegen* 'из-за' и вариативности его управления показано, почему имеющиеся средства в корпусных менеджерах оказываются недостаточными; представлен эксперимент, в котором данные, полученные из корпусов, автоматически размечаются с помощью разработанного алгоритма. Анализ корпусного материала показал, что генитив остается базовым управлением предлога *wegen* (около 60% от всех употреблений), а датив, признаваемый современной лингвистикой узаконенным вариантом управления, встречается относительно редко (5%). Анализ выявил, что в 35% случаев у существительного после *wegen* отсутствует указание на падеж, что не упоминается в соответствующих словарных статьях всех известных лексико-графических источников. Выявить долю подобных случаев, пользуясь стандартной разметкой рассмотренных корпусов, не представлялось возможным, так как разметка не предполагала категории беспадежного управления.

Ключевые слова: вариативность, грамматическая норма, корпусный анализ, немецкий язык

1 Введение

Вариативность грамматической нормы свойственна практически всем языкам, один из существующих вариантов может признаваться кодифицированной нормой, а другой – (еще) нет. При этом следует иметь в виду, что норма – понятие динамическое, хотя и более консервативное, чем узус. Смена узуса с течением времени приводит к изменению нормы или к ее модификации в том смысле, что нормативными признаются несколько вариантов. Т.е. понятия универсальной нормы не существует (Hennig and Müller, 2009).

Если говорить о немецкой грамматической норме, то здесь существуют два взгляда на норму: 1) дескриптивная норма, ориентированная на реальное употребление языка, в том числе на институциональных уровнях, и 2) прескриптивная норма, предписанная регулирующими институтами.

Важно, что в Германии отсутствует некая инстанция нормы на государственном уровне¹. Исключение составляет только „Rat der deutschen Rechtschreibung“, занимающийся исключительно вопросами правописания. В сомнительных случаях носитель языка обращается к авторитетным словарям или справочникам, например, Duden online², DWDS³, Дуден-9 и Вариг-5 (Duden 9, 2021; Wahrig, 2003).

По этой причине в сфере грамматической нормы и ее вариантов на сегодняшний день наблюдается ряд трудно решаемых проблем. Так, свойственная многим языкам – и немецкий здесь не исключение – вариативность падежного управления предлогов затрудняет процесс однозначного определения языковой нормы. Отдельные падежи (как например, родительный/генитив и дательный/датив в немецком), управляемые большим количеством предлогов и конкурирующие друг с другом на протяжении всего исторического развития, представляют собой особую трудность. Определение единой нормы как таковой для них вряд ли вообще возможно. Однако вполне реально определить допустимые варианты нормы с учетом количественных данных, сложившихся под воздействием систематических исторических процессов и определяющих не только и не столько диалектальное, сколько литературное варьирование падежного управления.

На сегодняшнем этапе развития информационных технологий решение обсуждаемой проблемы становится возможным благодаря, с одной стороны, наличию больших корпусов, с другой – благодаря наличию инструментов для обработки естественного языка.

Продемонстрировать это можно на примере немецкого предлога *wegen* 'из-за, по причине', нормативно допускающего вариативность управления.

2 Постановка задачи

Среди всех германских языков падежная система немецкого языка представляет собой промежуточный вариант между архаичной исландской и ярко выраженной аналитической ан-

¹Ср.: „Jeder in einer Grammatik kodifizierte Sprachgebrauch kann zur Norm erhoben werden, auch wenn die Autoren reine Deskription reklamieren.“ (Zifonun et al., 1997, p. 6) – Каждая зафиксированная любой из существующих грамматик особенность употребления языка может получить статус нормативной, даже если авторы формулировали свою задачу как чисто описательную.

²<https://www.duden.de>

³<https://www.dwds.de>

глейской, включая в себя, таким образом, две подсистемы – флективную и аналитическую (Dal, 1960). Этот факт не может не влиять на динамические процессы в морфологической системе языка. Редукция падежной системы в немецком – типичная для большинства германских языков – привела, в частности, к изменению семантики существующих падежей, что в свою очередь способствовало вытеснению их морфологических признаков предложными конструкциями.

Предлоги с генитивным управлением, будучи производными, непервообразными, стоят в предложной системе немецкого языка особняком: они возникли – по историческим меркам – относительно недавно, образовавшись преимущественно от основ существительных в ранненововерхненемецкий период (1350-1650), что и объясняет их управление: первичная функция падежа при них представляла собой адноминальный атрибутивный генитив. В течение продолжительного времени падежное управление при новых предлогах не было стабильным: генитив конкурировал с дативом, что не является случайным. В результате редукции падежной системы в ней остаются и развивают свою семантику падежи с наиболее абстрактным значением, их абстрактно-грамматический характер, а также многозначность – результат позднейшего обобщения на основе слияния форм и значений. Для немецкого языка максимально синкретическим оказался дательный падеж (Zhirmunskiy, 2020). Однако уже с начала XX в. генитив вновь занимает господствующее положение, что касается всей группы этих предлогов. Так, *wegen* в литературном языке употребляется в это время, как правило, с генитивом (Admoni, 1986, p. 119).

Глобальные процессы в падежной системе обусловили не только явление падежного синкретизма, но и способствовали развитию тенденции к проявлению аналитических признаков на грамматическом уровне и фразеологизации немаркированных форм в определенных видах дискурса (Zhirmunskiy, 2020; Schmid, 2017; Nübling, 2006). Весьма существенной становится роль предлогов, их количество и значения увеличиваются, а функции расширяются. С этим, в частности, связано и явление беспадежного употребления существительного после предлога, которое не тождественно явлению падежного синкретизма. Сам факт предложного управления с немаркированным падежом – явление не новое и характерное не только для исследуемых конструкций.

Сегодня благодаря доступности больших корпусов этот способ управления стал более заметен и не может игнорироваться грамматическим описанием, несмотря на то что традиционные словари и грамматики не упоминают это явление. Именно благодаря анализу больших объемов данных стала возможной дифференциация случаев морфологического синкретизма и беспадежного употребления существительных в составе предложной группы.

Изучение вариативности управления немецкого предлога *wegen* не представляется нам сегодня неразрешимой задачей: корпусные исследования вариативности немецких предлогов и, конкретно, предлога *wegen* уже проводились, и результаты представлены, например, на сайте проекта Grammis⁴ Института немецкого языка в г. Мангейме (далее – IDS), где обобщена информация, представленная в корпусе Deutsches Referenzkorpus⁵ (далее – DeReKo). В соответствующей словарной статье указано на возможность употребления предлога *wegen* как с генитивом, так и с дативом, а также без маркировки падежа (*ohne Kasusmarkierung*), однако данных о распределении частотности употреблений с разными падежами не приводится.

В исследовании предлогов, допускающих варьирование (Di Meola, 2009), проведенном также на основе DeReKo, были проанализированы вручную до 1000 контекстов на один предлог. Полученные данные автор разделил на следующие категории: 1) однозначное указание на генитив (для предлога *wegen* это 45,3% от общего числа проанализированных контекстов), 2) однозначное указание на датив (3,3% для *wegen*), 3) употребление в форме,

⁴<https://grammis.ids-mannheim.de>

⁵<https://cosmas2.ids-mannheim.de/cosmas2-web/>

которая может интерпретироваться как генитив (Genitiv-kompatibel) (49,3% для *wegen*), и 4) употребления в форме, которая может интерпретироваться как датив (сама форма исключает генитив, но маркеры датива отсутствуют) (2,1% для *wegen*). Этот анализ не выявляет, однако, доли беспадежного употребления и не позволяет однозначно определить, какие случаи попадают в категории (3) и (4). К тому же, анализ проводился вручную, а потому имеется ограничение на число контекстов.

Существуют также инструменты, с помощью которых можно выявить набор существительных, употребляемых с рассматриваемым предлогом. Сбор такой информации на основе текстов, полученных из корпуса, возможен с помощью Lexpan⁶, разрабатываемого также в IDS. Более того, имеется база данных по предложным группам Prepcon⁷, в которой представлен и предлог *wegen* с наиболее частотными управляемыми им существительными.

Тем не менее, вышеприведенные результаты исследований и использованные инструменты не позволяют получить готовый ответ на вопрос о том, каково распределение доли генитивного, дативного и беспадежного управления. Для решения этого вопроса необходимо обращение к крупному корпусу, снабженному морфологической разметкой. На первый взгляд кажется, что задача может быть сведена к нескольким запросам в выбранных для исследования корпусах, в которых исследователь выявляет количество генитивного и дативного управления при рассматриваемом предлоге и наличие иных вариантов. На практике, однако, задача оказывается сложнее.

Материал для исследования управления предлога *wegen* основывался на данных нескольких корпусов, имеющихся в Sketch Engine⁸, DeReKo и DWDS⁹ (в основном анализировался немецкий язык в Германии без учета региональных вариантов). В последних двух корпусных менеджерах отсутствуют возможности поиска по падежам, только по частеречным тэгам. В Sketch Engine доступен поиск и по частеречным тэгам, и по падежам. Однако для поставленной исследовательской задачи возможностей поиска оказалось недостаточно, и разметка также местами спорная. К примеру, если в Sketch Engine в корпусе German Web 2018 запрашивать случаи генитивного управления *wegen* (запрос: [lemma="wegen"] [1,3 [case="Gen"]]), то в результатах поиска среди первых десяти контекстов будут: 1) случаи употребления с генитивом (*wegen seiner guten Englischkenntnisse* 'из-за его хорошего знания английского языка'), 2) случаи, в которых дифференциация генитив vs. датив не представляется возможной, так как у существительного женского рода в единственном числе формы генитива и датива совпадают (*wegen der riesigen Nachfrage* 'из-за огромного спроса'), 3) случаи с дативом (*wegen Mängeln der Leistung* 'из-за плохой успеваемости'), 4) случаи без указания на падеж (далее – беспадежное употребление), т. е. с отсутствующим артиклем и с отсутствующим окончанием у существительного либо же с окончанием существительного без показателей генитива или датива (*wegen Sex und Drogen* 'из-за секса и наркотиков'). В случае (4) разметка *Drogen* как формы генитива спорная, так как формальные характеристики генитива отсутствуют. В случае (3) *der Leistung* стоит в форме, идентичной и генитиву, и дативу, т. е. разметка также спорная, кроме того существительное *Leistung* является зависимым не предлога *wegen*, а существительного *Mängeln* (что соответствует дистанции, указанной в запросе, – до трех токенов между предлогом и словом в генитиве). Схожие результаты можно получить при аналогичном поиске для датива.

Можно было бы предположить, что надо смотреть на первое существительное после предлога. Однако и такой подход ведет к ошибкам. После предлога может идти расширенный причастный оборот (*wegen vom Regen aufgeweichter Straßen* 'из-за мокрых от дождя дорог'). Предлог *wegen* здесь управляет генитивом *aufgeweichter Straßen*, хотя первое существительное после него стоит в дативе, но управляется не предлогом *wegen*, а предлогом *von*: *vom Regen*. Более того, предлог *wegen* может встречаться также в постпозиции (*der tiefgrünen*,

⁶<http://uwv.ids-mannheim.de/lexpan>

⁷http://uwv.ids-mannheim.de/prepcon/prepcon_online.html

⁸<https://www.sketchengine.eu>

⁹<https://www.dwds.de/d/korpora/spk>

satten Farbe wegen ‘из-за насыщенного зеленого цвета’). Случаи употребления *wegen* в постпозиции относительно редки, и поиск существительного в генитиве или дативе, после которого идет *wegen*, дал бы слишком много нерелевантных контекстов. Чтобы учесть такие случаи, необходим поиск по синтаксическим поддеревьям, недоступный в Sketch Engine.

Таким образом, основными препятствиями для работы с корпусом в Sketch Engine стали невозможность поиска по синтаксическим поддеревьям, где вершиной являлся бы предлог *wegen*, и отсутствие в используемой системе тэгов обозначений для случаев, где нельзя точно сказать, идет ли речь о генитиве или дативе (например, обозначение типа Gen/Dat), и для случаев беспадежного употребления (так мы называем случаи, в которых нет никакого указания на падеж: артикля или окончания у существительного). Поэтому было решено использовать имеющиеся корпуса исключительно как источник материала и обрабатывать полученные контексты с помощью скриптов на языке программирования Python и с помощью библиотеки spaCy¹⁰.

3 Материал и метод

Были проанализированы данные следующих корпусов:

1. Корпус немецких газет (размер корпуса – 7 млрд. токенов), DeReKo;
2. Корпус устной речи *Gesprochene Sprache* (размер – 3 млн. токенов), DWDS;
3. Корпус интернет-текстов *German Web 2018* (размер – 5,3 млрд. токенов), Sketch Engine.

Первый корпус – подкорпус из архива корпусов *W – Archiv der geschriebenen Sprache*, DeReKo, составлен на основе имеющихся корпусов письменной речи и полностью состоит из публицистики. Для чистоты эксперимента были исключены австрийские и швейцарские источники и оставлены только немецкие. В виртуальный корпус вошли немецкие газеты с 1990 г. по 2019 г. Размер корпуса при этом составил 7 млрд. токенов.

Также были проанализированы данные корпуса DWDS устной речи (*Gesprochene Sprache*). Корпус охватывает года с 1900 г. по 2001 г. В него входят транскрипты официальных речей, протоколы парламентских заседаний и интервью. Размер корпуса небольшой (менее 3 млн. токенов).

В качестве третьего корпуса выступает корпус *German Web 2018*, представленный в Sketch Engine. Он достаточно крупный (5,3 млрд. токенов) и состоит из текстов, собранных в интернете с разных сайтов.

Здесь следует сделать небольшую оговорку относительно состава последних двух корпусов. Хотя корпус DWDS *Gesprochene Sprache* называется корпусом разговорной речи, это несколько условное обозначение. В нем в основном представлены официальные речи, то есть не спонтанная речь, а подготовленные и, скорее всего, прошедшие редактуру тексты. В то же время корпус *Sketch Engine German Web 2018*, являясь корпусом письменной речи, может отражать и некоторые явления разговорной речи, так как в него входят, в том числе, форумы и блоги, хотя назвать его корпусом разговорной речи тоже нельзя. Поэтому выводы, полученные в результате данного корпусного исследования, будут актуальны прежде всего для письменной речи. Возможно, на корпусе спонтанной устной речи данные будут отличаться.

В выше указанных корпусах объявлялся поиск на случаи употребления словоформы *wegen*. Никаких дополнительных ограничений в запросе не было. Из DeReKo и Sketch Engine было скачано из каждого корпуса до 10 тыс. контекстов (максимально допустимое количество для скачивания), из корпуса DWDS – 465 контекстов.

Все контексты делились на четыре категории:

1. генитив: в эту группу попадали все те случаи, где есть четкий показатель генитива (артикль, окончание прилагательного или окончание у существительного);
2. датив: случаи с четким показателем датива (артикль, окончание прилагательного или окончание у существительного);

¹⁰<https://spacy.io>

3. беспадежное употребление (далее – null): нет артикля, у существительного нет окончания, которое бы указывало на генитив или датив, прилагательное отсутствует;
4. генитив или датив (далее – Gen/Dat): существительные женского рода в единственном числе, у которых эти две формы совпадают.

Последняя группа в анализе игнорируется, так как не имеет значения для данного исследования, а первые три (генитив, датив и беспадежное употребление) составляют 100%. В группу с беспадежным употреблением входят как употребления в единственном (*wegen Körperverletzung* ‘из-за телесных повреждений’), так и во множественном числе (*wegen Kreislaufschwierigkeiten* ‘из-за проблем с кровообращением’). Они объединены в одну группу именно по формальным признакам, то есть по отсутствию указания на падеж. Надо отметить, что в размеченных данных есть возможность отдельно рассматривать случаи употребления в ед. ч. и во мн. ч.

Чтобы получить статистику по управлению предлога *wegen*, которой можно было бы доверять, понадобилось сделать скрипт, который анализирует полученный из корпуса материал¹¹. Этот скрипт написан на языке программирования Python с использованием библиотеки *sraCu*, в которой есть языковые модели для немецкого языка. С помощью нее можно в том числе синтаксически проанализировать текст и получить морфологическую информацию по элементам текста.

В процессе обработки корпусного материала собирается морфологическая информация о каждом элементе синтаксического поддерева, вершиной которого является предлог *wegen*, и определяется, какое существительное зависит непосредственно от него. Для нахождения этого существительного можно пользоваться синтаксическим разбором *sraCu*, но в ряде случаев он ошибочен. Особенно частотны ошибки при постпозиции предлога и при перечислениях, поэтому для определения зависимого существительного мы пользовались в дополнение к разбору *sraCu* собственным алгоритмом.

Особое внимание уделяется случаям с перечислениями, так как в них могут встречаться разные падежи. Например, во фразе *wegen gemeinschaftlicher schwerer räuberischer Erpressung, Freiheitsberaubung, Verstößen* ‘из-за коллективного вымогательства с отягчающими обстоятельствами, лишения свободы, нарушений’ присутствует Gen/Dat (*wegen gemeinschaftlicher schwerer räuberischer Erpressung*), который игнорируется в подсчетах, беспадежное употребление (*Freiheitsberaubung*) и датив (*Verstößen*). Разные падежи или беспадежное употребление внутри одного перечисления подсчитываются отдельно¹². *sraCu* также автоматически определяет падеж, но он не всегда верный, поэтому указанный *sraCu* падеж мы не учитывали.

Отдельно анализируются конструкции с постпозицией (*von +*) Gen + *wegen*. Они редко встречаются в корпусе, и у них ограниченное число заполнителей слота. Чаще всего это фраза *von Amts wegen* ‘по долгу службы’.

Сложность представляют также распространенные причастные обороты. Если между предлогом *wegen* и первым существительным после него есть еще один предлог, то существительным, зависящим от *wegen*, скорее всего будет являться последнее существительное в выделенном синтаксическом поддереве (например, *wegen der in seiner Heimat herrschenden Kriegswirren* ‘из-за царящих на его родине военных беспорядков’).

Когда найдено существительное, непосредственно зависящее от предлога *wegen*, определяется искомый падеж. Для этого, помимо информации о роде, числе существительного и его окончании, требуется информация о том, имеется ли у него артикль, и если да, то в какой форме, и стоит ли перед ним прилагательное, если да, то с каким окончанием. После

¹¹ Данные и скрипты представлены по ссылке: https://github.com/ln-kr/german_prepositions

¹² Встретилось менее 1% случаев, когда внутри группы однородных членов встречаются генитив и датив. Гораздо чаще некоторые из однородных существительных стоят в генитиве или дативе, а другие – с немаркированным падежом, ср: *wegen Förderung der Volks- und Berufsbildung sowie der Studentenhilfe und der Förderung des Umweltschutzes* ‘из-за продвижения национального и профессионального образования, а также помощи студентам и поощрения защиты окружающей среды’. Подобные контексты нашлись в корпусах письменной речи DeReKo и German Web.

DeReKo, корпус немецких газет с 1990 г.	%
генитив	60%
датель	4%
null	36%
DWDS, корпус <i>Gesprochene Sprache</i>	%
генитив	66%
датель	6%
null	28%
Sketch Engine, корпус <i>German Web 2018</i>	%
генитив	51%
датель	12%
null	37%

Таблица 1: Данные по употреблению в генитиве, датель и с беспадёжным управлением, %

учета всех этих параметров определяется падеж – генитив, датель, Gen/Dat или отмечается случай беспадёжного употребления.

Как можно видеть, определение падежа, которым управляет предлог, допускающий варьирование, является часто нетривиальной задачей. В выше разобранном случае с *wegen* задачу усложняет возможность нахождения предлога в постпозиции, наличие перечислений, внутри которых могут встречаться существительные в разных падежах или в беспадёжном употреблении, расширенных причастных оборотов и наличие предложных групп, зависящих не от предлога, а от других элементов поддерева.

4 Результаты и дискуссия

Подсчет проанализированных контекстов дал следующие результаты (см. табл. 1).

Таким образом, генитив встречается примерно в 60% случаев, беспадёжное употребление – в 35%, датель – в 5%¹³.

В целом можно сказать, что датель в проанализированных корпусах представлен слабо, минимум 4% в публицистике, максимум 12% в корпусе интернет-текстов. Гораздо более значимым с точки зрения вариативности нормы оказывается беспадёжное употребление. Оно значительно чаще встречается, чем употребление в датель, в трети случаев, о чем не говорится в учебниках и словарях.

Помимо прочего, полученные данные позволяют проанализировать контекстные условия, в которых встречается рассматриваемый предлог. Рассмотрим ниже употребления с датель и беспадёжное употребление.

По данным корпусов из Sketch Engine и DeReKo существительные, встретившиеся в датель больше одного раза, в основном имеют форму мн.ч. (например, *wegen Mängeln* ‘из-за недостатков’, *wegen Verstößen* ‘из-за нарушений’) и означают либо какие-либо недостатки и проблемы (*wegen Problemen* ‘из-за проблем’, *wegen Vorwürfen* ‘из-за упреков’), либо какие-либо нарушения и преступления (*wegen Drogendelikten* ‘из-за преступлений в сфере оборота наркотиков’, *wegen Eigentumsdelikten* ‘из-за имущественных преступлений’). Соответственно, существительные в датель после *wegen* встречаются в основном в текстах юридической направленности.

Наиболее частотные случаи беспадёжного управления – это существительные женского

¹³Личные местоимения после *wegen* употребляются практически всегда в датель в современном узусе, ср. *wegen mir* ‘из-за меня’, *wegen dir* ‘из-за тебя’ и пр. (адвербиальные формы с исходно генитивным элементом типа *meinetwegen* имеют другое значение).

рода, обозначающие либо нарушения и преступления (*wegen Beihilfe* ‘из-за пособничества’, *wegen Beleidigung* ‘из-за оскорбления’, *wegen Sachbeschädigung* ‘из-за повреждения имущества’), либо травмы (*wegen Verletzung* ‘из-за травмы, повреждения’). В основном существительные в беспадежном употреблении встречаются в текстах юридической направленности и в спортивных новостях (например, сообщения о том, что из-за травмы какой-либо игрок не участвует в соревнованиях).

Возможно, в корпусе устной спонтанной речи доля датива среди существительных оказалась бы выше. Об этом свидетельствуют и результаты других исследований (Babenko, 2016; Marossek, 2016).

Основные лексикографические источники современного немецкого языка Duden и DWDS, доступные также в Сети, сообщают, что предлог *wegen* может управлять как генитивом, так и дативом, однако акценты касательно модели управления этого предлога расставлены в этих словарях по-разному. Duden указывает, что оба падежа – генитив и датив – представляют собой абсолютно равноправные употребления, не нарушающие грамматические нормы немецкого языка, причем употребление датива с *wegen* в письменной речи является стандартным разговорным вариантом. Duden, не вдаваясь в подробности, замечает на полях, что после *wegen* зависимое существительное может также употребляться в несклоняемой форме.

DWDS выделяет только два режима предложного управления для предлога *wegen*: основной с генитивом и альтернативный с дативом. Про беспадежное управление *wegen* пользователь этого корпусного словаря не получает вообще никакой информации. Дативные употребления DWDS снабжает дополнительно пометами «разговорный», «южно-немецкий», «австрийский». Представляется, что подобная характеристика употребления *wegen* несколько устарела, поскольку сегодня дативное управление *wegen* более не имеет территориальных ограничений и достаточно широко представлено в практике носителей немецкого языка в немецкоязычных странах.

Как видно из вышеупомянутых описаний, картина языковой нормы в отношении употреблений предлога *wegen* остается несколько неясной, особенно для неносителя немецкого языка. Основной отечественный двуязычный словарь под руководством О.И. Москальской (как и DWDS) сообщает пользователю, что *wegen* управляет генитивом, а датив представлен лишь как разговорный вариант в южных регионах Германии. Корпусные данные это опровергают.

Приведем фрагмент словарной статьи предлога *wegen* из «Немецко-русского словаря актуальной лексики», находящегося в данный момент в печати. Этот словарь, как представляется, учитывает современное словоупотребление и с помощью различных метаязыковых комментариев помогает пользователю разобраться в особенностях употребления предлога *wegen*. Сильной стороной этого словаря является также наличие основанных на корпусных данных иллюстративных примеров, что особенно важно для лексикографического представления служебной лексики.

wégen *ppr* (*G*, в некоторых формах и разг. *D*) *тж.* в *постпозиции* *ради*; *из-за*; *вследствие*; *благодаря*; ~ *éines Mótorschadens könnte die Fahrt nicht fórtgesetzt wéren* *из-за* [вследствие] поломки двигателя поезде пришлось прервать; *séines Gejámmer* ~ [séinem Gejámmer ~] *máchten wir éine Páuse* *из-за* его нытья мы сделали перерыв; с личными местоимениями *тж.* *D*: ~ *dir* *из-за* тебя; в *препозиции* возможно беспадежное управление – в случае отсутствия при суц. артикля или определяющих слов: ~ *Kránkheit* по болезни; ~ *Betrúg* по причине обмана

5 Заключение

Получение достоверных статистических данных по вариативному управлению предлогов с помощью имеющихся для немецкого языка корпусных менеджеров оказывается затруднительным по ряду причин. В DeReKo и DWDS нет разметки падежей, а в Sketch Engine

эта разметка не всегда корректна. Имеющийся набор тэгов для падежей не предполагает случаев, в которых форма слова может одновременно соответствовать двум падежам (дативу и генитиву) или где вообще нет никакого указания на падеж. Синтаксический анализ полученного из корпусов текстового материала с помощью иных средств (библиотеки *sraСу*) помогает приблизиться к цели, однако сам по себе не является достаточным. По этой причине для получения информации об управлении предлога необходимо дополнить набор *sraСу* собственными алгоритмами, в которых будут учтены все детали (возможность нахождения предлога в постпозиции, особая обработка перечислений и расширенных причастных оборотов) и соблюдена выбранная классификация (генитив, датив, беспадёжное употребление).

На примере предлога *wegen* мы показали возможности современных эмпирических методов исследования, которые позволяют внести ясность в проблематику вариативности грамматической нормы, актуальной для построения корпусной грамматики разных языков.

Новизна полученных результатов состоит в том, что впервые при классификации употреблений предлога *wegen* в особую группу выделяются случаи беспадёжного употребления существительных и именных групп, управляемых этим предлогом. Полученные статистические данные по различным типам управления являются ценным лингвистическим результатом, который используется в лексикографии и преподавании немецкого языка. Теоретический вывод из данного исследования состоит в том, что полученные результаты отражают общие типологические тенденции в грамматике современного немецкого языка.

Acknowledgements

Авторы благодарят анонимных рецензентов за внимательное прочтение работы, конструктивную критику и замечания, учтенные в окончательной версии статьи.

References

- Vladimir G. Admoni. 1986. *Teoreticheskaya grammatika nemetskogo yazyka (na nemetskom yazyke)* [Theoretical grammar of the German language (in German)]. Prosveshcheniye, Moskva, 4 edition.
- Natalia S. Babenko. 2016. Yeshche raz o kategorii padezha: 'obshchiy padezh' v sovremennom nemetskom yazyke [Once again about the category of case: 'common case' in modern German]. *Voprosy psikholingvistiki* [Questions of psycholinguistics], 28:38–43.
- Ingerid Dal. 1960. Entwicklungstendenzen im germanischen Kasussystem. *Untersuchungen zur germanischen und deutschen Sprachgeschichte*, 2:158–170.
- Claudio Di Meola. 2009. Rektionsschwankungen bei Präpositionen – erlaubt, verboten, unbeachtet. *Deutsche Grammatik - Regeln, Normen, Sprachgebrauch*, P 195–221.
- Duden 9. 2021. *Sprachliche Zweifelsfälle: das Wörterbuch für richtiges und gutes Deutsch*. Dudenverlag, 9. überarbeitete und erweiterte Auflage edition.
- Mathilde Hennig and Christoph Müller. 2009. *Wie normal ist die Norm? Sprachliche Normen im Spannungsfeld von Sprachwissenschaft, Sprachöffentlichkeit und Sprachdidaktik*. Kassel University Press, Kassel.
- Diana Marossek. 2016. *Kommst du Bahnhof oder hast du Auto? Warum wir reden, wie wir neuerdings reden?* Hanser Berlin im Carl Hanser Verlag, München.
- Doris Nübling. 2006. *Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels*. Gunter Narr Verlag, Tübingen.
- Hans Ulrich Schmid. 2017. *Einführung in die deutsche Sprachgeschichte*. J. B. Metzler Verlag, Stuttgart.
- Wahrig. 2003. *Fehlerfreies und gutes Deutsch. Das zuverlässige Nachschlagewerk zur Klärung sprachlicher Zweifelsfälle*. Bertelsmann, Gütersloh / München.
- Viktor M. Zhirmunskiy. 2020. *Istoriya nemetskogo yazyka* [History of the German language]. Stereotypical edition URSS.
- Gisela Zifonun, Ludger Hoffmann, et al. 1997. *Grammatik der deutschen Sprache (GDS)*. De Gruyter, Berlin / New York.

Сентенциальные актанты ментальных предикатов с союзом *когда* (по данным Национального корпуса русского языка)

Кустова Г. И.

Институт русского языка
им. В. В. Виноградова РАН;
Московский педагогический
государственный университет,
Москва, Россия
galinak03@gmail.com

Аннотация

Известно, что матричные предикаты со значением эмоции и оценки присоединяют сентенциальные актанты не только с союзом *что* (*Обидно / плохо, что команда проиграла*), но и с союзами *когда* и *если*: *Плохо, когда / если команда проигрывает*. Данные НКРЯ показывают, что предикаты других семантических классов, которые не упоминаются в грамматиках, тоже могут присоединять клаузы с союзами *когда* и *если*. В статье обсуждаются примеры *когда*-предложений с ментальными предикатами (*знать, помнить, понимать / понятно*): *Я помню, когда по Бородинскому мосту ходили трамваи; Понятно, когда клетки формируются в ходе развития зародыша, но во взрослом организме?*

Ключевые слова: сложное предложение; ментальные матричные предикаты; сентенциальный актант; союз

DOI: 10.28995/2075-7182-2022-21-358-366

Sentential arguments of mental predicates with conjunction *kogda* 'when' (according to the Russian National Corpus)

Kustova G. I.

Vinogradov Russian Language Institute
of the Russian Academy of Sciences;
Moscow Pedagogical State University,
Moscow, Russia
galinak03@gmail.com

Abstract

It is known that the matrix predicates of emotion and evaluation subordinate sentential arguments not only with conjunction *что* 'that': *Obidno / plokho, chto komanda proigrala* 'It's a shame / Too bad the team lost', but also with conjunctions *kogda / esli* 'when' / 'if': *Plokho, kogda / esli komanda proigravaet* 'It's bad if the team loses'. The NRC data show that predicates of the other semantic classes that are not mentioned in grammars can also subordinate clauses with conjunctions *kogda* 'when' and *esli* 'if'.

The article discusses examples of *kogda*-sentences with mental predicates (*znat'* 'to know', *pomnit'* 'to remember', *ponimat'* / *poniatno* 'to understand'): *Ya pomnyu, kogda po Borodinskomu mostu khodili tramvai* 'I remember when trams ran along the Borodino bridge'; *Ponyatno, kogda kletki formiruyutsya v khode razvitiya zarodysha, no vo vzrosлом organizme?* 'It is clear when cells are formed during the development of the embryo, but in an adult organism?'

Keywords: complex sentence; mental matrix predicates; sentential argument; conjunction

1 Гибридные предложения с *когда/если* придаточными

Сентенциальные актанты (актантные придаточные, в традиционной терминологии – изъяснительные придаточные), выражающие пропозициональную валентность (валентность

содержания), в стандартном случае присоединяются к матричному предикату союзами *что* (*Знаю, что они вошли*), *чтобы* (*Хочу, чтобы они вошли*), *как* (*Видел, как они вошли*), *будто* (*Показалось, будто кто-то вошел*) и нек. др., а также относительными местоимениями (косвенный вопрос и косвенное восклицание). Кроме того, есть нестандартный способ присоединения сентенциального актанта – союзами *когда* и *если*:

(1) **Жаль, когда** картины, мысленно предназначенные для Родины, уходят в музеи на чужбине [Н. К. Рерих. Листы дневника (1946)] ≈ 'Жаль, что картины уходят на чужбину'

(2) **Ведь богам тоже обидно, если кто-то умнее их** [Михаил Успенский. Там, где нас нет (1995)] ≈ 'Богам обидно, что кто-то умнее их'

По поводу этих союзов в литературе существует консенсус: считается, что они, хотя и присоединяют актантные придаточные, частично сохраняют исходные значения времени и условия, ср. [RG-80: §2793]; [Letuchii 2021: 210].

По поводу матричных предикатов, присоединяющих *когда/если*-придаточные, тоже существует консенсус: это предикаты эмоции (*приятно, обидно, любить, ненавидеть* и под.) и оценки (*хорошо, плохо, странно* и под.), ср. [Letuchii 2021], [Maksimov 2011], [Serdobol'skaya 2011], [Skoblikova 2006]. В [RG-80: §2793] они названы предикатами «психических реакций», однако речь идет о той же лексике.

В работе [Letuchii 2021] наряду с «предикатами эмоции и оценки» фигурирует также термин «фактивные предикаты». Н.В. Сердобольская [Serdobol'skaya 2011: 432] упоминает, что кроме эмоций и оценок возможны и другие классы матричных предикатов. Однако в приводимых примерах фигурируют только предикаты эмоции и оценки. Какие еще семантические классы матричных предикатов могут присоединять нестандартные *когда/если*-придаточные, в литературе не говорится, и тем более не приводится лексика, обладающая таким свойством. В данной работе мы ставим задачу найти другие группы матричных предикатов, присоединяющих *когда/если*-придаточные, и подтвердить их примерами из Национального корпуса русского языка (НКРЯ).

Предикаты эмоции и оценки мы будем называть основными группами матричных предикатов для *когда/если*-придаточных, остальные группы – периферийными.

На сегодняшний день мы обнаружили две периферийные группы матричных предикатов – перцептивные: *видеть / видно, слышать / слышно, заметно* – и ментальные. Сразу оговоримся, что в группе ментальных предикатов традиционно выделяются два больших класса – мнение (нефактивные) и знание (фактивные), ср. [Apresjan 1995]. Предикаты мнения (*думать, считать, предполагать* и под.) не присоединяют актантных *когда*-придаточных, поэтому далее речь пойдет только о ментальных предикатах знания – *знать, помнить, понимать / понятно*. К этой группе примыкают также *видеть / видно, заметно, чувствовать* в ментальном значении 'знать, понимать, догадываться' (ср.: *Я же вижу, что ты врешь; Видно, что он врет; Заметно, что он волнуется / что он чего-то боится; Я чувствую, что он что-то скрывает*), но их свойства не отличаются от исходных предикатов знания, поэтому из соображений экономии мы ограничиваемся предикатами *знать, помнить, понимать / понятно*.

Еще одно уточнение. Речь будет идти только о *когда*-придаточных. *Если*-придаточные даже при основных матричных предикатах эмоции и оценки встречаются достаточно редко. Что касается предложений с предикатами периферийных групп, то они и сами по себе немногочисленны, а *если*-придаточные вообще единичны. В НКРЯ удалось обнаружить лишь несколько примеров (ср.: *Ему и стыдно, но он наглый, гордый. Видно, всё ещё прячется в лесах, иначе как объяснить, что взрослый, сильный мужчина украл козла. Понятно, если бы он угнал лошадь, быка. Это лихость. А тогда козла украсть — всё равно что сейчас курицу украсть* [Фазиль Искандер. Чик чтит обычаи (1967)]), и еще несколько примеров было найдено в коллекции <https://books.google.ru>.

Общая схема декомпозиции предложений с матричными предикатами основной группы такова: Когда P, V / Praed, что P:

(3) **Неприятно, когда** тебя обманывают

'Когда тебя обманывают, неприятно, что тебя обманывают'

Аналогичная схема декомпозиции у предложений с ментальными предикатами (в приводимом примере *заметно* употребляется в ментальном значении):

(4) *Всегда заметно, когда человек чего-то боится*

‘Когда человек чего-то боится, всегда заметно, что он чего-то боится’

В работе [Letuchii 2021: 214] для предложений с *когда*-придаточными предлагается несколько иная схема: *Меня обижает, когда со мной так говорят* – ‘В каждом случае, когда со мной так говорят, меня это обижает’ (компонент ‘когда Р’ назван аспектуальной рамкой). Однако для нас важна идея, что предложения с актантными *когда*-придаточными можно рассматривать как контаминацию, компрессию двух типов придаточных. Исходя из этих соображений, далее мы будем для краткости называть предложения с актантными *когда*-придаточными гибридными предложениями. Кроме того, нам важно показать, с какими союзами в «нормальных» предложениях соотносится *когда* в гибридных предложениях. В приведенной схеме декомпозиции *когда* коррелирует со *что*, однако, как будет показано далее, он может также коррелировать с союзом *как* и даже с некоторыми относительно-вопросительными местоимениями (союзными словами). При этом не имеется в виду местоимение *когда* – косвенный вопрос с *когда* в некоторых случаях может быть второй интерпретацией *когда*-предложений, но нас этот случай не интересует. Речь идет о местоимении *почему* (см. ниже).

2 Матричный предикат *знать*

Предложениям с предикатом *знать* мы посвятили отдельную работу, так что здесь лишь коротко опишем их свойства. Речь идет о предложениях вида:

(5) *В принципе, я всегда знаю, когда мне врут или выдумывают. Мне даже не нужно на человека смотреть... Я не могу объяснить это чувство... это просто ощущения – вот здесь ложь... это интуиция* [Интернет-форум]

Предложения с союзом *когда* следует отличать от косвенных вопросов с наречием *когда*: *Я знаю, когда люди врут* = ‘Я знаю, в каких ситуациях, в каких обстоятельствах люди врут’.

Предложения с союзами имеют обычную для гибридных предложений с *когда*-придаточными интерпретацию: ‘Когда мне врут, я всегда знаю [осознаю / понимаю], что мне врут’.

Основываясь на общем тезисе современной семантики, что поверхностное поведение языковой единицы обусловлено ее семантической структурой (ср. [Wierzbicka 1985; 1988], ср. также [Bolinger 1977]; [Cristofaro 2003]; [Franks 1995]), нужно предположить, что появление нестандартного союза *когда* в контексте матричного предиката *знать* сигнализирует о некоторой семантической модификации.

Глагол *знать* имеет целый спектр значений, которые могут смещаться от собственно знания (некоторого центра условной шкалы) как в сторону понимания, так и в сторону мнения (уверенности) (см. подробно [Apresjan 1995]). В высказываниях вида (5) речь идет именно о понимании неочевидных вещей (мы еще вернемся к специфике понимания в разделе 4). Кроме того, придаточное имеет обобщенное значение (как и при основных предикатах эмоции и оценки, ср. [Serdobol'skaya 2011: 432]; [Letuchii 2021: 211]) и обозначает повторяющуюся ситуацию (благоприятным условием для употребления актантных *когда*-придаточных является наличие кванторных слов типа *всегда* при матричном предикате *знать*). Тем самым в целом предложения вида *Мать всегда знает, когда ребенок врёт* (\approx ‘Когда ребенок врёт, мать всегда понимает, чувствует, что ребенок врёт’) обозначают своего рода свойство субъекта, характеристику, способность: ‘X-а не обманешь; X-а невозможно ввести в заблуждение; X способен распознать ложь’. Таким образом, у гибридных предложений с матричным *знать* более узкая сфера охвата ситуаций, чем у предложений со стандартными (основными) матричными предикатами (эмоциями и оценками), т.к. стандартные предикаты могут вводить и единичные ситуации (ср.: *Было неприятно, когда он пожаловался начальнику*).

3 Матричный предикат *помнить*

Следующий случай – *когда*-придаточные с глаголом *помнить*.

В качестве фона для сравнения приведем примеры с косвенным вопросом (т.е. с наречием *когда*):

(6) — *А я что-то не помню, когда мы пили на брудершафт* [Василий Белов. Воспитание по доктору Споку (1976)]

(7) *Я не помню, когда и как заснул* [Владимир Войнович. Москва 2042 (1986)]

Для косвенных вопросов характерно отрицание (*не помню*), хотя это не обязательно: *Я помню, когда мы пили на брудершафт*.

ПРИМЕЧАНИЕ. Разумеется, союз *когда* может вводить и придаточное времени: **Помню, когда я написал «Не хлебом единым», роман вызвал мощную волну народного отклика** [Владимир Дудинцев. Добро не должно отступать // «Труд», 1989]; **Помню, когда я шел домой, меня охватило какое-то усталое безразличие** [Михаил Шишкин. Всех ожидает одна ночь (1993-2003)]. В приведенных примерах *помню* – вводное слово, за которым следует сложное предложение с придаточным времени.

Актантные придаточные с союзом *когда*:

(8) *Даже небо не свободно от города. — Люсин кивнул на белые клубы, застывшие над высокими трубами ТЭЦ. — Я еще помню, когда по Бородинскому мосту ходили трамваи. Кажется, тридцать первый и сорок второй* [Ермей Парнов. Александрийская гемма (1990)]
= 'Помню, что [когда-то] ходили трамваи'.

(9) — *Раньше чиновники избегали «светиться» с бандитами, — кивнул Шабанов в сторону дальнего столика. <...> — Ты прав. Я помню, когда за «неделовые контакты» увольняли с должностей* [Даниил Корецкий. Менты не ангелы, но... (2011)]
= 'Помню, как увольняли' / 'Помню, что увольняли'

Можно предположить, что такие предложения являются редуцированным вариантом предложений типа: *Я помню то время / те времена, когда по мосту ходили трамваи*. Дело в том, однако, что *времена* – это абстрактная, почти местоименная лексема. Она сама по себе не может заполнить валентность *помнить* без придаточного Р (?*Я помню то время*; исключение составляют случаи анафорической отсылки, ср.: *А потом меня вывезли оттуда. /.../ Еще маленьким. Я смутно помню то время* [НКРЯ]). Т.е., в конечном счете, пропозициональную валентность глагола *помнить* все равно заполняет ситуация Р 'По мосту ходили трамваи'. Таким образом, актантное *когда*-придаточное (*когда Р*) и блок *то время, когда Р* – это просто синонимичные варианты выражения некоторого смысла – пропозициональной валентности предиката *помнить*.

В наших примерах придаточное имеет обобщенный статус, как и в предложениях с основными предикатами вида *Обидно, когда увольняют с должности; Хорошо / правильно / справедливо, когда увольняют за неделовые контакты*.

Есть примеры, в которых придаточное вводит единичную ситуацию, что не характерно для актантных *когда*-придаточных с матричным предикатом *знать*:

(10) *«Зачем бы Исидору ехать? Разве не метят ли его в митрополиты? Но, мне кажется, он не годится. Я помню, когда он в первый раз приезжал в Москву, за милостынею для Афонских монастырей. Он нечистого православия и чуть ли не волк в овечьей шкуре* [Н. А. Полевой. Клятва при гробе Господнем (1832)]

(11) *Колхознику-то уж что достанется. Помню, когда первый раз в колхоз вступали. Куриц и тех собрали в одно место, овец, одне коты по домам остались* [Василий Белов. Плотницкие рассказы (1968)]

(12) *Я помню первый день их встречи. Помню, когда он приехал в Куоккалу и сказал мне, что теперь для него начинается новая жизнь, — так как он встретил единственную женщину — навеки — до смерти* [Г. Д. Катанян. Азорские острова (1955-1990)]

(13) *При таких истериках я или успокаивала его, или сердилась на него и умоляла не мучить и не пугать меня. Но бывали случаи, когда я боялась за него, когда он, казалось мне, близок к катастрофе. Помню, когда он пришел из Госиздата, где долго ждал кого-то, стоял в очереди в кассу, доказывал что-то, не требующее доказательств. Придя домой, он бросился на тахту во всю свою длину, вниз лицом и буквально завыл: я — больше — не могу...* [Л. Ю. Брик. Из воспоминаний (1956-1977)]

Как видно по датам, такие конструкции встречаются в текстах не только XX в., но и XIX в., т.е. они являются периферийными, но не являются новыми для русского языка.

Весьма показательны примеры, которые иллюстрируют симметричность *когда*- и *как*-придаточных:

(14) *Я помню, как хоронили Брежнева, помню, когда советская номенклатура ездила по Москве с охраной. Где они все? Сейчас о них никто не вспоминает, если только с презрением. А они считали себя очень важными, влиятельными людьми* [«Вестник США», 2003.09.03]

= ‘Помню, как номенклатура ездила с охраной’

Заметим, что союз *как* характерен прежде всего для предикатов восприятия (*видеть* / *видно*, *слышать* / *слышно*). Он предполагает личное присутствие, личное наблюдение, непосредственное восприятие ситуации Р. То же самое предполагают приведенные примеры с *когда*: говорящий повествует о личном опыте. По-видимому, *когда*-предложения не являются распространенными, в частности, в силу того, что прототипическим для таких конструкций является союз *как*. Тем не менее *когда*-придаточные и с семантической, и со стилистической точки зрения являются правильными и находятся в рамках литературной нормы.

Актантные *когда*-придаточные при ментальных предикатах – ситуация достаточно редкая и необычная, поэтому естественно, что здесь есть ряд ограничений. В частности, они не допускают отрицания – в этом случае предложение превращается в косвенный вопрос: (8а) *Я не помню, когда по мосту ходили трамваи*; (10а) *Я не помню, когда он приезжал в Москву*.

Однако сложное предложение как целое может иметь вопросительный статус:

(15) *А помнишь, когда ты мистическим образом проспала скачки в честь Стаса и тебя разбудила птица? Тогда он словно понимал, что надо участвовать в борьбе* [Екатерина Маркова. Каприз фаворита (1990-2000)]

(16) *Я даже к дороге никак не могла привыкнуть, мы едем, едем, а я все чувствую, она железная, она тяжелая... Помнишь, когда я чуть не отстала?.. Меня звали, руки протягивали, а я оцепенела!..* [М. М. Рощин. Эшелон (1972)]

Такие предложения не допускают понимания в качестве косвенного вопроса (‘Помнишь, когда [именно] это было?’), речь идет о том, что субъект помнит факт Р: *Помнишь, как ты проспала?*; *Помнишь, как я чуть не отстала?* (собственно, это не настоящий вопрос, а напоминание, поскольку речь идет об общем фонде знаний собеседников). В таких случаях союз можно опустить без потери информации: *Помнишь, ты проспала?*; *Помнишь, я чуть не отстала?* Если бы это был косвенный вопрос, вопросительное слово невозможно было бы изъять, ср.: *Ты помнишь, когда он вернулся* [‘в какой день’]? и *Помнишь, он вернулся?* – два разных вопроса.

4 Матричные предикаты *понимать* / *понятно*

Наконец, особый, наиболее сложный и парадоксальный тип актантных *когда*-придаточных встречается при матричных предикатах *понимать* / *понятно*. Если в предыдущих случаях союз *когда* при перифразировании соотносился с союзами *что* или *как*, то здесь при перифразировании приходится использовать не только союзы, но и местоименные наречия.

Внутри этой группы есть разные подгруппы.

Во-первых, предложения с актантными *когда*-придаточными при *понимать* теоретически могут иметь, так сказать, стандартную интерпретацию, аналогичную предложениям с матричным *знать*. При этом в предложениях, извлеченных из НКРЯ, стандартная интерпретация встретилась только один раз:

(17) *Мы всегда понимаем, когда нам идут навстречу, когда делают незаслуженные подарки, от которых, однако, нет сил отказаться* [Сергей Лукьяненко. Ночной дозор (1998)]:

= ‘Когда нам идут навстречу, делают незаслуженные подарки, мы понимаем, что нам пошли навстречу, сделали незаслуженный подарок’.

Семантическая структура таких предложений аналогична семантической структуре стандартных предложений типа *Приятно, когда идут навстречу, делают подарки* – ‘Когда идут навстречу, делают подарки, приятно, что идут навстречу, делают подарки’, а также структуре предложений с матричным *знать*: ‘Каждый раз, когда Р, X знает, что Р’

Все остальные найденные нами предложения с *когда*-придаточными при *понимать* / *понятно* имели нестандартные интерпретации. При этом выяснилось, что таких интерпретаций две.

(а) Первый вариант – эмоциональная реакция 'X-у не нравится'.

В таких предложениях матричный предикат имеет при себе отрицание и семантически смещается в зону эмоций, ср.:

(18) — *Я не понимаю, когда вот так вот деньгами разбрасываются, — осуждающе и обиженно заявил Валерий* [Алексей Иванов (Алексей Маврин). Псоглавцы (2011)]

≈ 'Меня возмущает / мне не нравится, что разбрасываются деньгами'.

В принципе, такие предложения имеют стандартную декомпозицию, как у предложений с предикатами основной группы, ср.: *Мне неприятно, когда разбрасываются деньгами* = 'Когда деньги расходуются неразумно, мне неприятно, что они расходуются неразумно'. Нестандартным здесь является значение *не понимаю* – происходит семантический сдвиг из ментальной зоны в эмоциональную, чего не было в случаях *знать* и *помнить*: эти глаголы не приобретали эмоционально-оценочного значения в контексте *когда*-придаточного.

В принципе, *понимать* / *понятно* претерпевает этот сдвиг и в других конструкциях:

Зачем тратить деньги впустую? Я этого не понимаю / Мне это непонятно.

Я не понимаю, почему ты так себя ведешь!

≈ 'Это нельзя понять, потому что это неразумно' → 'Это нельзя одобрить' → 'Я это осуждаю, мне это не нравится'.

В основе таких конструкций лежит риторический вопрос с отрицательной оценкой Р: *Зачем тратить деньги впустую?! → 'Не надо тратить деньги впустую. Это плохо'; Почему ты так себя ведешь! → 'Не надо так себя вести. Это плохо'*. Риторические *зачем/почему*-вопросы и соответствующие гибридные предложения с *не понимаю, когда* являются синонимами: *Зачем разбрасываться деньгами?! ~ Я не понимаю, когда разбрасываются деньгами* (ср. также: *Я не понимаю, зачем разбрасываться деньгами*).

Оценочное значение является результатом естественного семантического сдвига, основанного на имплицатуре: если поведение человека (которое должно быть разумным), нельзя обосновать, объяснить (сделать понятным), то его нельзя одобрить, оно неприемлемо, предосудительно. Такой сдвиг в сферу оценочных значений происходит, например, с прилагательными *понятный* и *непонятный*: *понятный* значит 'разумный, допустимый, приемлемый' (*Это понятный компромисс*), *непонятный* – 'сомнительный, предосудительный, неприемлемый, недопустимый' (ср.: *Он занимается какими-то непонятными делишками*).

Вообще, для прилагательных развитие качественно-оценочного значения является стандартным семантическим сдвигом. У многих существительных в семантическую парадигму также входит качественное значение, ср.: *дипломат* – 'сотрудник дипломатического ведомства' и 'человек, умеющий улаживать конфликтные ситуации, сглаживать противоречия и т.п.'; *ходок* – 'тот, кто ходит пешком' / 'тот, кто занимается ходьбой' / 'человек, посланный с каким-то поручением' и 'мужчина, который «ходит налево»'. При этом выясняется, что у глаголов и предикативов тоже может развиваться такое эмоционально-оценочное значение.

ЗАМЕЧАНИЕ. В экспрессивных конструкциях участвуют и другие ментальные глаголы, ср.: *Знаем мы, как вы плохо играете*, – однако оценочное значение у них не появляется.

(б) Второй вариант: предложение входит в особую – двухчастную – конструкцию, обычно с союзом *но* (реже – с союзом *а*):

(19) *Духи!!! Человек пятнадцать на бортовой машине! <...> Такого я еще не видел. Понятно, когда они передвигаются по дорогам, но сейчас их машина явно остановилась в аккурат на спуске к нашей заставе, перегородив въезд* [С. Погодаев. Письмо другу об афганской войне (2004)]

(20) *Сколько русских парней погибло от рук таджикских боевиков и афганских моджахедов. Понятно, когда погибает человек, добровольно приехавший в этот чужой мир воевать. Он хотел таким путем решить свои финансовые проблемы. Но когда гибнут пацаны, призванные в армию после школы и безжалостно брошенные почти сразу же в «горячие точки», — это преступление, это трагедия, это вечная боль и слезы родителей погибших* [Алексей Грачев. Ярый против видеооператоров (1999)]

В этой конструкции встречаются не только предикативы, но и глаголы: *Я еще могу понять, когда ребенок верит в эти сказки, но ты-то как мог поверить?*

Если в контексте *знать когда*-придаточное может в каких-то случаях иметь две интерпретации – как союзное и как косвенный вопрос, то в контексте *понимать* (в обсуждаемых примерах) такой неоднозначности не возникает – *когда* имеет статус союза и не может быть понято как наречие в косвенном вопросе. Другое дело, что сама эта конструкция – двухчастная, и она требует «продолжения»:

(21) *Известно, что клетки крови недолговечны. Максимальный срок их жизни не превышает двух месяцев. Так откуда же берутся новые? Понятно, когда клетки формируются в ходе развития зародыша, но во взрослом организме?* [«Знание — сила», 2003]

В таких предложениях невозможно рассматривать *когда*-придаточные как косвенный вопрос: ‘Нам понятно, когда именно [= в какой момент] формируются клетки крови’, – потому что ответ содержится в самом предложении: в ходе развития зародыша. Смысл этого предложения другой: ‘Ситуация с развитием клеток крови зародыша понятна, ситуация с клетками крови взрослого непонятна’. При этом смысл *когда*-придаточного не эквивалентен смыслу *что*-придаточного: *Понятно, что клетки крови формируются в ходе развития зародыша.*

Аналогично:

(22) *Понятно, когда мумифицированные останки находят в жарких и сухих пещерах Средней Азии, но в Сумгане?* [Константин Серафимов. Экспедиция во мрак (1978-1996)].

Смысл этой конструкции можно передать разными способами:

(а) ‘То, что мумифицированные останки находят в жарких и сухих местах, понятно / объяснимо / естественно / нормально, ожидаемо, но откуда они в нежарких и влажных местах? Это непонятно’.

(б) ‘Понятно, почему мумифицированные останки находят в пещерах Средней Азии, – там жарко и сухо, но почему их находят в Сумгане, где нет таких условий?’.

В общем виде семантическую структуру таких предложений можно представить следующим образом:

‘Имеет место ситуация Р в условиях С1 и ситуация Р в непохожих условиях С2.

То, что ситуация Р произошла в условиях С1, объяснимо / понятно (она произошла из-за С1). То, что Р произошла в условиях С2, необъяснимо / непонятно’.

Или: ‘Почему Р произошла в условиях С1, объяснимо / понятно; почему Р произошла в условиях С2, необъяснимо / непонятно’.

Итак, в контексте *понимать / понятно когда*-придаточное синонимично не *что*-придаточному и не *как*-придаточному, а косвенному вопросу, но не вопросу *когда*, а вопросу *почему*.

С одной стороны, это довольно парадоксально (на фоне других матричных предикатов в гибридных предложениях), с другой – напротив, закономерно. В предложениях с *понимать / понятно* присутствует характерная для *когда*-придаточных идея узуальности, обобщенности, за которую отвечает союз *когда*. Но добавляется идея причинности, и за нее отвечает уже другой компонент этой структуры – матричный предикат *понятно / понимать*. Как и в случае *знать*, с которым происходило семантическое изменение (смещение *знать* в семантическую зону ‘понимать / чувствовать / догадываться’), с *понимать / понятно* тоже происходит некоторое семантическое изменение, а именно – появляется акцент на причине, что не удивительно, т. к. *понимать* системно связано с причиной.

Как показано в [Апресжан 1995], *понимание*, в отличие от *знания*, не сводится к получению информации о некоторой ситуации, а предполагает анализ, размышления, умозаключения и подобные ментальные операции, а сама получаемая в процессе понимания информация является нетривиальной, неочевидной. Т.е. понимают вещи, которые нельзя узнать путем непосредственного наблюдения. В типичном случае *понять* связано с наличием объяснения, т.е. со знанием причины, механизма возникновения данной ситуации: человек, который понимает, может объяснить, как ситуация возникла, т.е. сообщить ее причину. Таким образом, *понять* системно и закономерно связано с *почему*, т.е. с выяснением причины, с объяснением: *понял* → ‘могу объяснить’. Тем самым *понимать / понятно* семантически сближается со смыслами *объяснимо, закономерно, естественно, нормально, ожидаемо, не удивительно, ср.:*

Не удивительно, что клетки формируются в ходе развития зародыша ('понятно, почему / каким образом'), *удивительно, что они формируются и во взрослом организме*.

Естественно, эта конструкция идиоматизирована, и у нее есть ограничения. В общем случае ментальные глаголы (при наличии видовой пары) могут употребляться в обоих видах: *Я понял, что он обиделся / Я понимаю, что он обиделся; Я понял, когда он взял кошелек / Я понимаю, когда он взял кошелек*. В нашей гибридной конструкции употребляется только несовершенный вид: *Я еще понимаю, когда ты отпрашиваешься на три дня, но на месяц – это уже слишком VS. ?Я еще понял, когда ты отпрашился на три дня, но на месяц – это уже слишком*.

Интересно, что некоторые высказывания данной модели сближаются с типом (а): *Я еще понимаю, когда человек отпрашивается с работы раз в месяц, но каждый день – это уже слишком VS. Я не понимаю, когда человек отпрашивается каждый день!*. Однако это отдельная тема, в которую мы не можем углубляться в силу ограниченности объема статьи.

В целом можно констатировать, что гибридные *когда*-придаточные не «привязаны» только к матричным предикатам эмоции и оценки и должны рассматриваться в более широком контексте конструкций с союзом *когда*.

5 Конструкции с нестандартными матричными предикатами с точки зрения автоматического анализа текста

Очевидно, что главная проблема автоматического анализа текста, связанная с данными конструкциями, – различить актантные клаузы с союзом *когда* и местоимением *когда*. Как видно из приведенного выше материала, это нетривиальная задача, т.к. далеко не всегда в тексте есть однозначные маркеры различий. Тем не менее иногда такие маркеры обнаруживаются. Суммируем эти показатели. *Когда*-придаточное интерпретируется как косвенный вопрос (с наречием *когда*) в следующих случаях: 1) вхождение *когда* в сочинительную группу с другим местоимением (*Понятно, когда и как следует повышать пенсии* [«Советская Россия», 2003.09.01]); 2) наличие при *понятно* вспомогательных глаголов прошедшего или будущего времени (*Уже тогда стало понятно, когда ты покинешь «Дэвилз»* [Советский спорт, 2013.07.16]; *25 июня готовится совещание, по результатам которого станет понятно, когда граждане смогут полноценно пользоваться новым сервисом* [Известия, 2014.06.17]); 3) наличие отрицания при матричных предикатах *знать, помнить* (*Это человек, который сам не знает, когда он врет* [Б. С. Житков. Виктор Вавич (1941)]).

Когда-придаточное при матричных предикатах *понятно, понимать* интерпретируется как актантное с союзом *когда* при наличии в ближайшем контексте союзов *но* или *а*, вводящих вторую часть конструкции (*Понятно, когда из-за изменившихся торговых потоков умирает торговый город. Но Детройт всего полвека назад был автомобильной столицей США* [Комсомольская правда, 2013.09.12]; – *Понятно, когда раскручивают «Ледниковый период». Купил коньки, пошел в парк Горького – и ты уже фигурист. А какой прок от «Вышки»?* [Советский спорт, 2013.06.21]).

Эти правила могли бы использоваться в фильтрах для автоматического распознавания актантных придаточных, но, разумеется, они способны распознать лишь какой-то процент употреблений, и, как и в других случаях омонимии, может потребоваться экспертная оценка лингвиста.

Acknowledgements

Исследование выполнено при финансовой поддержке РФФИ и Национального научного фонда Болгарии, проект № 20-512-18005.

The reported study was funded by RFBR and National Science Foundation of Bulgaria (NSFB), project number 20-512-18005.

References

- [1] Apresyan Yu. D. The problem of factivity: to know and synonyms [Problema faktivnosti: znat' i ego sinonimy] // «Voprosy yazykoznanija». 1995. № 4; perepechatano v: Apresyan Yu. D. Selected Works, v. 2 [Izbrannye trudy, t. 2]. — Moscow, YaSK Publ., 1995. — P. 403–433.
- [2] Bolinger D. Meaning and Form. English Language Series, 11. — London: Longman, 1977.
- [3] Cristofaro S. Subordination. — Oxford: Oxford University Press, 2003.
- [4] Franks S. Parameters of Slavic morphosyntax. — New York; Oxford: Oxford University Press, 1995.
- [5] Letuchii A. B. Russian language about situations. Constructions with sentential actants in Russian [Russkii yazyk o situatsiyakh. Konstruktsii s sententsial'nymi aktantami v russkom yazyke]. — Saint Petersburg, Aleteiya Publ., 2021.
- [6] Maksimov L. Yu. Multidimensional classification of complex sentences [Mnogomernaja klassifikacija slozhnopodchinennykh predlozhenij]. — Stavropol', SGU Publ., 2011.
- [7] RG-80 – Russian grammar. Vol. 2. Ed. by Shvedova N. Yu. [Russkaya grammatika. T. 2. Pod red. Shvedovoy N. Yu.]. — Moscow, Nauka Publ., 1980.
- [8] Serdobol'skaya N. V. [On the typology of expressing the meaning of a generic event in constructions with sentential actants] // Acta linguistica petropolitana. Trudy instituta lingvisticheskikh issledovaniy. Tom VII, chast' 3 [Acta linguistica petropolitana. Transactions of the Institute for Linguistic Studies]. Vol. X. Part 2. — Saint Petersburg, Nauka Publ., 2011.
- [9] Skoblikova E.S. Modern Russian language. The syntax of a complex sentence [Sovremennyi russkii yazyk. Sintaksis slozhnogo predlozheniya]. — Moscow, Flinta: Nauka Publ., 2006.
- [10] Wierzbicka A. Lexicography and conceptual analysis. — Ann Arbor: Karoma, 1985.
- [11] Wierzbicka A. The semantics of grammar. — Amsterdam: Benjamins, 1988.

Sweet ideal

Irina Levontina

Russian Language Institute RAS

irina.levontina@mail.ru

Abstract

The work is devoted to the semantics, the structure of polysemy, the syntax of the word *ideal*. The word *ideal* at first glance does not seem complicated either from the point of view of the semantics and structure of polysemy, or from the point of view of the actant structure. *X is the ideal of Y from the point of view of Z* - means that the object X, either real or imaginary, belongs to class Y and fully corresponds to Z's idea of what an object of class Y should be. Moreover, this is such a complete correspondence that in the life of such practically does not happen. The paper considers some non-trivial features of this word. In particular, it turns out that the genitive form of the word *ideal* can occupy three positions: not only the ideal of Petya (=Petya's ideal) and the ideal of wife (the idea of what a wife should be), but also the ideal of service (*ideal sluzhenija*) - in the sense that service is the content of the ideal. In addition, it turns out that the semantics and structure of the polysemy of the word *ideal* have changed during its existence in the Russian language. In Pushkin's expression *Tatyany milyj ideal* the sweet ideal does not imply that Tatyana has some kind of ideal, but that the ideal is Tatyana herself. More precisely, the *ideal of Tatiana* here is the same as the *image of Tatiana*. Different words with this root have various structure of polysemy.

Keywords: polysemy structure; actant structure; ambiguous filling of valencies; evaluation

DOI: 10.28995/2075-7182-2022-21-367-374

Милый идеал

Левонтина И. Б.

ИРЯ РАН им. В. В. Виноградова РАН

irina.levontina@mail.ru

Аннотация

Работа посвящена семантике, структуре многозначности, модели управления слова идеал. Слово *идеал* на первый взгляд не кажется сложным ни с точки зрения семантики и структуры многозначности, ни с точки зрения актантной структуры. *X является идеалом Y-а с точки зрения Z* – значит, что объект X, реальный или воображаемый, принадлежит к классу Y и полностью соответствует представлению Z о том, каким должен быть объект класса Y. Причем это настолько полное соответствие, что в жизни такого практически не бывает. В работе рассматриваются некоторые нетривиальные особенности этого слова. В частности, обнаруживается, что форма родительного падежа при слове *идеал* может замещать три позиции: не только *идеал Пети* (=Петин идеал) и идеал жены (представление о том, какой должна быть жена), но и *идеал служения* – в том смысле, что служение и есть содержание идеала. Кроме того, выясняется, что семантика и структура многозначности слова *идеал* за время его существования в русском языке изменились. В пушкинской формулировке *Татьяны милый идеал* подразумевается не то, что у Татьяны есть какой-то идеал, а то, что идеал – это сама Татьяна. Точнее, *идеал Татьяны* здесь – это то же, что *образ Татьяны*. У разных слов с этим корнем структура многозначности складывалась по-разному.

Ключевые слова: структура многозначности; актантная структура; неоднозначное заполнение валентностей; оценка

1. Семантика слова идеал

Слово *идеал* на первый взгляд не кажется сложным ни с точки зрения семантики и структуры многозначности, ни с точки зрения актантной структуры. *X является идеалом Y-а с точки зрения Z* – значит, что объект X, реальный или воображаемый, принадлежит к классу Y и полностью соответствует представлению Z о том, каким должен быть объект класса Y. Причем это настолько полное соответствие, что в жизни такого практически не бывает. Конечно, это огрубленная экспликация, она нуждается в некоторых уточнениях. Это особенно заметно при сопоставлении существительного *идеал* с прилагательным *идеальный*, значение которого несколько шире.

Так, можно сказать *идеальное преступление* и даже *идеальное убийство*:

- 1) *Оставалось найти недорогого киллера, деньги на его гонорар и уж тогда спланированное мною **идеальное** убийство станет очередной из неразрешимых загадок.* [Андрей Курков. Милый друг, товарищ покойника (2001)]

Однако *идеал преступления <убийства>* – едва ли. Значит, *идеал* в гораздо большей степени связывается с чем-то положительно оцениваемым. Ср. также странное *идеал круга* при естественном *идеальный круг*. Геометрическая фигура не предполагает никакой оценки, поэтому существительное *идеал* здесь не годится.

Уточним, что оценки у разных людей различаются, и идеалы не у всех одинаковы, как в пушкинском *Какая ужасная разница между идеалами бабушки и внушек!* Тем не менее, идеал подразумевает положительную оценку со стороны его носителя. Просто у другого человека могут быть свои идеалы, чуждые и неприятные говорящему:

- 2) *А после
Заводил о том, что духу чужд
Подлый идеал всеобщей пользы,
Низких нужд.*
[А. П. Межиров. Эта встреча, может быть, последняя... (1971)]
- 3) *Мы призываем ученых Украины, наших старых добрых коллег по совместной научной работе, не верить **лживым идеалам**, не поддаваться на провокации националистов и помнить нашу великую общую историю* (Обращение РФФИ 06.03.2022).

Надо, прочем, отметить, что два последних примера выглядят несколько парадоксально.

Кроме того, прилагательное *идеальный* может характеризовать бытовые объекты небольшой значимости; ср. *идеальный пельмень*, *идеальная заплатка*. Существительное *идеал* так обычно не используется (странно – *идеал пельменя <заплатки>*), оно предполагает некоторую масштабность и возвышенность:

- 4) *Если в сердце твоём оскорблен **идеал**,
Идеал человека и света,
Если честно скорбишь ты и честно устал, —
Отдохни над страницей поэта.*
[С. Я. Надсон. «Если душно тебе, если нет у тебя...» (1880)]
- 5) *Она меня очаровала,
Я в ней нашел все красоты,
Все совершенства **идеала**
Моей возвышенной мечты.*
[Н. М. Языков. Элегия (1825.04.01)]

Заметим, что первое место здесь не занято: свойство быть *идеалом* никому не приписывается, а существует как самостоятельная умозрительная сущность.

В целом, однако, семантика слова *идеал* кажется вполне прозрачной.

2. Актантная структура слова идеал

Прозрачна и его актантная структура. *Идеал* подразумевает три семантических актанта: сам оцениваемый объект (*Ты (мой) идеал*), класс, к которому принадлежит объект (*идеал ученого*), и субъекта оценки (*мой идеал*).

В приведенных примерах первый семантический актант не выражается зависимым самого слова *идеал*, а присоединяется к нему через связку. Но действительно ли у этого слова реализуются только два синтаксических актанта?

Оказывается, что это не так. Первый актант тоже может реализоваться при слове *идеал* синтаксическим зависимым. Дело в том, что форма родительного падежа при слове *идеал* может замещать три позиции: не только *идеал Пети (=Петин идеал)* и идеал *жены* (представление о том, какой должна быть жена), но и *идеал служения* – в том смысле, что служение и есть содержание идеала:

- б) *Среди идеализированных образов прошлого особо привлекал уклад средневековых рыцарских сообществ с их почитанием чести, благородства, с*

их идеалами служения, мужества и неизменным блеском ритуалов. [Маргарита Силаева. «Дон Кихот самодержавия» // «Родина», 2008]

- 7) *Разработка этического кодекса чиновника. Идеал служения обществу, государству, человеку. Нормы иерархической подчиненности, законопослушности, добросовестного профессионального труда* [Н. Л. Захаров. Система регуляторов социального действия российских государственных служащих (теоретико-социологический анализ) (2002)]
- 8) *Полюбишь идеал изящного душой,
И скажешь: вот поэт, природы друг — и мой!*
[П. И. Шаликов. Наши стихотворцы (1812)]
- 9) *Какое непонятное цветенье
Возвышенных и нежных идеалов
Гражданственности, мудрости, закона!*
[В. А. Луговской. Акрополь (1932-1933)]

Однако природа этого актанта неочевидна. С одной стороны, кажется, что слово *идеал* в этом отношении устроено аналогично слову *идея*. Можно сказать *идея статьи Якобсона*, и это будет валентность темы. А можно – *идея поездки на море*, то есть *идея поехать на море*, и это будет валентность содержания.

В случае слова *идеал* дело обстоит не совсем так. *Идеал служения* нельзя перефразировать как *идеал служить*. Это различие не случайно. Сочетания типа *идеал гражданственности*, с другой стороны, сближаются с сочетаниями *чувство любви, свойство транзитивности, грех чревоугодия* и т. п. Ср.:

- 10) *И сама жизнь, явление жизни, дар жизни так захватывающе нешуточны!*
[Б. Л. Пастернак. Доктор Живаго (1945-1955)]

Особенность подобных сочетаний в том, что, хотя формально здесь имеет место подчинение, но содержательно слабо представлен важный признак подчинения – сводимость именно к главному компоненту (*Мне нравится идея поехать на море* = *Мне нравится идея*). Здесь естественнее свернуть сочетание можно к формально зависимому: *испытывать чувство любви* = *испытывать любовь, обладать свойством транзитивности* = *обладать транзитивностью*). В этом отношении очень показателен приведенный выше пример из Пастернака, где через запятую идут *жизнь, явление жизни и дар жизни*: все это синонимы, просто сначала дается прямое указание на жизнь, а во втором и третьем случае в качестве управляющего слова отдельно выносятся имена классов, к которым относится жизнь.

Необходимо отметить, что такой способ выражения сильно лексикализован: можно сказать *грех самоубийства*, но не **преступление убийства*.

Что же касается слова *идеал*, оно в этом отношении занимает промежуточное положение. Трудно сказать, *верен идеалу гражданственности <служения>* – это скорее *верен идеалу* или *верен гражданственности <служению>*?

Ситуация с двумя слабо противопоставленными валентностями слова *идеал* вписывается в круг проблем, которые активно обсуждаются в последнее время, в частности, в работах Л. Л. и Б. Л. Иомдиных [2020]. В этом отношении давно известны, например, валентности темы и содержания, а в указанных работах рассматриваются, в частности, валентности стимула и реакции в таких глаголах, как *возражать, возмущаться, извиняться*: валентность стимула выражает положение дел, валентность реакции вводит речевой акт, совершаемый субъектом в качестве отклика на это положение дел и предлагающий объяснение; ср. *Иван извинился, что не пришел на мой день рождения* [положение дел], *Иван извинился, что плохо себя чувствовал*¹ [речевая реакция Ивана на положение дел].

¹ Некоторые из подобных употреблений, включая этот, конечно, находятся на грани нормы. Как отмечает рецензент «Диалога», «В НКРЯ есть несколько примеров, где *извинился* означает ‘сказал (в оправдание своего/чьего-то отсутствия в актуальном времени)’», и здесь везде в придаточном наст. время (т.е. это как бы прямая речь): *Он извинился, что ему надо домой, и стал прощаться* (Л. Н. Толстой, Воскресение); *Наконец, она извинилась, что мужу нездоровится, и все стали усаживаться за стол*. [Михаил Шишкин. Всех ожидает одна ночь (1993-2003)]; *Извинившись, что неважно себя чувствую, я ушел*. [Михаил Шишкин. Всех ожидает одна ночь (1993-2003)].

3. Устаревшее значение слова идеал

Вернемся к конструкциям вида *служить идеалу добра* и функции в них формы родительного падежа. Обращает на себя внимание то, что в этой позиции трудно встретить слова с предметным значением. Однако это так лишь для современного языка. В пушкинской формулировке *Татьяны милый идеал* подразумевается не то, что у Татьяны есть какой-то идеал, а то, что идеал – это сама Татьяна:

11) *Но те, которым в дружной встрече
Я строфы первые читал...
Иных уж нет, а те далече,
Как Сади некогда сказал.
Без них Онегин дорисован.
А та, с которой образован
Татьяны милый идеал...
О много, много рок отъял!*

Точнее будет сказать, что *идеал Татьяны* здесь – это то же, что *образ Татьяны*. Такие употребления слова *идеал* в пушкинское время встречались:

12) *Признался бы, что он подобной вам
Обязан был прелестным идеалом
Своих богинь.*

[А. А. Дельвиг. К Е.А. Кильштетовой (1818)]

То есть, образы богинь создавал по образцу красивых женщин. При этом и слово *идеал*, и слово *образ* могли метонимически указывать и на сам объект, в котором этот образ воплотился:

13) *На милый идеал пастушки
Лорнет наводит селадон,
Платок свой алый расправляет,
Вздыхает раз, вздыхает два,
И к ней, кобенясь, обращает
Он следующие слова:
«Приветствую мольбой стократной
Гебею zdeшней стороны!»*

[П. А. Вяземский. Первый отдых Вздыхалова (1811)]

Наводит лорнет на *идеал пастушки* – то есть смотрит на саму пастушку. Аналогично и в стихотворении М. Ю. Лермонтова Хаджи Абрек, герой которого, чтобы отомстить убийце брата, убивает не его самого, а его возлюбленную:

14) *«Он что-нибудь да в мире любит:
Найду любви его предмет,
И мой удар его погубит!»
Свершилось наконец. Пора!
Твой час пробил еще вчера.
<...>
Когда сегодня в первый раз
Я увидел твой образ нежный,
Тоскою горькой и мятежной
Душа, как адом, вся зажглась.*

Таким, в ранних примерах мы видим подтверждение того, что у слова *идеал* поначалу было еще одно значение – визуализация какой-то идеи, представления, свойства, а также воплощение этого представления или свойства в материальном объекте. Интересно, что позитивная оценка этого представления не обязательна:

15) *Ужель бездушный идеал
Неотразимого разврата
Тебя, как жертву каземата,
Рукой поносной оковал?
[А. И. Полежаев. Раскаяние (1832-1833)]*

16) *Графиня же была погружена
В самой себе, в волшебстве моды новой,*

*В своей красе надменной и суровой.
Она казалась хладный идеал
Тщеславия. Его б вы в ней узнали;
Но сквозь надменность эту я читал
Иную повесть: долгие печали,
Смиренье жалоб.... В них-то я вникал,
Невольный взор они-то привлекали....*
[А. С. Пушкин. Домик в Коломне (1830)]

Заметим, что у слова *идеал* в этом смысле – другая актантная структура. У него нет субъектной валентности (**Она казалась моим идеалом тщеславия*) и нет валентности класса. Сейчас это значение устарело. Как это часто бывает в подобных случаях, мы часто принимаем старые употребления за реализацию современного значения (то есть, мы воспринимаем *Татьяны милый идеал* не как «Татьяны милый образ», а как «Татьяна – милый идеал»).

В свое время В. В. Виноградов довольно подробно описывал функционирование слова *идеал* в языке Пушкина, однако под совершенно другим углом зрения. Его интересовало, как в истории этого слова отразилось движение Пушкина от «романтических» эстетических представлений к «реалистическим»: «В период расцвета своей работы над национально-реалистическим стилем Пушкин решительно противопоставляет «высокопарным мечтаньям» своей весны, теории возвышенного предмета «прозаические бредни» реализма, поэзию живой жизни. В этом отношении любопытна сложная и противоречивая эволюция употребления слова *идеал* в поэтическом языке Пушкина. <...> распространившееся под влиянием идеалистической эстетики великих философов начала XIX века слово *идеал* в стиле Пушкина выходит далеко за пределы его первоначального романтического употребления. Связывая с этим словом представление о наиболее полном и совершенном отражении действительности, Пушкин присоединяется к тезису, что «цель искусства есть идеал, а не нравоучение», и констатирует: "Мелочная и ложная теория, утвержденная старинными риториками, будто бы польза есть условие и цель изящной словесности, сама собою уничтожилась" (Мнение М. Е. Лобанова о духе словесности). Тут закладываются основы реалистической эстетики, противопоставляющей классицизму и натурализму принцип всестороннего отражения действительности в художественном слове с точки зрения глубоководержательного "идеала". В поэтическом языке Пушкина слово *идеал* применяется как к возвышенным, так и к низменным предметам; по мысли Пушкина, целью поэзии может быть любой предмет». [Виноградов 1999]

Действительно, судя по пушкинским примерам, слово *идеал* было приметой романтического стиля, над которым Пушкин в «Евгении Онегине» уже иронизирует (*темно и вяло*):

17) *И наконец перед зарею,
Склонясь усталой головою,
На модном слове – идеал,
Тихонько Ленский задремал...*

Тогда становится понятно, что в других известных строках из того же произведения Пушкин намеренно сталкивает слово *идеал* с нарочито бытовыми, крестьянскими деталями:

18) *Мой идеал теперь – хозяйка,
Мои желанья – покой,
Да щей горшок, да сам большой.*

Между тем, уже по приводимым Виноградовым многочисленным примерам видно, что не только стилистическая окраска, но и семантика, и грамматика этого слова в первой половине XIX были иными. Некоторые из таких отличий уже обсуждались выше, приведем еще только одно:

19) *Муж-мальчик, муж-слуга, из жениных пажей —
Высокий идеал московских всех мужей.*
[А. С. Грибоедов. Горе от ума (1824)]

Эти хрестоматийные строки с точки зрения современной нормы должны были бы пониматься в том смысле, что московские мужья имеют такое идеальное представление (субъектная валентность). Между тем, очевидно, что имеется в виду другое – то, что в современном языке выражалось бы с использованием формы единственного числа: *высокий идеал московского мужа* (валентность класса).

4. Структура многозначности слов с корнем идеал

Итак, мы установили, что структура многозначности слова *идеал* сильно изменилась: в первой половине XIX в. у него было также значение 'образ', которое впоследствии устарело. При этом у некоторых родственных слов соответствующие два значения полноценно сформировались и существуют до сих пор. Так, у слова *идеалист* есть значение 'сторонник нематериалистической философии' и значение 'человек в розовых очках или борец с ветряными мельницами'² (тот, кто склонен *идеализировать* или придавать слишком большое значение отвлеченным идеалам).

- 20) *Понятно, что, скажем, философу-мистику, идеалисту в условиях диктатуры пролетариата и марксизма делать нечего...* [Даниил Гранин. Зубр (1987)]
- 21) — *Жалко, Лев Николаич помер — послушал бы хоть. Тоже был идеалист То на публичной лекции о своём кумире Ньютоне прогудит из-под жёлтых усов: — Тут мне прислали записку: "Маркс написал, что Ньютон — материалист, а вы говорите — идеалист." Отвечаю: Маркс передергивает. Ньютон верил в Бога, как всякий крупный ученый.* [Александр Солженицын. В круге первом, т.1, гл. 1-25 (1968) // «Новый Мир», 1990]
- 22) *безнадёжный. — В отношении мужичка — да, был идеалист.* [Василий Шукшин. Печки-лавочки (1970-1972)]
- 23) *Взаимоотношения парткома и профкома, секретаря обкома и директора крупного треста, мастера и рабочего были известны ему до тонкостей, и фантазия — отказавшийся от незаработанной премии бригадир-идеалист — выростала на прочном основании.* [Алексей Филиппов. Продавцы воздуха. Производственная тема-2002 в Новом драматическом театре (2002) // «Известия», 2002.09.23]

В (20) и (21) представлено философское значение, в (22) – указание на человека, склонного видеть только хорошее, в (23) – указание на человека, который руководствуется высокими идеалами, а не практическими соображениями.

Аналогично устроено и слово *идеализм*. А вот слово *идеализировать* соотносится только с *идеалом* как совершенством:

- 24) *Не хочу идеализировать Колеватова: у него были свои недостатки.* [И. Э. Кио. Иллюзии без иллюзий (1995-1999)]
- 25) *Впрочем, — добавил он, вспомнив свои двадцать пять лет лагерей, — наше время я тоже не идеализирую.* [Владимир Войнович. Иванькиада, или рассказ о вселении писателя Войновича в новую квартиру (1976)]
- 26) *Достоевский никогда не идеализировал народ и не поклонялся ему как кумиру.* [В. С. Соловьев. Три речи в память Достоевского (1881-1883)]

Интересно обстоит дело со словами *идеальный/идеально*. Они устроены похоже на слово *идеал*: в основном они соотносятся с идеей совершенства, однако иногда в них реализуется и возможность, не получившая полного развития, когда *идеальный* означает трансцендентный, умозрительный. Чаще всего это происходит в философских текстах при противопоставлении материальное/идеальное:

- 27) *Таким образом, двойственность и противоположение идеального и материального мира, истины и факта, здесь остается неразрешенною, примирения нет.* [В. С. Соловьев. Чтения о Богочеловечестве (1878)]
- 28) *Однако "основной вопрос философии" получает своеобразное преломление во взглядах Витгенштейна на математику: сознание делается здесь частью бытия, а мир идеальный — частью мира материального.* [В. А. Успенский. Витгенштейн и основания математики (2002)]

Ср., однако, знаменитый фрагмент из «Москвы-Петушков» Вен. Ерофеева:

- 29) *Оказывается, если вы уже выпили пятую, вам надо и шестую, и седьмую, и восьмую, и девятую выпить сразу, одним махом — но выпить идеально, то есть выпить только в воображении. Другими словами, вам надо одним волевым усилием, одним махом — не выпить ни шестой, ни седьмой, ни восьмой, ни девятой. А выдержав паузу, приступить непосредственно к десятой.*

² В БТС это три значения, но мы здесь второе и третье объединяем.

В этом примере хорошо видно, что столь изысканное употребление слова *идеально*, тем не менее, совершенно понятно, поскольку оно реализует имеющийся у слова семантический потенциал.

Уместно будет также вспомнить знаменитые строки посвящения к «Евгению Онегину»:

30) *Но так и быть — рукой пристрастной
Прими собранье пестрых глав,
Полусмешных, полупечальных,
Простонародных, **идеальных**,
Небрежный плод моих забав,
Бессонниц, легких вдохновений,
Незрелых и увядших лет,
Ума холодных наблюдений
И сердца горестных замет.*

Здесь мы опять сталкиваемся с тем, что хрестоматийность строк мешает задуматься об их содержании. В каком смысле Пушкин называет свой роман *идеальным*? Конечно, не в смысле совершенства: напротив, Пушкин шутливо принижает свое произведение (оно и недостойно адресата, и пестрое, и небрежное, и т. д.). По-видимому, под *идеальным* имеется в виду, что у автора слишком разыгралось воображение. Кстати, и *простонародный* здесь – не социальная характеристика, а функционально-стилистическая (*слог простонародный* для Пушкина – это живая речь образованного общества).

5. Семантика числовых форм слова идеал

Следующее наблюдение касается семантических приращений, которые у слова идеал появляются в форме множественного числа:

31) *Когда в степях, где жгутся скалы
Огнем полуденных лучей,
Где загорались **идеалы**
Его божественных речей; —*

[К. К. Случевский. Влад. Серг. Соловьеву, перед отъездом его в Палестину (1898.03.28)]

32) *Меж гражданских живя высот
и **общественных идеалов**,
всяких сладостей и красот
наша юность не признавала.*

[Я. В. Смеляков. Прогулка [Фрагменты из второй части повести в стихах «Строгая любовь», 4] (1953-1956)]

Очень часто форма *идеалы* понимается как указание не на несколько отдельных идеалов, а на нерасчлененную совокупность представлений о прекрасном и должном, как это и происходит в приведенных примерах. Это совершенно стандартный смысловой сдвиг, который происходит у многих абстрактных слов в форме множественного числа; ср. описание слов *план* и *намерение* в НОССе (статья составлена мной):

У слов *намерение* и *план* форма МН может обозначать не только несколько разных объектов (*Все его планы один за другим срывались*), но и один объект, т. е. употребляться применительно к единичному действию. Обычно в этом случае форма МН имеет значение неопределенности. Ср. *Какие у тебя планы? В магазин идешь?* и *Какой у тебя план?* В последнем случае говорящий считает, что собеседник имеет некоторую цель, и предполагает, что должен быть план ее достижения. <...> Аналогично высказывание *Каковы ваши намерения?* означает, что спрашивающий не обязательно считает, что у адресата есть определенные намерения.

В подобных случаях формы единственного и множественного числа полностью или частично взаимозаменяемы:

33) *И если я наполнил жизнь борьбою
За **идеал** добра и красоты,
И носит песнь, слагаемая мною,
Живой любви глубокие черты —*

О мать моя, подвигнут я тобою!

[Н. А. Некрасов. Из поэмы «Мать» (1850-1877)]

В таком контексте свободно может использоваться и форма множественного числа. Стоит, впрочем, отметить, что некоторое различие между числовыми формами слова *идеал* сохраняется и в подобных случаях, и здесь снова можно обратиться к примеру из «Евгения Онегина»:

34) *Одну Россию в мире видя,
Преследуя <лаская в ней> свой идеал,
Хромой Тургенев им внимал
И, плети рабства ненавидя,
Предвидел в сей толпе дворян
Освободителей крестьян.*

Смысл здесь не в том, что Россия – идеал Николая Тургенева, а в том, что он лелеял мечту о прекрасной России будущего – без крепостничества. Это не туманные обобщенные *идеалы*, а нечто гораздо более конкретное.

Семантика неопределенности, которая проявляется у слова *идеал* в форме множественного числа, приводит к образованию у него яркого специфического круга употреблений. *Идеалы* в этом случае – это нечто неконкретизируемо позитивное. Это делает форму *идеалы* исключительно пригодной для манипулятивного дискурса. Попытка конкретизации того, в чем, собственно, состоят эти идеалы, нередко приводит к появлению неловких и бессмысленных формулировок, как в тексте «Основ государственной политики по сохранению и укреплению традиционных российских духовно-нравственных ценностей»:

35) *Перечень разрушительных и чуждых идей и ценностей <...>*

отрицание идеалов патриотизма

отрицание идеалов служения Отечеству

отрицание идеалов продолжения рода

отрицание идеалов созидательного труда

отрицание идеалов позитивного вклада России в мировую историю

отрицание идеалов позитивного вклада России в культуру

<https://meduza.io/feature/2022/01/24/minkult-rossii-sostavil-reestry-traditsionnyh-tsennostey-i-chuzhdyh-idey>

Как, например, понять формулировку *идеалы позитивного вклада России в культуру*? Никакого позитивного вклада нет, есть лишь мечта об этом? Причем эту мечту еще кто-то отрицает?

Заключение

Мы установили, что русское слово *идеал* имеет множество интересных особенностей – семантических, синтаксических, словообразовательных, а также, что на протяжении своей более чем двухвековой истории оно претерпело любопытные изменения.

References

- [1] Apresjan Yu. D. (2006). Correspondence types of semantic and syntactic actants // Problemy tipologii i obshchey lingvistiki. — SPb., — P. 15–27.
- [2] Vinogradov V. V. History of words: Approximately 1500 words and expressions and more than 5000 words, connected with them // Ros. akad. nauk. Otd-niye lit. i yaz. Nauch. sovet "Rus. yaz.". In-t rus. yaz. im. V. V. Vinogradova. - M., 1999. - 1138 P. ISBN 5-88744-033-3
- [3] Iomdin B. L., Iomdin L. L. Valency structure of some speech predicate words: new findings // Komp'yuternaya lingvistika i intellektual'nyye tekhnologii: Po materialam yezhegodnoy mezhdunarodnoy konferentsii «Dialog» (Moskva, 17–20 iyunya 2020 g.). Vyp. 19 (26). S. 385–400.
- [4] Iomdin B. L., Iomdin L. L. Valency of Fact and Valency of Argumentation // Trudy IRYA RAN, 2020, №3, P. 51–63. <https://doi.org/10.31912/pvrli-2020.3.4>

Artificial Text Detection with Multiple Training Strategies

Bin Li^{1*}, Yixuan Weng^{2*}, Qiya Song^{1*}, Hanjun Deng³

¹ College of Electrical and Information Engineering, Hunan University, Changsha, China

² National Laboratory of Pattern Recognition Institute of Automation,
Chinese Academy Sciences, Beijing, China

³ Experimental High School Affiliated to Beijing Normal University, Beijing, China
{libincn, sqyunb}@hnu.edu.cn, wengsyx@gmail.com, Hanjun_Deng@gmail.com

Abstract

As the deep learning rapidly promote, the artificial texts created by generative models are commonly used in news and social media. However, such models can be abused to generate product reviews, fake news, and even fake political content. The paper proposes a solution for the Russian Artificial Text Detection in the Dialogue shared task 2022 (RuATD 2022) to distinguish which model within the list is used to generate this text. We introduce the DeBERTa pre-trained language model with multiple training strategies for this shared task. Extensive experiments conducted on the RuATD dataset validate the effectiveness of our proposed method. Moreover, our submission ranked **second place** in the evaluation phase for RuATD 2022 (Multi-Class).

Keywords: Artificial Text Detection, Pre-trained Language Model, Multiple Training Strategies.

DOI: 10.28995/2075-7182-2022-20-375-381

Обнаружение искусственного текста с несколькими стратегиями обучения

Бин Ли

Хунаньский университет
Район Юэлу, город Чанша
libincn@hnu.edu.cn

Кийя Сон

Хунаньский университет
Район Юэлу, город Чанша
sqyunb@hnu.edu.cn

Йи Суан Вен

Китайская академия наук
Район Хайдянь, Пекин
wengsyx@gmail.com

Хан Джун Дэн

Средняя школа при
Пекинском педагогическом университете
Район Сичэн, Пекин
Hanjun_Deng@gmail.com

Аннотация

С быстрым распространением глубокого обучения все чаще тексты, созданные с помощью генеративных моделей, начинают использоваться в новостях и социальных сетях. Однако такими моделями можно злоупотреблять для создания отзывов на продукты, фейковых новостей и даже фальшивого политического контента. В этой работе предлагается решение для задачи классификации модели, использованной для генерации текста на русском языке, в рамках соревнования Dialogue RuATD-2022. Мы применили предобученную языковую модель DeBERTa с несколькими стратегиями обучения. Обширные эксперименты с данными RuATD подтверждают эффективность предложенного нами метода. Кроме того, представленные нами результаты заняли **второе место** на этапе оценки RuATD 2022 (Мультиклассовая классификация).

Ключевые слова: Обнаружение сгенерированного текста, предобученная языковая модель, несколько стратегий обучения.

1 Introduction

With the rapid development of AI technologies, a growing number of methods the ability to generate realistic artifacts. For instance, amount of texts generated by recent text generation methods the transformer

*These authors contribute equally to this work.

encoder-decoder framework are very close to the text written by humans, including lots of security issues (de Rosa and Papa, 2021; Topal et al., 2021). Extensive Transformer-based text generation models, such as Bert-style (Devlin et al., 2018a), GPT-style (Radford et al., 2019), have achieved excellent results on a large number of NLP tasks. (Keskar et al., 2019) proposed the conditional transformer language model (CTRL) with 1.63 billion parameters to control the text generation. The model is trained with different codes that control task-specific behavior, entities, specify style and content. (Zellers et al., 2019) introduced a controllable text generation model named Grover, which can overwrite propaganda papers. For example, given a headline "Discovering the link between vaccines and autism," Grover could generate a description article for this title. Humans rated this generated text as more trustworthy than human-written text. However, the success of natural language generation has drawn dual-use concerns. On the one hand, applications such as summary generation and machine translation are positive. On the other hand, related techniques may also enable adversaries to generate neural fake news, targeted propaganda and even fake political content. Therefore, several researchers have made many attempts to develop artificial text detectors (Jawahar et al., 2020). (Solaiman et al., 2019) used the pre-trained language model RoBERTa for the downstream text detection task and achieved the best performance in recognizing web pages generated by the GPT-2 model. (Kushnareva et al., 2021) proposed a novel method based on Topological Data Analysis (TDA). The interpretable topological features that can be derived from the attention map of any transformer-based language model are introduced for the task of artificial text detection. (Shamardina et al., 2022) originated two tracks on the RuATD 2022 Dialogue Shared task to solve the problem of automatic recognition of generated texts. In this paper, we adopt DeBERTa method with multiple training strategies for the Russian artificial text detection in the Dialogue shared task 2022 (Multi-Class). More details about our system are introduced in the following sections.

2 Main Method

This section will elaborate on the main method for the Russian artificial text detection dialogue shared task, where we adopt the pre-trained model with multiple training strategies, such as adversarial training, child tuning, and intrust loss function.

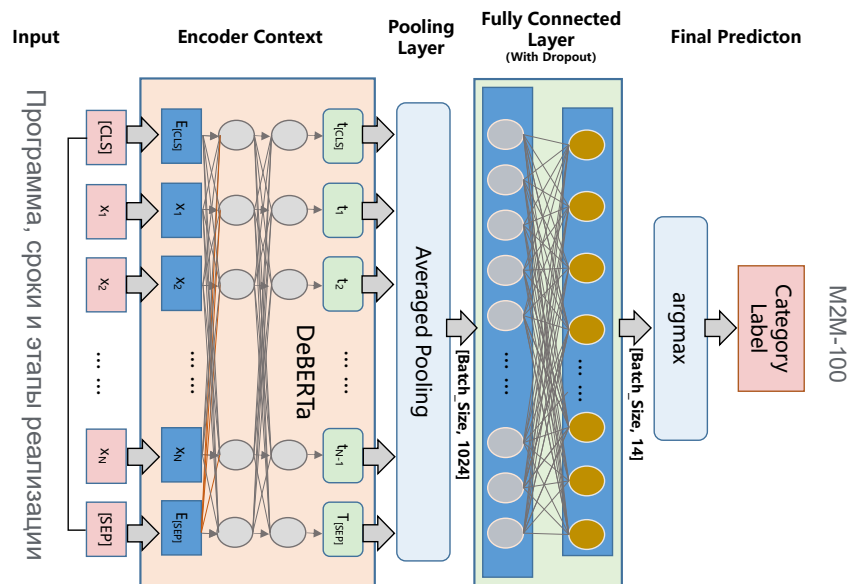


Figure 1: The model architecture of the submitted system.

2.1 Overview of Pre-trained Model

It is noted that the pre-trained model has a solid ability to differentiate the results generated by different models (Qiu et al., 2020; Naseem et al., 2021), we resort to the state-of-the-art (SOTA) pre-trained

language model for better prediction. As shown in Figure 1, we present the main model architecture for the shared task. Specifically, we adopt the DeBERTa family, i.e., mDeBERTa (He et al., 2020) and DeBERTa (He et al., 2021a), for the category classification. The pooling and fully connected layers are at the top of the pre-trained language model for leveraging global semantics. Finally, the argmax is performed after the 14 categories classification to obtain the final results.

2.2 Multiple Training Strategies

2.2.1 Adversarial Training

The common method in adversarial training is the Fast Gradient Method (Nesterov, 2013; Dong et al., 2018). The idea of FGM is straightforward. Increasing the loss is to increase the gradient so that we can take

$$\Delta x = \epsilon \nabla_x L(x, y; \theta) \quad (1)$$

where x represents the input, y represents the label, θ is the model parameter, $L(x, y; \theta)$ is the loss of a single sample, Δx is the anti-disturbance.

Of course, to prevent Δx from being too large, it is usually necessary to standardize $\nabla_x L(x, y; \theta)$. The more common way is

$$\Delta x = \epsilon \frac{\nabla_x L(x, y; \theta)}{\|\nabla_x L(x, y; \theta)\|} \quad (2)$$

2.2.2 Child-tuning Training

The efficient Child-tuning (Xu et al., 2021) method is used to fine-tuning the backbone model in our method, where the parameters of the Child network are updated with the gradients mask. For this shared task, the task-independent algorithm is used for child-tuning. When fine-tuning, the gradient masks are obtained by Bernoulli Distribution (Chen and Liu, 1997) sampling from in each step of iterative update, which is equivalent to randomly dividing a part of the network parameters when updating. The equation of the above steps is shown as follows

$$w_{t+1} = w_t - \eta \frac{\partial \mathcal{L}(w_t)}{\partial w_t} \odot B_t \quad (3)$$

$$B_t \sim \text{Bernoulli}(p_F)$$

where the notation \odot represents the dot production, p_F is the partial network parameter.

2.2.3 Ensemble Method

Once obtaining the pre-trained model, we need to maximize the advantages of each model. So, we use ensemble each model with the Bagging Algorithm (Skurichina and Duin, 2002) via voting on the predicted results of the trained models. The Bagging algorithm is used during the prediction, where this method can effectively reduce the variance of the final prediction by bridging the prediction bias of different models, enhancing the overall generalization ability of the system.

3 Experiment

We will introduce the RuATD dataset, evaluation indicators, implementation details and method description.

3.1 RuATD

However, some people may use these models with malicious intent to generate false news, automatic product reviews, and false political content. RuATD 2022 proposes a new task, which requires judging whether a sentence is generated by the model (binary classification) or even which model it is generated by (multi-class classification). More task details can be found in the website¹.

¹<https://www.kaggle.com/c/ruatd-2022-multi-task>

Model	FGM	ChildTune	In-trust loss	Backbone
DeBERTa-large	61.27	61.23	61.54	61.42
mDeBERTa-base	61.89	61.68	62.21	62.06

Table 1: The experimental results of the text detection.

Methods	Accuracy
Random sample	19.927
Tf-idf	44.280
BERT fine-tuning	59.813
Ours	64.731

Table 2: Comparison with baselines in official test set.

3.2 Evaluation

In the multi-classification task, accuracy is used as the evaluation index. The task requires the model to judge whether a sample is written by humans or generated by other generation models.

$$Acc = \left(\frac{Right}{All} \right)$$

3.3 Baseline Introduction

Tf-idf: With the help of sklearn (Pedregosa et al., 2018), the organizers connect TF-IDF, SVD, standard-scaler and logistic regression in turn for training.

BERT Fine-tuning: BERT (Devlin et al., 2018b) is a model designed for natural language understanding task. It uses MLM pre-training method and has strong semantic feature understanding ability. The organizer added a 14 category linear layer after the output layer of BERT-base, and Cross-Entropy loss is used to fit.

3.4 Implementation Details

We train the model using the Pytorch² (Paszke et al., 2019) on the NVIDIA RTX3090 GPU and use the hugging-face³ (Wolf et al., 2020) framework. For all uninitialized layers, We set the dimension of all the hidden layers in the model as 1024. The AdamW (Loshchilov and Hutter, 2018) optimizer which is a fixed version of Adam (Kingma and Ba, 2014) with weight decay, and set β_1 to 0.9, β_2 to 0.99 for the optimizer. We set the learning rate to $1e - 5$ with the warm-up (He et al., 2016). The batch size is 32. We set the maximum length of 280, and delete the excess. Linear decay of learning rate and gradient clipping is set to $1e - 4$. Dropout (Srivastava et al., 2014) of 0.1 is applied to prevent over-fitting. All experiments select the best parameters in the valid set. Finally, we report the score of the best model (valid set) in the test set.

We use the mDeBERTa-base (He et al., 2021c; He et al., 2021b) as our pretrained model, and fine-tune the model⁴. The mDeBERTa⁵ model comes with 12 layers and a hidden size of 768. And it was trained with the CC100 (Conneau et al., 2020) multilingual data .

4 Case Study

We counted and analyzed the mispredicted samples, and the distribution of error types is shown in Figures 2. We chose the top 100 samples with the most significant difference from the ground truth as the analysis

²<https://pytorch.org>

³<https://github.com/huggingface/transformers>

⁴You can reproduce the baseline code from here <https://github.com/dialogue-evaluation/RuATD/blob/main/Baseline.ipynb>

⁵[microsoft/mdeberta-v3-base](https://github.com/microsoft/mdeberta-v3-base)

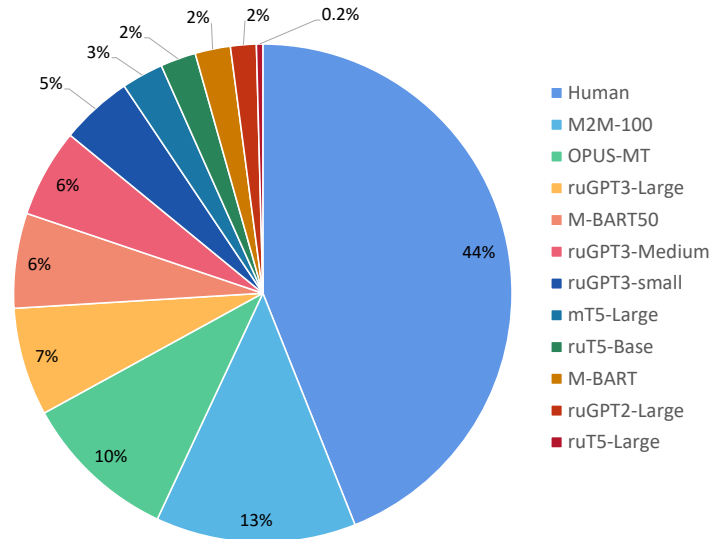


Figure 2: Case study in the shared task.

object. As we can see from Figure 2, the most mispredicted type in the classification task was “Human”, with 44%, followed by “M2M-100” with 13%, etc, the “ruT5-Large” obtains the least error with 0.2%. Further conclusions can be that sorting from high to low actually shows the capability performance of the model. The higher the error rate, the better the performance of the model, and the effects like M2M/GPT3 are better. Then the bigger the model, the harder the target is to distinguish.

5 Result and Discussion

As shown in Table 1, we implement the DeBERTa-large and mDeBERTa-large with multiple training strategies. It can be further concluded that the in trust loss method with the pre-trained model can achieve the best results in artificial text detection. It may be the reason that the model is trained through In-trust training. can be more robust. Moreover, we found that the mDeBERTa outperforms the original version, which indicates that the multi-lingual can provide differentiated knowledge for this text detection. Table 2 also presents the comparison between ours and baselines, where our method outperforms the BERT baseline by 2.397 in accuracy score on the official test set.

6 Conclusion

This paper illustrates our contributions for Russian Artificial Text Detection Dialogue Shared task (Multi Class). We use the DeBERTa pre-trained language model with multiple training strategies to distinguish which model from the list was used to generate this text. In the evaluation phase, our submission achieves **second place**.

References

- Sean X Chen and Jun S Liu. 1997. Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica*, P 875–892.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 8440–8451, Online, July. Association for Computational Linguistics.
- Gustavo H de Rosa and Joao P Papa. 2021. A survey on text generation using generative adversarial networks. *Pattern Recognition*, 119:108098.

- 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. *arXiv preprint arXiv:1810.04805*.
- Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. // *Proceedings of the IEEE conference on computer vision and pattern recognition*, P 9185–9193.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. // *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, P 770–778.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021b. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021c. Deberta: Decoding-enhanced bert with disentangled attention. // *International Conference on Learning Representations*.
- Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*.
- 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*.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning*.
- Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. *arXiv preprint arXiv:2109.04825*.
- Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. // *International Conference on Learning Representations*.
- Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35.
- Yu Nesterov. 2013. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. // H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in python.
- Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

- Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian. // *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue"*, volume 21.
- Marina Skurichina and Robert PW Duin. 2002. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2):121–135.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askeff, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
- Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.
- M Onat Topal, Anil Bas, and Imke van Heerden. 2021. Exploring transformers in natural language generation: Gpt, bert, and xlnet. *arXiv preprint arXiv:2102.08036*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.
- Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 9514–9528.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Method and Software Model for Evaluating the Statistical Characteristics of a Speech Melody

Boris Lobanov

United Institute of Informatics
Problems NAS Belarus
lobbormef@gmail.com

Vladimir Zhitko

United Institute of Informatics
Problems NAS Belarus
zhitko.vladimir@gmail.com

Abstract

A method for estimating the statistical characteristics of speech melody is proposed. The procedure of constructing histograms of the frequency distribution of the pitch frequency over sufficiently long intervals of speech is described. A distinctive feature of the method is that the discrete values of the pitch of speech are measured only at intervals of vowels. Two options of the pitch scales used for analysis of the melody characteristics are selected, namely: linear for speech and logarithmic for singing. A method for estimating three parameters of the histogram is proposed: register - R, range - D, asymmetry - A. Numerous examples are given showing the effectiveness of the proposed method in assessing the individuality of the melody of the speaker's speech, as well as his emotional state. A description of the prototype of the Speech Melody Meter (SMM) system is given (see also: <https://intontrainer.by>). SMM is a software implementation of the proposed method for assessing the statistical characteristics of speech melody.

Keywords: speech prosody, speech melody, pitch frequency, vowels detection, software model.

DOI: 10.28995/2075-7182-2022-21-382-395

Метод и программная модель оценки статистических характеристик мелодики речи

Б.М. Лобанов

Объединенный институт проблем
информатики НАН Беларуси
lobbormef@gmail.com

В.А. Житко

Объединенный институт проблем
информатики НАН Беларуси
zhitko.vladimir@gmail.com

Аннотация

Предлагается метод оценки статистических характеристик мелодики речи. Описана процедура построения гистограмм распределения частот встречаемости дискретных значений частоты основного тона (ЧОТ) на достаточно длительных интервалах речи. Отличительной особенностью метода является то, что дискретные значения ЧОТ речи измеряются только на интервалах гласных звуков. Выбраны два варианта шкал ЧОТ, используемые при анализе характеристик мелодики, а именно: линейная для речи и логарифмическая для пения. Предложен метод оценки трех параметров гистограммы: регистр - R, диапазон - D, асимметрия - A. Приводятся многочисленные примеры, показывающие эффективность предложенного метода при оценке индивидуальности мелодики речи диктора, а также его эмоционального состояния. Дано описание прототипа системы Speech Melody Meter (SMM), являющейся программной реализацией предложенного метода оценки статистических характеристик мелодики речи (см. также: <https://intontrainer.by>).

Ключевые слова: просодия речи, мелодия речи, частота основного тона, программная модель.

1 Введение

Мелодика — основной компонент интонации речи человека. С акустической точки зрения, мелодика — это изменения во времени частоты основного тона (ЧОТ), измеряемой в герцах (Гц) или в музыкальных интервалах. Мелодика речи занимает важное место в структуре языка. Она участвует в реализации двух основных функций интонации — семантической и экспрессивной.

Семантическая функция мелодики выполняет следующие достаточно хорошо изученные задачи:

- организует фразу, расчлняя её на синтагмы и акцентные группы и связывая её части;
- различает коммуникативные типы высказывания (вопрос, побуждение, повествование и другие).

Семантическая функция реализуется путем специфических локальных изменений ЧОТ на отдельных элементах фразы (слог, акцентная группа, синтагма). Описанию закономерностей этих изменений для различных языков посвящены многочисленные работы, обзор которых не входит в задачу данной статьи.

Экспрессивная функция мелодики выполняет следующие задачи:

- выражает персональные особенности интонирования речи путем формирования специфического мелодического рисунка, обуславливающего своеобразность, стиль, выразительность и нормативность речи;
- определяет субъективное отношение к содержанию сообщения или к собеседнику путем выражения различного вида эмоций.

Необходимо отметить также, что степень владения навыками управления экспрессией речи является, пожалуй, одним из важных показателей эмоционального интеллекта человека — способностью управлять своими эмоциями и эмоциями других людей.

Экспрессивная функция мелодики, в отличие от семантической, реализуется главным образом за счет специфических интегральных характеристик траектории ЧОТ на достаточно длительном временном отрезке речи. Физические свойства мелодической кривой, обуславливающие особенности экспрессивной функции в сравнении с семантической функцией, исследованы в существенно меньшей степени. В определенной мере к такого рода исследованиям можно отнести результаты работ, направленных на решение ряда практически важных задач [(M. Farru's, 2009), (M. Leena, 2008), (C. Leyns, 2010), (D. Gharavian, 2010), (Cahill, 2016), (T. Banziger, 2005), (Peppé, 2009), (A. Norton, 2009), (Celine De Looze, 2017)], таких, как оценка:

- индивидуальных особенностей речи диктора [(M. Farru's, 2009), (M. Leena, 2008), (C. Leyns, 2010);
- эмоционального состояния говорящего [(D. Gharavian, 2010), (Cahill, 2016), (T. Banziger, 2005)];
- физиологических и патологических дефектов речи [(Peppé, 2009), (A. Norton, 2009), (Celine De Looze, 2017)].

В работе (M. Farru's, 2009) индивидуальные особенности речи диктора оцениваются путем измерения «дрожания и мерцания» ЧОТ и амплитуды. В работе (M. Leena, 2008) контур F0 представлен с помощью следующих оценок: значение максимума dF0, расстояние пика F0 относительно начала, наклон и продолжительность наклона кривой ЧОТ. В работе (C. Leyns, 2010) возраст и пол диктора определялся с помощью акустического анализа четырех интонационных параметров (общий интонационный сдвиг, общий диапазон, конечный интонационный сдвиг и показатель вариации ЧОТ).

В работе (D. Gharavian, 2010) показано, что в эмоциональных предложениях средние частоты основного тона и величина этого увеличения зависит от типов групп фонем и эмоций. Работа (Cahill, 2016) посвящена анализу особенностей реализации эмоций в африканском языке Knni. В работе (T. Banziger, 2005) описываются результаты, полученные для эмоциональных выражений актера. Показано, что средний уровень и диапазон F0 на контурах сильно различаются в зависимости от степени активации изображаемой эмоции. Напротив, было сравнительно мало свидетельств появления качественно разных форм контуров для разной эмоции.

В работе (Peppé, 2009) отмечается, что для болезни Паркинсона характерна низкая высота тона, низкая скорость и малый диапазон изменений ЧОТ. Научно обоснованный метод лечения – терапия с использованием мелодической интонации (МИТ) – описывается в работе (A. Norton, 2009). Метод использует интонирование при чтении и пении для улучшения выразительности речи людей с афазией. В работе (Celine De Looze, 2017) исследовано влия-

ние когнитивных нарушений на просодические параметры речеобразования при рассеянном склерозе. Замечено, что наклон высоты тона в высказывании тем круче, чем короче высказывание.

Как следует из обзора, в цитируемых работах по анализу индивидуальных особенностей, эмоционального состояния и патологических дефектов речи используются разнородные методики и инструменты анализа ЧОТ речевого сигнала. В данной работе предпринята попытка разработки единого метода и программной модели оценки статистических характеристик мелодики речи, пригодных для решения различных прикладных задач.

Настоящая работа базируется на результатах статьи (B. Lobanov, 2021), в которой описан метод расчета статистических оценок комплекса просодических характеристик темпа речи и алгоритмы его программной реализации. В данной работе использованы те же методы предварительной обработки речевого сигнала и тот же метод детектирования гласных звуков.

2 Предварительная обработка и анализ речевого сигнала

Оценка статистических характеристик мелодики речи базируется на предварительном определении границ гласных звуков, на интервалах которых строится траектория движения ЧОТ анализируемого отрезка речи. Гласные звуки, в сравнении с согласными и паузами, характеризуются в среднестатистическом смысле значительной интенсивностью, высокой степенью периодичности и специфическим частотным спектром. Опора на гласные звуки позволяет, как хорошо известно, выделить наиболее значимые для восприятия участки траектории ЧОТ и в то же время избежать ряда артефактов, возникающих на интервалах согласных и пауз.

Кроме задачи автоматического определения границ гласных звуков, необходимо было также выбрать наиболее эффективный алгоритм анализа ЧОТ речевого сигнала. К настоящему времени разработаны и находятся в свободном доступе (Group, 2021) четыре наиболее эффективных алгоритма: RAPT, SWIPE, REAPER и WORLD. После тщательной проверки и тестирования на большом речевом материале наилучшие результаты (минимум ошибок и сбоев) показал алгоритм RAPT [(Talkin, 1995), (Talkin, 2015)], который был выбран нами для решения поставленных задач.

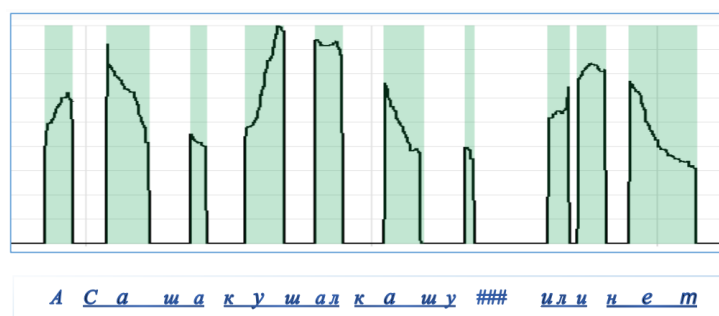


Рис. 1: Результат расчета траектория ЧОТ фразы: «А Саша кушал кашу или нет?»

На рисунке 1 приведен пример результата определения границ гласных и расчета траектории ЧОТ алгоритмом RAPT для произнесенной фразы: «А Са+ша ку+шал кашу, и+ли нет?» (знаком + указано положение словесного ударения).

3 Построение гистограмм статистического распределения ЧОТ речевого сигнала

Гистограмма в математической статистике — это распределение частот встречаемости дискретной случайной величины, Графически представляется в виде набора столбцов равной ширины. Высоты столбцов пропорциональны частоте их встречаемости, а количество столб-

цов задается требуемой точностью анализа в выбранном диапазоне измерения. Диапазон измерения ЧОТ и количество столбцов выбираются исходя из специфики решаемой задачи.

Так, при анализе статистических характеристик ЧОТ пения целесообразно воспользоваться логарифмической музыкальной шкалой частот в диапазоне не менее пяти октав. На рисунке 2 представлена клавиатура фортепиано с указанием названий октав и нот.

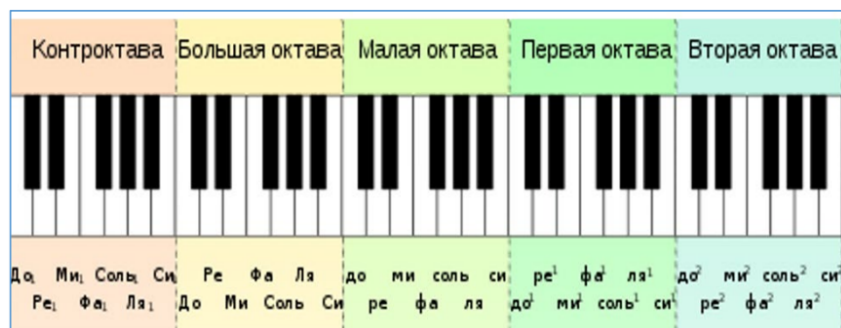


Рис. 2: Клавиатура фортепиано

Номера столбцов гистограммы, их количество и значения ЧОТ в Герцах, соответствующих клавиатуре фортепиано, представлены в таблице 1.

Контр октава	1	2	3	4	5	6	7
	32.70	36.95	41.21	43.65	49.00	55.00	61.74
Большая октава	8	9	10	11	12	13	14
	65.41	73.91	82.41	87.31	98.00	110.00	123.48
Малая октава	15	16	17	18	19	20	21
	130.82	147.83	164.81	174.62	196.00	220.00	246.96
Первая октава	22	23	24	25	26	27	28
	261.63	293.66	329.63	349.23	349.23	392.00	440.00
Вторая октава	29	30	31	32	33	34	35
	493.88	523.25	587.32	659.26	784.00	880.00	987.75
Ноты	До	Ре	Ми	Фа	Соль	Ля	Си

Таблица 1: Номера столбцов и значения ЧОТ в Гц (логарифмическая музыкальная шкала)

Формирование каждого из столбцов гистограммы осуществляется следующим образом. Значения ЧОТ измеряются на множестве гласных исследуемой выборки речевого сигнала в каждый момент дискретного времени – порядка 1 мсек. В эти моменты времени измеренная ЧОТ речевого сигнала сравнивается со всеми частотами таблицы 1. В ту из них, которая окажется ближе всего к измеренной ЧОТ, добавляется 1, и так далее, до конца анализируемого отрезка речевого сигнала. Результатом выполнения этих операций является построение гистограммы статистического распределения нот и, соответственно, значений ЧОТ.

Пример результатов расчета гистограмм, построенных с использованием пяти октавной логарифмической шкалы, представлены на рисунке 3 («Нашид» в исполнении Майкла Джексона и песня «Ямайка» в исполнении Робертино Лоретти). По горизонтальной оси отложены номера музыкальных нот, по вертикальной оси – относительные значения частоты их встречаемости. Гистограмма отображает относительное количество различных музыкальных нот, использованных певцом при исполнении данных песен.

На рисунке 4 показаны гистограммы «Здравной арии» из оперы Дж. Верди «Травиата» в исполнении известных итальянских певцов – тенора и сопрано. Очевидно нотное подобие представленных гистограмм. Их различие состоит лишь в сдвиге распределений нот на одну

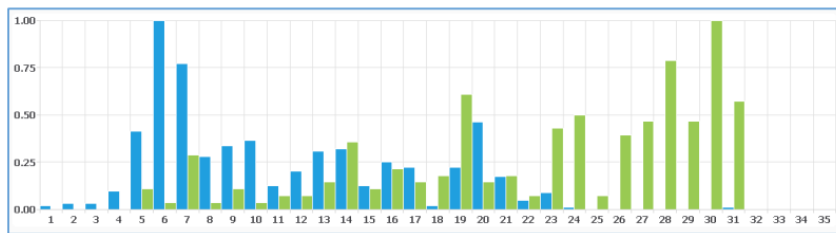


Рис. 3: Гистограммы статистических распределений нот – «Нашид» в исполнении Майкла Джексона (синие столбики) и песня «Ямайка» в исполнении Робертино Лоретти (зеленые).

октаву.

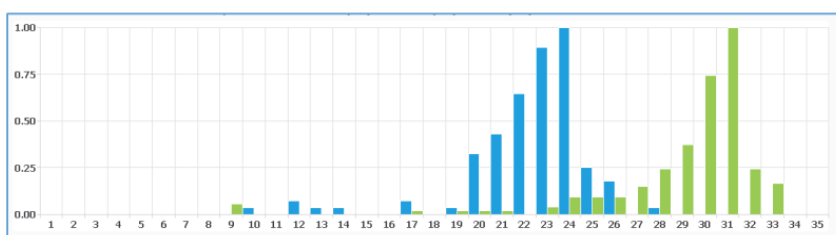


Рис. 4: Гистограмма статистического распределения нот - «Заздравная ария» в исполнении тенора (синие столбики) и сопрано (зеленые).

При анализе статистических характеристик ЧОТ речи, в отличие от пения, целесообразно воспользоваться не логарифмической, а линейной шкалой частот (Hermes and van Gestel., 1991). Рекомендуемый диапазон частот – от 60 до 400 Гц. В таблице 2 представлены номера столбцов гистограммы и значения ЧОТ в Герцах.

Номер столбца	1	2	3	4	5	6	7
Частота в Гц	60	70	80	90	100	110	120
Номер столбца	8	9	10	11	12	13	14
Частота в Гц	130	140	150	160	170	180	190
Номер столбца	15	16	17	18	19	20	21
Частота в Гц	200	210	220	230	240	250	260
Номер столбца	22	23	24	25	26	27	28
Частота в Гц	270	280	290	300	310	320	330
Номер столбца	29	30	31	32	33	34	35
Частота в Гц	340	350	360	370	380	390	400

Таблица 2: Номера столбцов и значения ЧОТ в Гц (линейная речевая шкала)

Пример результатов расчета гистограмм, построенных с использованием линейной частотной шкалы (таблица 2), представлен на рисунке 5.

На рисунке 5 по горизонтальной оси отложены номера отсчетов речевой шкалы, по вертикальной оси – относительные значения частоты встречаемости ЧОТ, соответствующей номеру столбика шкалы. Зная номер столбца, легко подсчитать соответствующее ему значение в Герцах: ЧОТ [Гц] = [(номер столбца) × 10] + 50.

Ниже будут рассмотрены многочисленные примеры гистограмм с использованием линейной шкалы (таблица 2), построенных для речевых сигналов различного текстового и

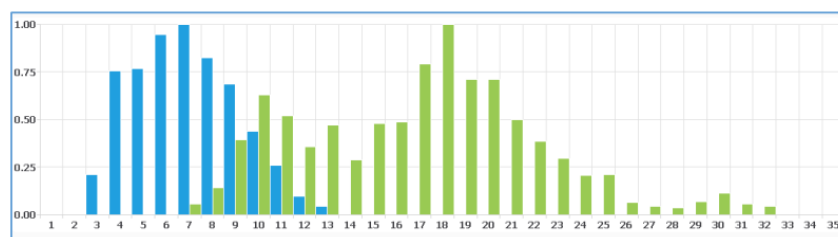


Рис. 5: Гистограмма статистических распределений ЧОТ английской речи двух профессиональных дикторов мужского пола – нейтральной (синие) и эмоциональной (зеленые).

эмоционального содержания в исполнении различных дикторов – мужчин и женщин.

4 Иллюстрация проявления экспрессивной функции мелодики речи на примерах гистограмм ЧОТ

Гистограммы ЧОТ речевого сигнала могут служить в качестве полезных признаков проявления экспрессивной функции мелодики речи различного рода: дикторской индивидуальности (гендерной и внутри-гендерной), эмоционального состояния говорящего, актерских навыков, стиля речи и др. В качестве примеров рассмотрим серию гистограммы, рассчитанных для реальных реализаций речевого сигнала (см. рисунки 6 – 12).

Рисунок 6 иллюстрирует гендерные различия. На рисунке приведены гистограммы статистических распределений ЧОТ для 2-х отрезков речевого сигнала с одинаковым текстовым содержанием (около 200 слов), произнесённых дикторами мужчиной и женщиной. Отличие гистограмм очевидно. Гистограммы различаются формой распределений, центрами концентрации, шириной диапазонов и др.

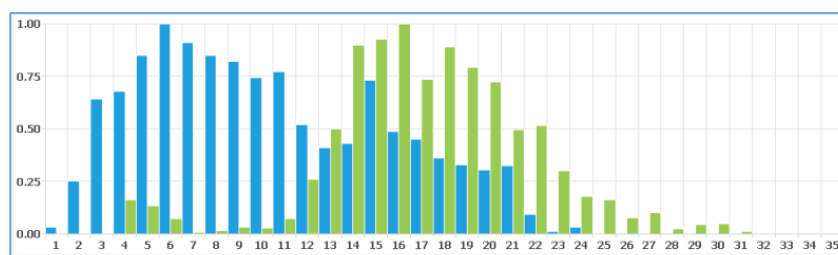


Рис. 6: Гистограммы статистических распределений ЧОТ для двух отрезков речи, произнесённых мужчиной (синие столбики) и женщиной (зеленые).

Рисунки 7 и 8 иллюстрируют наличие внутри гендерных различий. На рисунке 7 приведены гистограммы ЧОТ двух дикторов-мужчин (актеры И. Змеев и А. Канцевич), а на рисунке 8 – двух дикторов-женщин (актриса Ю. Тархова и специалист по фонетике русского языка Г. Зимовина). В их исполнении озвучен отрывок в 200 слов из рассказа А.П. Чехова «Дама с собачкой». На рисунке показаны гистограммы ЧОТ мужчин И. Змеева и А. Канцевича, а на рисунке 7 – гистограммы ЧОТ женщин - Ю.Тарховой и Г.Зимовиной. Визуальное сравнение полученных гистограмм ясно свидетельствует о наличии существенных персональных особенностей вида гистограмм ЧОТ рассмотренных пар дикторов.

Важно отметить, что визуальный «портрет» гистограмм ЧОТ характеризует манеру чтения не только отдельных фрагментов, но и на протяжении всего произведения. На рисунке 9 представлены гистограммы двух фрагментов речевого сигнала (диктор А.Канцевич - мужчина) длительностью порядка одной минуты. Один из них (помечено синим) взят из начала рассказа «Дама с собачкой», а другой – из его середины.

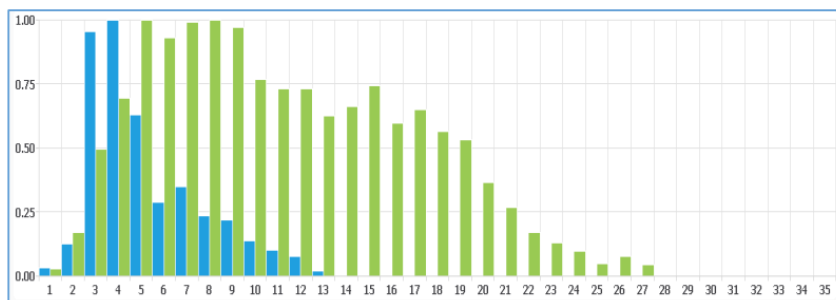


Рис. 7: Гистограммы ЧОТ мужчин: И. Змеев (синие) и А. Канцевич (зеленые)

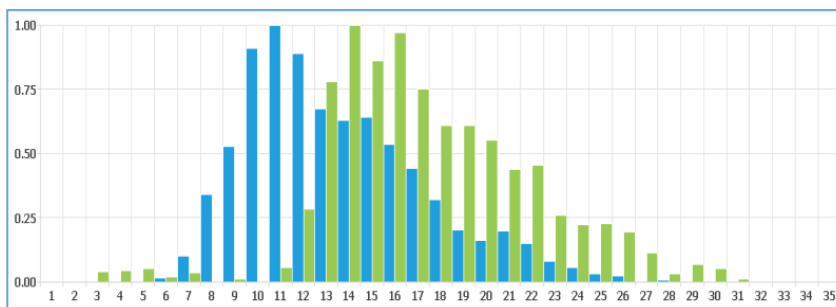


Рис. 8: Гистограммы ЧОТ женщин: Ю. Тархова (синие) и Г. Зимовина (зеленые)

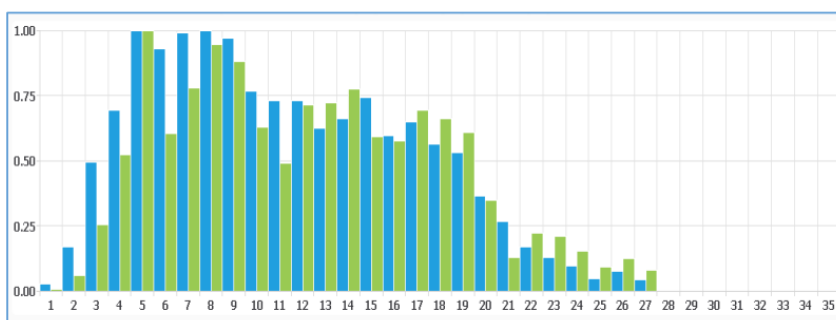


Рис. 9: Гистограммы ЧОТ мужского голоса: диктор А. Канцевич – в начале рассказа «Дама с собачкой» (синие) и в середине (зеленые)

Аналогичная картина, представленная на рисунке 10, получена для двух отрезков речевого сигнала диктора Г.Зимовиной – женщины. Один из фрагментов (помечено синим) взят из начала рассказа «Дама с собачкой», а другой – из его середины.

Интересно отметить, что вид гистограммы ЧОТ сохраняется даже при чтении диктором какого-либо другого произведения. На рисунке 10 представлены гистограммы двух фрагментов речевого сигнала длительностью порядка одной минуты. Оба фрагмента записаны в исполнении в профессионального диктора И. Змеева. Один из них (помечено синим) взят из рассказа «Дама с собачкой», а другой – из рассказа А.П. Чехова «Попрыгунья».

Представляет интерес гистограммы ЧОТ специфических стилей чтения, например, религиозных текстов. На рисунке 11 приведены гистограммы ЧОТ молитвы и проповеди.

Вид гистограмм, приведенных на рисунке 12, подтверждает известные каждому ощущение монотонности молитвы и напевности проповеди.

Любопытно увидеть также гистограммы ЧОТ речи известных исторических личностей

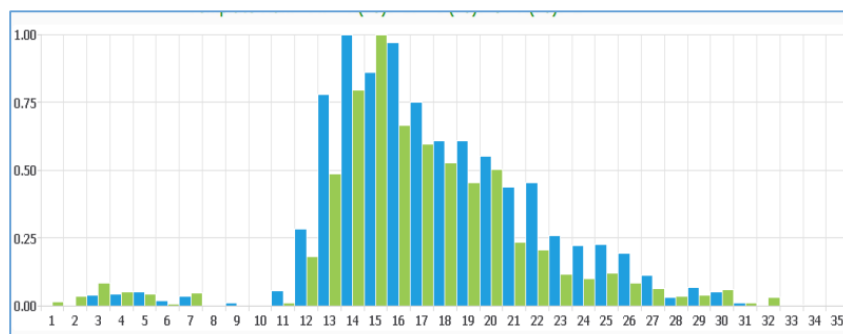


Рис. 10: Гистограммы ЧОТ женского голоса: диктор Г. Зимовина – в начале рассказа «Дама с собачкой» (синие) и в середине (зеленые)

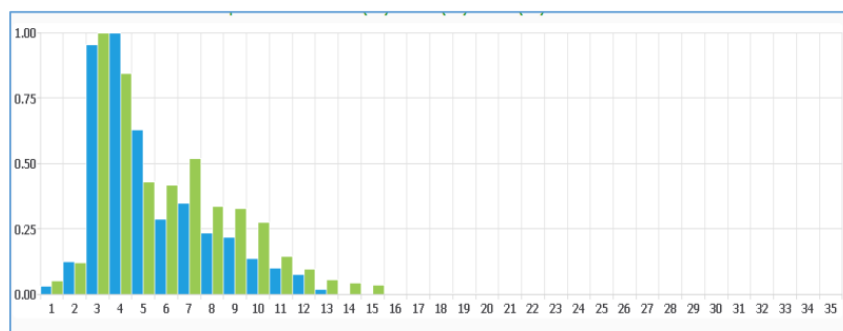


Рис. 11: Гистограммы ЧОТ речи диктора-мужчины И. Змеева – рассказы А.П. Чехова «Дама с собачкой» (синие) и «Попрыгунья» (зеленые)

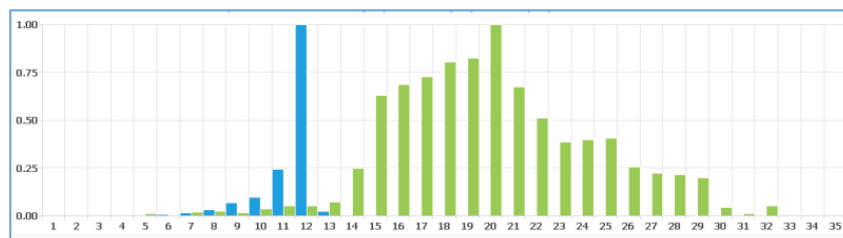


Рис. 12: Гистограммы ЧОТ молитвы (синие) и проповеди (зеленые)

(см. рисунки 13, 14), образцы речи которых приведены в ролике YouTube.com «Как звучали реальные голоса исторических деятелей в начала XX века»

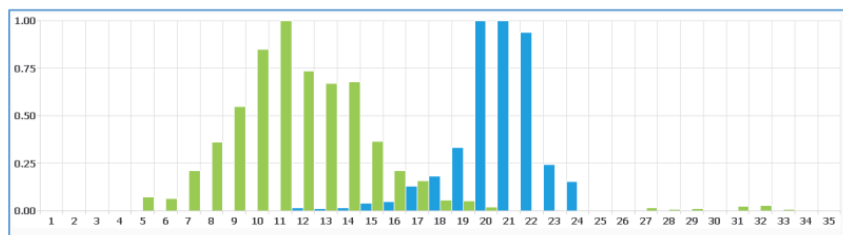


Рис. 13: Гистограммы ЧОТ выступлений: В. Ленин (синие) и И. Сталин (зеленые)

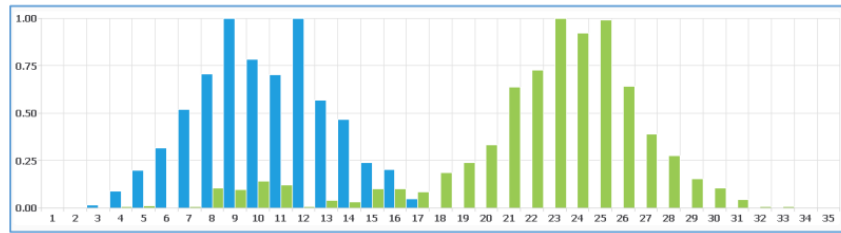


Рис. 14: Гистограммы ЧОТ выступлений: У. Черчилль (синие) и Б. Муссолини (зеленые)

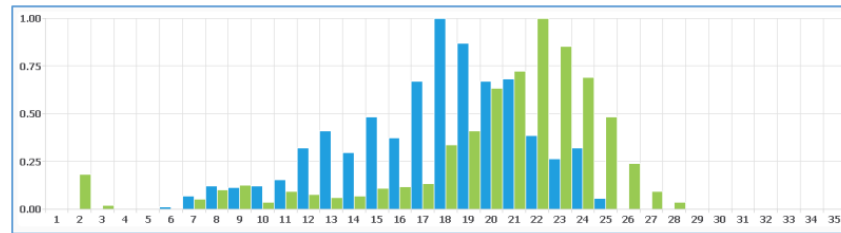


Рис. 15: Гистограммы ЧОТ выступлений: А. Вышинский (синие) и Л. Берия (зеленые)

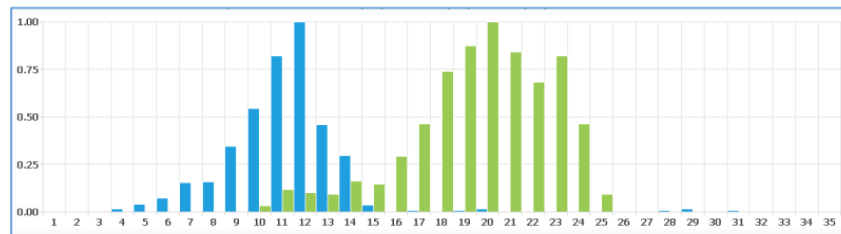


Рис. 16: Гистограммы ЧОТ выступлений: Ю. Левитан (синие) и Ю. Гагарин (зеленые)

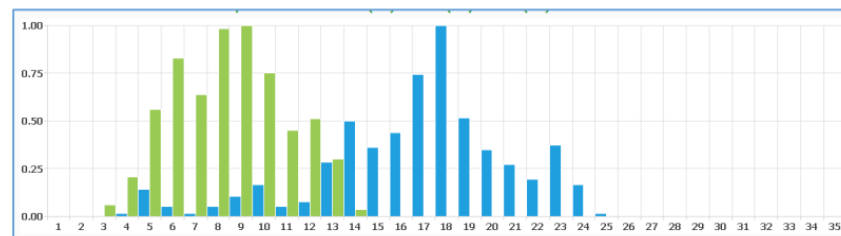


Рис. 17: Гистограммы ЧОТ выступлений: Н. Хрущев (синие) и Л. Брежнев (зеленые)

5 Оценка и сравнение статистических характеристик мелодики речи

Наиболее очевидным способом сравнения статистических характеристик мелодики речи двух дикторов является непосредственное сопоставление их гистограмм ЧОТ, вычисляя, например, векторное расстояние между ними или взаимную корреляцию. Для полимодальных распределений определенный интерес представляет также описание гистограмм коэффициентами Фурье.

В данной работе мы ограничимся рассмотрением способа оценки одномодальных распределений, характерных для большинства случаев. Наиболее простым в реализации и достаточно информативным является оценка гистограммы набором следующих 3-х параметров, предлагаемых в данной работе:

- центр концентрации распределения – Регистр – R,
- эффективный диапазон – Диапазон» - D,
- асимметрия распределения – Асимметрия – A.

Одним из наиболее простых и статистически эффективных способов определения центра гистограммы является расчет локализации центра «тяжести» гистограммы ЧОТ – $C(0)$, представленной 35-ю отсчетами: 1, 2, ... n ... 35.

Тогда: $C(0) = [\text{SUM}(n * F0n)] / [\text{SUM}(F0n)]$, (1)

Где $F0n$ – значения ЧОТ на n-м отсчете гистограммы. Рассчитанное значение $C(0)$ – условный мелодический регистр исследуемого речевого сигнала - R.

Аналогичным образом рассчитываются левый и правый центры тяжести гистограммы: $C(0-)$ и $C(0+)$. А именно: $C(0-)$ – от 1-го значения до $[C(0)-1]$, $C(0+)$ – от $[C(0)+1]$ до последнего 35-го значения.

Результатом расчёта являются: $C(0)$, $C(0-)$ и $C(0+)$ – целые числа, соответствующие номерам отсчетов гистограммы. Используя полученные значения, рассчитываются диапазон – D и асимметрии гистограммы – A:

$$D = C(0+) - C(0-). \quad (2)$$

$$A = [C(0+) + C(0-)] - 2C(0) \quad (3)$$

Отметим, что в результате расчета асимметрия A может принимать положительное значение, если в гистограмме преобладают высокочастотные компоненты и отрицательное – если низкочастотные.

Таким образом, используя формулы (1) – (3), получаем искомые оценки статистических характеристик мелодики речи дикторов.

На рисунке 18 представлены гистограммы ЧОТ записанных с помощью диктофона выступлений по ТВ В. Зеленского и А. Лукашенко, а на рисунке 19 – А. Невзорова и В. Жириновского, используемых в качестве примеров результатов расчета статистических характеристик мелодики речи.

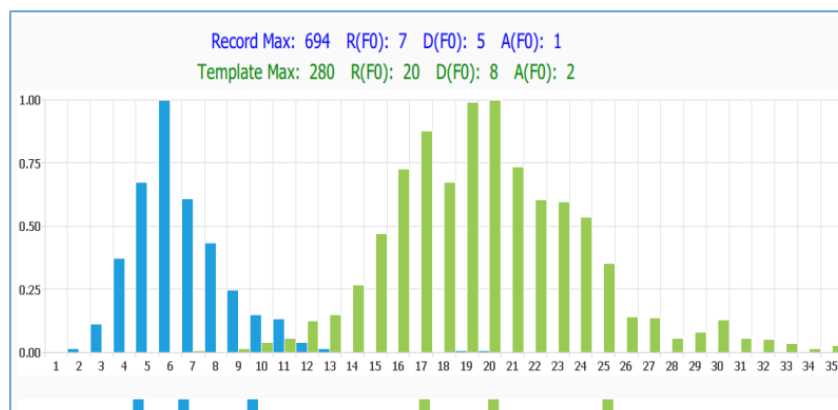


Рис. 18: Гистограммы ЧОТ выступления по ТВ В. Зеленского (синие) и А. Лукашенко (зеленые). Вверху рисунка показаны рассчитанные значения R, D, A. Внизу рисунка показаны метки позиций R, D, A на гистограммах

Полученные результаты расчётов R, D, A для указанных выше 4-х дикторов графически отображены на диаграмме (рисунок 20).

Рисунок 20 наглядно подтверждает достаточно высокую эффективность использования предложенных признаков в качестве статистических оценок персональных дикторских различий в мелодике речи.

Предложенный набор признаков может быть с успехом использован также в качестве статистических оценок эмоциональных различий в мелодике речи дикторов. На рисунках 21 и 22 представлены гистограммы ЧОТ выступлений на ТВ Ольги Скабеевой и Владимира

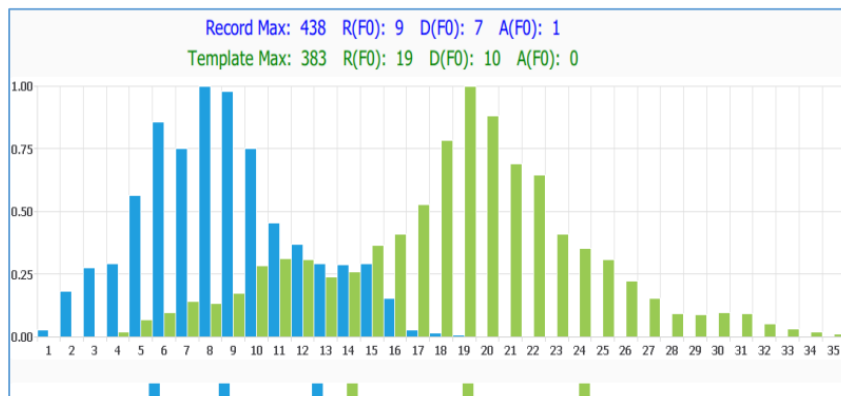


Рис. 19: Гистограммы ЧОТ выступления по ТВ А. Невзорова (синие) и В. Жириновского (зеленые). Вверху рисунка показаны рассчитанные значения R, D, A.

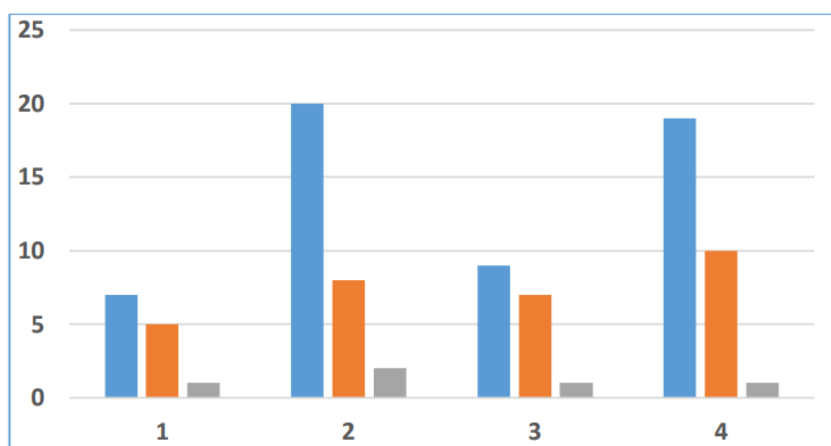


Рис. 20: Диаграмма расчетных значений R (ряд 1), D (ряд 2), A (ряд 3). Выступления по ТВ: 1 – В. Зеленский, 2 – А. Лукашенко, 3 – А. Невзоров, 4 – В. Жириновский

Соловьева в двух профессионально выраженных эмоциональных состояниях: нейтральном и в ярко-выраженном негативном.

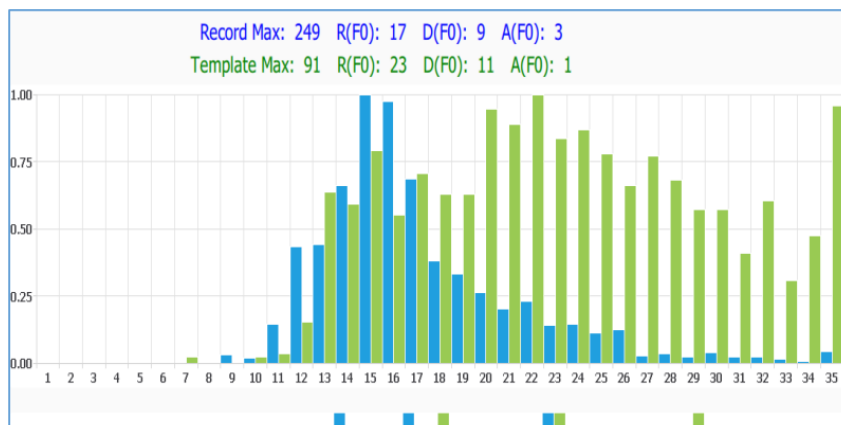


Рис. 21: Гистограммы ЧОТ выступления по ТВ Ольги Скабеевой в нейтральном эмоциональном состоянии (синие) и в ярко-выраженном негативном (зеленые)

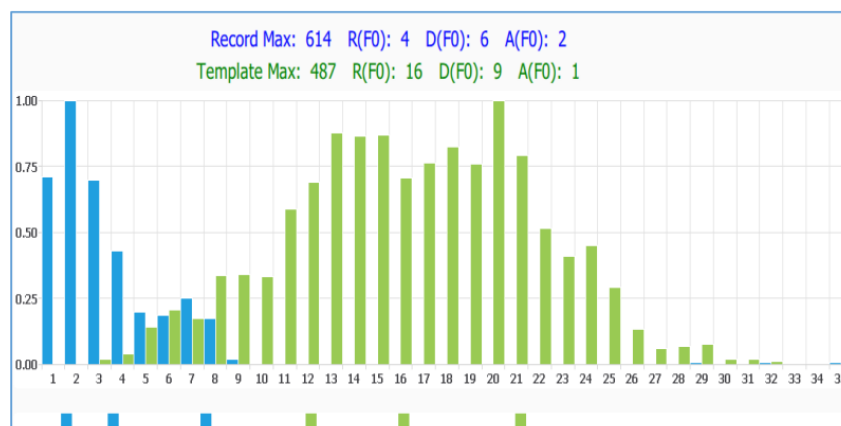


Рис. 22: Гистограммы ЧОТ выступления по ТВ Владимира Соловьева в нейтральном состоянии (синие) и в ярко-выраженном негативном (зеленые).

Полученные результаты расчётов R, D, A для указанных дикторов и их эмоционального состояния графически отображены на диаграмме (рисунок 23).

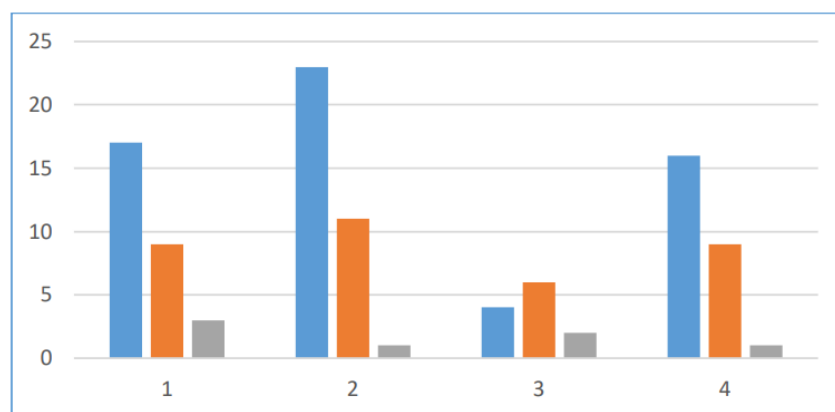


Рис. 23: Диаграмма расчетных значений R (ряд 1), D (ряд 2), A (ряд 3): 1 – О.С. (нейтр.), 2 – О.С. (негат.), 3 – В.С. (нейтр.), 4 – В.С. (негат.)

Рисунок 23 наглядно подтверждает достаточно высокую эффективность использования предложенных признаков в качестве статистических оценок эмоциональных дикторских различий в мелодике речи.

6 Программная реализация предложенного метода

Приведенные в разделах 2 – 5 результаты получены с использованием разработанного прототипа системы Speech Melody Meter (SMM), являющейся программной реализацией рассмотренного метода оценки статистических характеристик мелодики речи.

В основу программной реализации метода положен тот же подход, который был использован при разработке системы Speech Rate Meter (SRM) для статистической оценки просодических параметров темпа речи (В. Lobanov, 2021).

Прототип системы работает следующим образом. Сразу после запуска программы пользователю предоставляются следующие возможности (см. рисунок 24):

- Запись через микрофон достаточно длительного отрезка анализируемой речи (желательно не менее 30 с);
- Вызов пары заранее подготовленных тестовых аудиофайлов, хранящихся в папке “data-tests”.

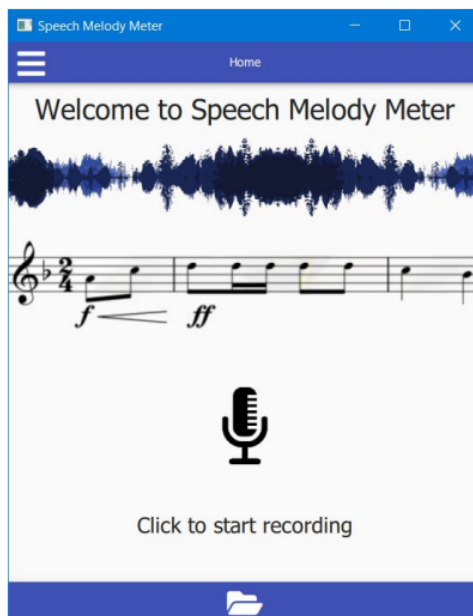


Рис. 24: Стартовое окно интерфейса программы

После того как анализируемый речевой сигнал введен с микрофона или из папки “data-tests”, осуществляется построение гистограммы статистического распределения значений ЧОТ речевого сигнала и расчет статистические характеристики ЧОТ: R, D, A с экранным отображением результатов расчета (см. рисунки 18, 19 и 21, 22).

Алгоритмы расчёта и обработки данных написаны на языке “C” с использованием библиотек GNU Scientific Library, Speech Signal Processing Toolkit (SPTK) и OpenAL. Разработан прототип системы для OS Windows – «Speech Melody Meter (SMM)». Программный модуль будет размещен на сайте <https://intontrainer.by> и общедоступен для бесплатной загрузки и тестирования. Планируется также разработка мобильной версии системы SMM в OS Android.

7 Заключение

- Предложен метод оценки статистических характеристик мелодики речи и описана процедура построения гистограмм распределения частот встречаемости дискретных значений частоты основного тона (ЧОТ) на достаточно длительных интервалах речи.
- Отличительной особенностью предложенного метода является то, что дискретные значения ЧОТ речи измеряются только на интервалах гласных звуков исследуемого речевого фрагмента.
- Выбраны два варианта шкал ЧОТ, используемые при анализе характеристик мелодики, а именно: линейная шкала для анализа речи и логарифмическая (музыкальная) – для пения.
- Предложен метод оценки трех параметров гистограммы ЧОТ: регистр – R, диапазон – D, асимметрия – A и показана их достаточно высокая различительная способность.
- Приведены многочисленные примеры, показывающие эффективность предложенного метода при оценке индивидуальности мелодики речи диктора, а также его эмоционального состояния.
- Дано краткое описание прототипа системы Speech Melody Meter (SMM), являющейся программной реализацией предложенного метода оценки статистических характери-

стик мелодики речи (см. также: <https://intontrainer.by>).

Программная реализация разработанного метода может быть полезной в следующих практически важных ситуациях:

- Для индивидуального тренинга при подготовке выступлений и презентаций для самоконтроля и обучения желаемым характеристикам мелодики речи;
- В качестве средства самоконтроля для устранения дефектов речи, связанных с различными физиологическими нарушениями и болезнями;
- Как средство мониторинга мелодики речи в реальных условиях работы операторов колл-центра, дикторов радио и телевидения.
- Определенный интерес представляет использование разработанной программы SMM для непрерывного самоконтроля особенностей мелодики речи – одного из важных показателей эмоционального интеллекта человека.

В заключение заметим, что в данной работе предприняты лишь первые шаги в направлении численной оценки экспрессивной функции мелодики речи. Дальнейшее развитие данного направления может быть связано как с использованием более тонких методов статистического анализа траекторий ЧОТ, так и с более тонкими информативными признаками гистограммам, например, описание коэффициентами ряда Фурье.

References

- S. Marchina G. Schlaug A. Norton, L. Zipse. 2009. *Melodic intonation therapy: Shared insights on how it is done and why it might help*. Annals of the New York Academy of Sciences.
- V. Zhitko B. Lobanov. 2021. *Method for Statistical Estimation of the Prosodic Parameters of Speech Tempo*.
- et al C. Leyns. 2010. *Age and gender differences in Belgian Dutch intonation*, volume 4, No. 1. Majlesi Journal of Electrical Engineering.
- M. Cahill. 2016. *Intonation and emotions in Knni: A preliminary study*. In Doris L. Payne, Sara Pacchiarotti Mokaya Bosire (eds.), *Diversity in African languages*, Berlin. Language Science Press.
- et al. Celine De Looze. 2017. *Effects of cognitive impairment on prosodic parameters of speech production planning in multiple sclerosis*. Journal of Neuropsychology. The British Psychological Society.
- M. Janipour D. Gharavian, M. Sheikhan. 2010. *Pitch in Emotional Speech and Emotional Speech Recognition Using Pitch Frequency*, volume 4, No. 1. Majlesi Journal of Electrical Engineering.
- SPTK Working Group. 2021. *pitch*.
- Dik J. Hermes and Joost C. van Gestel. 1991. *The frequency scale of speech intonation*, volume 90, No. 1. The Journal of the Acoustical Society of America.
- J. Hernando M. Farru´s. 2009. *Using Jitter and Shimmer in speaker verification*, volume 3, Iss. 4. IET Signal Process, Englewood Cliffs, NJ.
- B. Yegnanarayana M. Leena. 2008. *Extraction and representation of prosodic features for language and speaker recognition*, volume 50. Speech Communication.
- S. Peppé. 2009. *Why is prosody in speech-language pathology so difficult?*, volume 11 (4). International Journal of Speech-Language Pathology.
- K. Scherer T. Banziger. 2005. *The role of intonation in emotional expressions*, volume 46. Speech Communication.
- D. Talkin. 1995. *A robust algorithm for pitch tracking*. Speech Coding and Synthesis.
- D. Talkin. 2015. *REAPER: Robust epoch and pitch estimator*.

8 ПРИЛОЖЕНИЕ

Аудио файлы примеров, использованных при расчете приведенных в докладе гистограмм. См: <https://intontrainer.by/papers/PaperAudioAssets.zip>

DIALOG-22 RuATD Generated Text Detection

Narek Maloyan
Lomonosov MSU
maloyan.narek@gmail.com

Bulat Nutfullin
Lomonosov MSU
bulat15g@gmail.com

Eugene Ilyshin
Lomonosov MSU
john.ilyushin@gmail.com

Abstract

Text Generation Models (TGMs) succeed in creating text that matches human language style reasonably well. Detectors that can distinguish between TGM-generated text and human-written ones play an important role in preventing abuse of TGM. In this paper, we describe our pipeline for the two DIALOG-22 RuATD tasks: detecting generated text (binary task) and classification of which model was used to generate text (multiclass task) (Shamardina et al., 2022). We achieved 1st place on the binary classification task with an accuracy score of 0.82995 on the private test set and 4th place on the multiclass classification task with an accuracy score of 0.62856 on the private test set. We proposed an ensemble method of different pre-trained models based on the attention mechanism¹.

Keywords: Generated text, text classification, ensemble methods, multi-class classification task

DOI: 10.28995/2075-7182-2022-21-394-401

Аннотация

Модели генерации текста успешно синтезируют текст, который сложно отличить от написанного человеком текста. Детекторы, способные отличить текст, созданный автоматически, от написанного человеком текста позволяют предотвратить злоупотребление сгенерированными текстами. В этой статье мы описываем наше решение для задач DIALOG-22 RuATD по обнаружению сгенерированного текста и классификации с помощью какой модели был сгенерирован текст. Мы заняли 1-е место в задаче бинарной классификации с оценкой точности 0,829 в частном наборе тестов и 4-е место в задаче мультиклассовой классификации с оценкой точности 0,628 в тестовом наборе. Наше решение является ансамблем дообученных моделей, основанных на механизме внимания.

Ключевые слова: Сгенерированные тексты, классификация текстов, ансамблевые методы, задача многоклассовой классификации

1 Introduction

As the language neural nets got better at generating texts, it's getting harder and harder to distinguish the human-written text from generated one. So manual detection of these texts got impossible. For that reason, it's desirable to build a system that can automatically detect generated text.

The proposed system will use various features extracted from the text such as length, punctuation, word choice etc. To determine whether the text is human-generated or not. The accuracy of this system can be improved by using machine learning algorithms which will learn how humans generate texts and then use those features for detection purpose.

There are many ways to build such a system, but probably the most reliable one is based on machine learning algorithms. These algorithms can be trained on a large number of examples - both human-generated and computer-generated texts. After being trained, they should be able to identify which texts are computer-generated with high accuracy.

This approach already works in other areas, such as spam detection. Some early experiments have shown promising results and indicate that this approach works well for the detection of the generated text.

The DIALOG-22 shared task of RuATD 2022 had 2 tracks for binary classification and multiclass classification. In this report, we will describe the data, how we handled it, the models we used, and the ensembling technique.

¹github.com/maloyan/ruatd

2 Task Definition

For the binary classification task F_{binary} , we frame the generated text detection problem as follows: given a piece of text X , label it as either human-written or machine-generated $y_{binary} = \{H, M\}$.

$$F_{binary} : X \rightarrow y_{binary}$$

For the multiclass classification task $F_{multiclass}$, we set up the problem as follows: given a piece of text X , label it as one of the 14 classes that represent deep neural models

$$y_{multiclass} = \{M2M-100, Human, OPUS-MT, M-BART50, ruGPT3-Medium, ruGPT3-Small, mT5-Large, ruGPT3-Large, ruT5-Base-Multitask, mT5-Small, ruT5-Base, ruGPT2-Large, M-BART, ruT5-Large\}$$

$$F_{multiclass} : X \rightarrow y_{multiclass}$$

3 Datasets

Provided datasets offer the train and test splits. The part of the set was annotated automatically by different generative models. Various language models were fine-tuned on different tasks: machine translation, paraphrasing, summarization, simplification, and unconditional text generation - are used to generate texts. The texts written by a human were collected from open sources from different domains. (rua, a), (rua, b).

Text	Class
Обустройство тротуаров, мостовых (в том числе тротуарной плиткой).	Н
Минстрой обозначил способы снижения энергоемкости российской экономики.	М
В конце 1873 года военный суд вынес решение по делу Франциска Ахиллы Базейн.	М
увеличение правовой грамотности и развитие правосознания граждан.	Н

Table 1: Example of the binary classification data.

Text	Class
Прочла автобиографию Каутского, Одесса, 1905.	Human
Вы не можете быть в печи и в мушкетере.	M-BART50
Вторая попытка привела к тому же результату.	OPUS-MT
Сколько учеников в вашем классе?	M2M-100

Table 2: Example of the multiclass classification data.

4 Related Works

In this section, we will discuss various methods for detecting machine-generated texts.

Over the past years, many approaches appeared for generated text detection. The latest works are usually based on using transformer-based models. Either fine-tuning the proposed task or using probabilities distribution to make decision-based on them (Ippolito et al., 2019). Here we list some examples of different methods:

- First, we calculate the mean likelihood over all machine-generated sequences, then the mean of human-written ones. If the likelihood according to some language model is closer to the machine-generated mean likelihood, then we consider it as generated text (Solaiman et al., 2019);
- In GLTR (Gehrmann et al., 2019) described a method using a language model to compute the probability distribution of the next word given the text sequence. For each sequence position, we get the likelihood of the ground-truth next word within this list. Then these ranks are displayed on a histogram. Based on the distribution of bins, we can decide if this text is generated or not.

- Bert fine-tuning on the classification task. Having a label of text if it's machine-generated or not, we can fine-tune the language model to predict (Solaiman et al., 2019)
- Also, there is possible to use human-machine collaboration. Real or Fake Tool provides a game-like interface for humans to decide at what point the text begins to be written by a computer (Dugan et al., 2020).

More approaches are described in the survey (Jawahar et al., 2020).

5 Experiment Setup

In this section, we present the experiment configurations we use to solve binary and multiclass tasks.

5.1 Data preprocessing

We decided not to perform any preprocessing of the text itself. Regarding the data split:

- **Binary classification.** We concatenated the train set with the validation set. This new concatenated dataset gave us an option to perform a 5-fold cross-validation;
- **Multiclass classification.** We took the data as is without changing anything.

5.2 Models

All models we used for fine-tuning we got from transformers library (Wolf et al., 2020). Here we will describe these models:

- **sberbank-ai/sbert_large_nlu_ru**(hab, 2020), **sberbank-ai/ruBert-large**, **DeepPavlov/rubert-base-cased** are BERT models (Devlin et al., 2018). Sber-ai models were fine-tuned on a closed dataset collected by their research group (sbe,). DeepPavlov's RuBERT was fine-tuned on the Russian part of Wikipedia and news data.
- **IlyaGusev/mbart_ru_sum_gazeta** is a mBART model fine-tuned on the dataset for summarization of Russian news(Gusev, 2020). The original MBart model was pretrained on large-scale monolingual corpora in many languages using the BART objective. mBART is one of the first methods for pre-training a complete sequence-to-sequence model by denoising full texts in multiple languages, while previous approaches have focused only on the encoder, decoder, or reconstructing parts of the text (Liu et al., 2020).
- **MoritzLaurer/mDeBERTa-v3-base-mnli-xnli** (He et al., 2021) This multilingual model is suitable for multilingual zero-shot classification. The original model was pre-trained by Microsoft on the CC100 multilingual dataset (Wenzek et al., 2020). This model was fine-tuned on the XNLI development dataset and the MNLI train dataset. The XNLI development set consists of translated texts for each of the 15 languages (Conneau et al., 2018).
- **DeepPavlov/xlm-roberta-large-en-ru-mnli** is an XLM-RoBERTa model (Conneau et al., 2019) which was fine-tuned on the english-russian part of the MNLI (Williams et al., 2017) dataset.

5.3 Binary classification (with ensembling technique)

Five chosen models were used in the experiment: **sberbank-ai/sbert_large_nlu_ru**, **sberbank-ai/ruBert-large**, **IlyaGusev/mbart_ru_sum_gazeta** , **MoritzLaurer/mDeBERTa-v3-base-mnli-xnli**, **DeepPavlov/xlm-roberta-large-en-ru-mnli**

Let's describe steps for training each of these models:

1. We split our training dataset into non-overlapping 5-folds and performed cross-validation (Figure 1);
2. for each validation fold we predicted the target and as a result, we got out-of-fold predictions for training data;
3. for the test set, we predicted the target from every 5 models from cross-validation and averaged the result which is marked with red;

As a meta-model, we choose Logistic Regression, which was trained on out-of-fold predictions. Then we predict the final results for the test set.

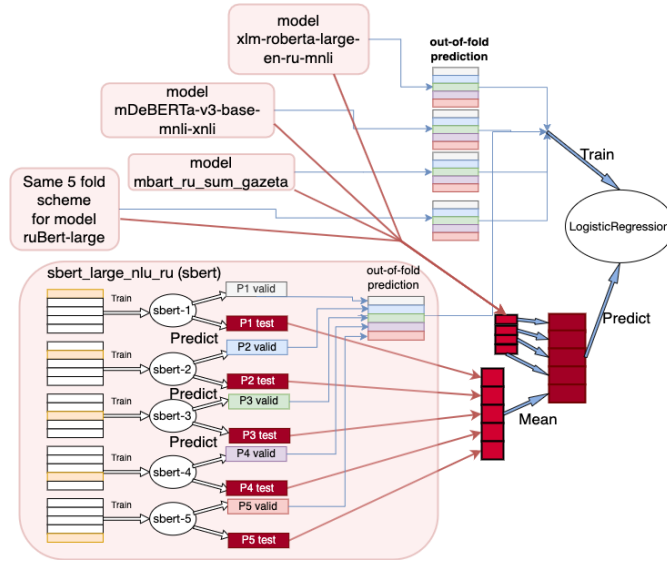


Figure 1: Ensembling scheme

5.4 Multiclass classification

For the multiclass classification we chose these models: **DeepPavlov/rubert-base-cased**, **DeepPavlov/xlm-roberta-large-en-ru**, **IlyaGusev/mbart_ru_sum_gazeta**. We fine-tune these models without cross-validation only on provided train set.

6 Results

Table 3 shows each model accuracy score on binary classification task. The multiclass classification accuracy is shown in Table 4. On both tasks, the best performing single model was **DeepPavlov/xlm-roberta-large-en-ru-mnli**. We managed to ensemble models in the binary task, so the ensemble of models showed the best accuracy.

Model name	Accuracy
sberbank-ai/sbert_large_nlu_ru	0.79986 ±0.003
sberbank-ai/ruBert-large	0.80154 ±0.002
IlyaGusev/mbart_ru_sum_gazeta	0.80566 ±0.001
MoritzLaurer/mDeBERTa-v3-base-mnli-xnli	0.80710 ±0.001
DeepPavlov/xlm-roberta-large-en-ru-mnli	0.81708 ± 0.002
Ensemble	0.82995²

Table 3: Results for binary classification for different models

Model name	Accuracy(validation)	Accuracy(kaggle public/private)
IlyaGusev/mbart_ru_sum_gazeta	0.6142	0.61459/0.61092
DeepPavlov/rubert-base-cased.	0.6045	0.60433/0.60472
DeepPavlov/xlm-roberta-large-en-ru-mnli.	0.6242	0.62856/0.62644

Table 4: Results for multiclass classification for different models

7 Conclusion

In this paper, we described our pipeline for the DIALOG-22 RuATD challenge. Our solution achieved 1st place in binary classification using ensembling techniques and 4th place for the multiclass classification task using only a single model. However, the proposed solution requires a lot of computational power, so it cannot be used in real-time systems to detect generated texts. But it gives us an understanding that we still need to upgrade methods to distinguish generated texts from human-written ones.

References

- Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.
- 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. *arXiv preprint arXiv:1911.02116*.
- 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*.
- Liam Dugan, Daphne Ippolito, Arun Kirubakaran, and Chris Callison-Burch. 2020. Roft: A tool for evaluating human detection of machine-generated text. *arXiv preprint arXiv:2010.03070*.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.
- Ilya Gusev. 2020. Dataset for automatic summarization of russian news. // *Conference on Artificial Intelligence and Natural Language*, P 122–134. Springer.
2020. Naturlal language model learning with bert and tensorflow link:<https://habr.com/ru/company/sberdevices/blog/527576/>, Nov.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Deberv3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*.
- Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*.
- 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.
- Ruatd 2022 (binary task). <https://www.kaggle.com/competitions/ruatd-2022-bi>. Accessed: 2022-03-28.
- Ruatd 2022 (multi task). <https://www.kaggle.com/competitions/ruatd-2022-multi-task>. Accessed: 2022-03-28.
- Sberbank-ai. <https://huggingface.co/sberbank-ai>.
- Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian. // *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue"*, volume 21, P xxx–xxx.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. // *Proceedings of the 12th Language Resources and Evaluation Conference*, P 4003–4012, Marseille, France, May. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

Russian neural morphological tagging: do not merge tagsets

Movsesyan A. A.

Institute for Information Transmission Problems (Kharkevich Institute)

Russian Academy of Sciences, Moscow, Russia

derise@iitp.ru

Abstract

There are multiple morphologically annotated corpora of Russian available. They have different tagsets and annotation guidelines, which makes them difficult to use together. We proposed a neural morphological tagger for Russian based on multitask learning technique which is able to predict morphological tags of words for different tagsets. We evaluated our model on various corpora and showed that utilising multiple corpora without merging them not only improves tagging performance but allows for scalable indirect conversion between multiple tagsets in all directions. Furthermore, we also showed that treating each corpus separately is more efficient than merging the corpora even if they share the same tagset.

Keywords: morphological tagging, tagset conversion, multitask learning

DOI: 10.28995/2075-7182-2022-21-402-411

Морфологический анализ русского языка на основе нейронных сетей: не объединяйте морфологические стандарты

Мовсесян А. А.

Институт проблем передачи информации РАН им. А. А. Харкевича

Москва, Россия

derise@iitp.ru

Аннотация

Для русского языка доступно множество аннотированных корпусов, снабженных морфологической разметкой. Различия между их морфологическими стандартами и схемами аннотации усложняют их совместное использование. Мы разработали модель морфологического анализатора для русского языка на основе нейронных сетей и многозадачного обучения. Модель позволяет снабжать слова морфологической разметкой для разных морфологических стандартов. Для оценки качества мы использовали ряд корпусов и показали, что использование нескольких корпусов без их слияния не только улучшает качество разметки, но и позволяет косвенно использовать модель для конвертации между несколькими морфологическими стандартами во все стороны, причем модель легко масштабируется на большее число стандартов. Кроме того, мы также показали, что использование каждого корпуса как отдельную единицу более эффективно, чем слияние корпусов, даже тогда, когда корпуса имеют общий морфологический стандарт.

Ключевые слова: морфологический анализ, морфологический стандарт, конвертация, многозадачное обучение

1 Introduction

Morphologically annotated corpora are valuable sources of data for linguistic research and natural language processing (NLP) tasks like morphological tagging and parsing. Such a corpus provides each word with a set of values of morphological categories¹ such as part-of-speech (POS), case or gender.

In the case of the Russian language, many corpora with morphological annotation exist. However, each corpus often has its own unique tagset (Hana and Feldman, 2010; Sharoff et al., 2008, to name a

¹Throughout the paper we will refer to each unique set of morphological features assigned to a word as a morphological tag.

few) and converting between them without mistakes and information loss is a challenging task. One clear example is morphological analysis contest MorphoRuEval-2017 (Sorokin et al., 2017). The organisers provided four different annotated corpora and automatically converted morphological tags to the Universal Dependencies (UD) v2.0 format (Nivre et al., 2020). But most participants ended up using only one dataset because adding others did not improve the performance of their models, especially models based on deep learning methods.

From a linguistic perspective, merging different corpora allows linguists to widen their research scope. From a statistical perspective, including machine learning and deep learning, more data would allow better performance of morphological processing tasks because it helps with the data sparseness problem.

Tagset conversion is challenging for multiple reasons:

1. Lack of parallel data. Russian corpora have little common texts, which makes it hard to create conversion rules, since each word’s tag depends on context. Training a supervised conversion model is also not possible under these circumstances.
2. Inter-annotator agreement. Even if two corpora share the same tagset, they might follow different annotation guidelines because some language phenomena are debatable. These differences might be crucial in terms of performance for the neural taggers. This problem to a lesser extent occurs within a single corpus when different annotators make different decisions because of the flaws in the guideline (Plank et al., 2014). Another challenge occurs when inter-annotator agreement score is high but all annotators make the same error in some cases (Bočarov et al., 2013).
3. Lack of annotated data. Some corpora are small and not representative enough to make plausible conversion results without the use of additional resources.

There are many approaches to this problem. We can divide them into two groups: direct and indirect. Direct approaches are mainly rule-based: for a given word in a source corpus, there is a rule to convert its tag to the target corpus format based on the word’s context (including annotation). Although some automated tools exist to provide multi-corpora tagset conversion², it is hard to cover all possible patterns using rules, and it requires manual correction, which is time-consuming. For example, in the process of converting syntactically tagged Russian text corpus SynTagRus (Inšakova et al., 2019) to the UD format (Droganova and Zeman, 2016) some sentences were omitted due to differences in the guidelines. Somewhat similar is the task of providing a unified tagset from a number of corpora’s tagsets for comparison purposes (standardisation). Such tagsets usually lack some morphological features because of conversion difficulties (Ljaševskaja et al., 2010; Lyashevskaya et al., 2017).

Indirect approaches are usually based on statistical morphological taggers. Such taggers, trained on the target corpus, intrinsically utilise source corpus annotation. These approaches are applicable to both tasks: morphological tagging and tagset conversion. One such approach (and some variations) aimed at tagset conversion trains a tagger to produce the so-called bundled tags (Li et al., 2015). Let T^s and T^t be the set of all possible tags in source and target corpus, respectively. Then the set of all bundled tags is a Cartesian product $T^s \times T^t$. During training, instead of predicting a correct label $t_i^t \in T^t$ the model predicts all labels in the set $\{t_i^t\} \times T^s$ thus making the labels ambiguous. That allows to predict labels from both tagsets at the same time. The authors tested the approach by training a POS tagger on two Chinese corpora. This approach is practically inapplicable to Russian because there are hundreds and thousands of different morphological tags possible in a given corpus compared to a few dozens of POS tags in Chinese, which keeps the Cartesian product small.

As for the Russian language, there is one indirect approach in the literature to our knowledge, and it is based on transfer learning technique (Andrianov and Mayorov, 2017). Namely, the authors trained multiple neural taggers (one tagger per source corpus in the case of multiple source corpora) and used their intermediate layers’ outputs as inputs to the main tagger trained on the target corpus.

All those indirect approaches have one essential drawback: scalability. We often need to be able to make the conversion in both directions, and the mentioned approaches are not easy to apply when the number of the target corpora is more than one.

The primary objective of this paper is to show how unrelated Russian morphological corpora can

²See, for example, <https://pypi.org/project/russian-tagsets/>

benefit each other on the morphological tagging task in a scalable manner. We train a neural morphological tagger in a multitask learning setting, treating each corpus' annotation separately but sharing the intermediate text representation. We do not use pretrained word embeddings or any other external data besides the corpora. We evaluate our model on a set of Russian corpora and also on the data provided in the MorphoRuEval-2017 contest for comparison. We show that utilising multiple corpora in a multitask setting improves tagging performance on each tagset, but it depends on the size of the corpus. We also show that treating multiple corpora sharing the same tagset separately instead of merging them leads to a better tagging performance.

The paper is organised as follows. Section 2 describes the proposed neural tagger model. Section 3 provides experimental results, which we discuss in section 4. Section 5 concludes the paper.

2 Methods

Our model receives a tokenised sentence in the form of word³ sequence $\{w_1, w_2, \dots, w_n\}$ as input features, and predicts a sequence of morphological tags $\{t_1^j, t_2^j, \dots, t_n^j\}$ for each tagset T^j . We provide detailed description of the model in the next sections.

2.1 Model architecture

The model has three basic blocks:

1. word embeddings
2. encoder layer
3. output layer.

We used GRU-based (Cho et al., 2014) character-level word embeddings, proven to be effective in various NLP tasks, including morphological tagging (Heigold et al., 2017; Lukovnikov et al., 2017). Each word w_i is represented as a sequence of its characters $\{c_1, c_2, \dots, c_k\}$. Each character is represented as a one-hot encoded vector over a predefined vocabulary V^{char} and passed to a character embedding layer:

$$c_i^{embed} = W^{embed} \cdot one_hot(c_i),$$

where $W^{embed} \in \mathbb{R}^{char_embedding_size \times |V^{char}|}$. All word's character embeddings are then passed to a unidirectional GRU layer:

$$\begin{aligned} r_i &= \sigma(W_r c_i^{embed} + b_r + U_r h_{i-1}^{char} + u_r), \\ z_i &= \sigma(W_z c_i^{embed} + b_z + U_z h_{i-1}^{char} + u_z), \\ n_i &= \tanh(W_n c_i^{embed} + b_n + r_i \odot (U_n h_{i-1}^{char} + u_n)), \\ h_i^{char} &= (1 - z_i) \odot n_i + z_i \odot h_{i-1}^{char}, \\ h_0^{char} &= 0, \end{aligned}$$

where σ is the sigmoid function, $W_r, U_r, W_z, U_z, W_n, U_n \in \mathbb{R}^{char_hidden_size \times char_embedding_size}$ and $b_r, u_r, b_z, u_z, b_n, u_n$ are the bias vectors, respectively. The final hidden state of the character sequence is the word embedding of the word w_i :

$$w_i^{embed} = h_k^{char}$$

As the encoder layer, we chose the Transformer model's encoder. Not only this model showed promising results in various sequence tagging tasks (Devlin et al., 2019) because of its receptive field, but also its architecture allows easier interpretation through visualisation compared to other encoder models including recurrent neural networks. We did not make any changes to the architecture besides hyperparameter tuning (we also did not use the decoder layer of the Transformer) so we refer the readers to the original paper (Vaswani et al., 2017) for more details. The output of the encoder layer is

³We treated punctuation marks as words.

$$w_i^{enc} = TransformerEncoder(w_i^{embed}),$$

where $w_i^{enc} \in \mathbb{R}^{d_{model} \times 1}$.

We made $|T|$ output layers where $|T|$ is the number of tagsets (corpora). Each output layer projects each encoder's output to a probability distribution over a predefined set of tags:

$$w_i^{out} = softmax(W_{out}^j w_i^{enc} + b_{out}),$$

where $W_{out}^j \in \mathbb{R}^{|T^j| \times d_{model}}$, $j = 1, 2, \dots, |T|$ and b_{out} is the bias vector. The predicted morphological tag in a given tagset for a given word is the tag with the highest probability. See Figure 1 for the graphical representation of the model.

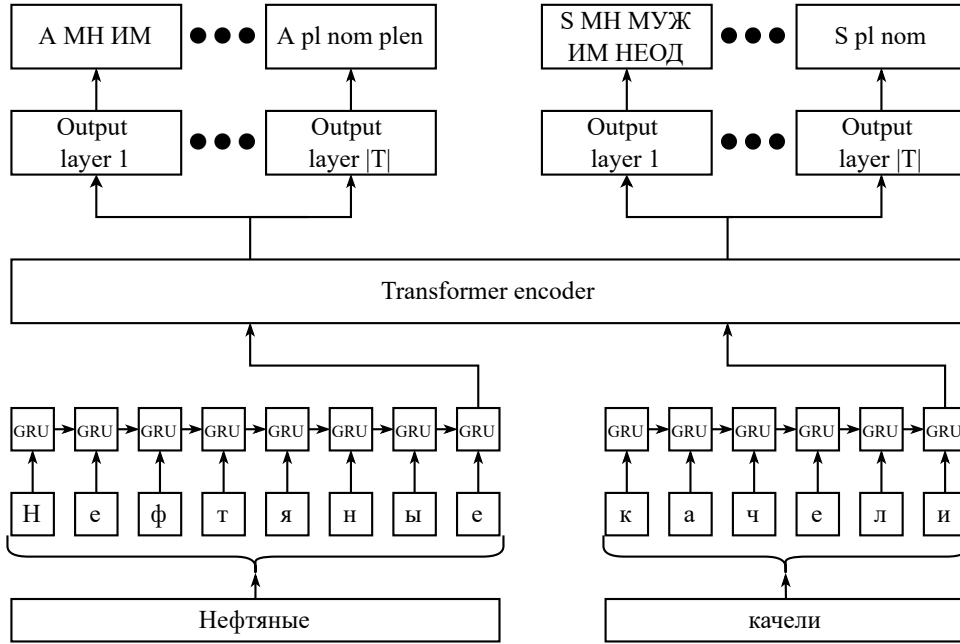


Figure 1: Graphical representation of the proposed neural tagger for the sentence *Neftjanye kačeli*.

2.2 Model hyperparameters

Table 1 shows the hyperparameters we used in our model. We fine-tuned these hyperparameters once and did not change them between the experiments.

Model part	Hyperparameter	Value
Word embeddings	$ V^{char} $	95
	$char_embedding_size$	32
	$char_hidden_size$	128
Encoder layer	d_{model}	128
	d_{ff}	512
	P_{drop}	0.1
Output layer	$ T^j $	Depends on the corpus

Table 1: Hyperparameters of the proposed model. We did not mention some hyperparameters in the section 2.1 and for them, we either use the notation proposed in the paper (Vaswani et al., 2017) or not mention them at all if we did not make any changes.

We used the Adam optimiser with weight decay (Loshchilov and Hutter, 2019). Its hyperparameters as well as learning rate function are almost identical to (Vaswani et al., 2017) except we chose *warmup_steps* to be 10% of the total number of steps.

One problem with our model is the training process. We used cross-entropy as the cost function, but each task (each output layer) has its own cost function and simply adding them up may affect performance since different corpora have different sizes. To overcome this issue, we adopted the approach proposed in (Cipolla et al., 2018). Namely, before adding up, it weighs each cost function by considering the homoscedastic uncertainty of each task.

3 Experiments

We chose eight different corpora to evaluate our model. We divided them into two parts to conduct two different sets of experiments. The first part consists of the manually (re)annotated corpora:

1. Syntactically tagged Russian text corpus SynTagRus (Inšakova et al., 2019). It is a subcorpus of the National Corpus of the Russian language. SynTagRus is supplied with several types of annotation, including fully disambiguated and manually corrected morphological and syntactic annotation.
2. Disambiguated subcorpus of the National Corpus of the Russian language (RNC) (Plungjan and Sičinava, 2004). This subcorpus was manually disambiguated, and it provides full morphological annotation.
3. Russian Universal Dependencies Treebank annotated and converted by Google (GSD)⁴. GSD is a small treebank automatically annotated and converted into UD format. The current version was manually reannotated and provides full morphological and syntactic annotation.
4. Russian Universal Dependencies Treebank based on data samples extracted from Taiga Corpus and MorphoRuEval-2017 and GramEval-2020 shared tasks collections (Taiga)⁵. It includes manually corrected morphological and syntactic annotation.

The second part consists of the corpora provided by the organisers of the MorphoRuEval-2017 contest (Sorokin et al., 2017):

1. UD SynTagRus. It is the SynTagRus corpus automatically converted into UD format.
2. RNC Open. It is a smaller part of the RNC corpus mentioned above being automatically converted into UD format.
3. GICR. It is a morphologically disambiguated part of the General Internet Corpus of Russian (Piperski et al., 2013). It was automatically annotated and then converted into UD format.
4. OpenCorpora. It is a morphologically disambiguated part of the OpenCorpora project⁶. It was manually annotated and then automatically converted into UD format.

We tackled some corpora differently from others. The first difference is how we split the corpora into training, development and test sets. GSD and Taiga corpora have predefined splits, so we left it as is. For SynTagRus and RNC, we used their intersection as test sets and split the remaining sentences randomly so that 10% of the sentences form a development set. For the remaining four corpora, the organisers of the MorphoRuEval-2017 contest provided a shared test set, so we split these corpora into train and development sets with the ratio 9:1, respectively.

The second difference is how we collected grammemes. We used the tagset descriptions provided with SynTagRus, RNC, GSD and Taiga and then omitted all non-inflectional features. For the remaining four corpora, we used only those grammemes which were counted at the testing phase of the MorphoRuEval-2017 contest.

To collect the tagset of a corpus, we followed the following algorithm:

1. Collect each word's tag from a corpus.
2. Exclude unused grammemes from each tag.
3. Remove duplicate grammemes from each tag (in case of annotation errors).

⁴https://universaldependencies.org/treebanks/ru_gsd/index.html

⁵https://universaldependencies.org/treebanks/ru_taiga/index.html

⁶<http://opencorpora.org/>

4. For SynTagRus and RNC: replace each tag in which any grammatical category has two or more different values with a special "erroneous" tag.
5. Sort grammemes in each tag.
6. Return unique preprocessed tags.

See Table 2 for the detailed statistics of each corpus.

Corpus name	#Sentences	#Words	#Grammmemes	#Tags ($ T^j $)
SynTagRus	97138	1685273	45	470
RNC	519726	7961784	62	1285
GSD	5030	98000	52	652
Taiga	17871	197001	54	683
UD SynTagRus	50116	931075	41	237
RNC Open	98892	1344875	41	492
GICR	83148	1086148	41	292
OpenCorpora	38508	457583	41	366

Table 2: Corpora statistics. We treat punctuation marks as words and POS features as grammemes.

We conducted two series of experiments. The first series concerns 4 corpora: SynTagRus, RNC, GSD and Taiga. They have different sizes, tagsets and annotation guidelines. We trained 15 different neural taggers using different subsets of corpora (one tagger for each of the 4 corpora, one tagger for each of the 6 pairs, one tagger for each of the 4 triples and one tagger trained on all 4 corpora) and compared their performance.

The second series of experiments concerns the remaining 4 corpora: UD SynTagRus, RNC Open, GICR and OpenCorpora. These corpora share the same tagset, they are similar in size, but they follow different annotation guideline. We trained and evaluated 15 different neural taggers in the same way as in the first series, but because the tagsets are the same, we were able to train another *combined* tagger using a single merged corpus which consists of all 4 corpora. For that final experiment, we also merged the corpora’s tagsets.

Each tagger has the same model architecture described in section 2.1. We trained each tagger for 10 epochs and chose the final parameters based on the best development set performance. We did not use fixed mini-batch size because different sentences vary in size dramatically. Instead, each mini-batch contained some sentences of the same length from the same corpus with the overall limit of 2048 words per mini-batch. Since each corpus has morphological annotation for only one output layer, we froze the weights of other output layers during training, depending on to which corpus the sentences from the current mini-batch belong.

4 Results

To compare the taggers, we used per-word and per-sentence accuracy. The word is tagged correctly if the tag predicted by tagger is the same as in the gold standard (it means that the tags’ grammemes also match). The sentence is tagged correctly if each word’s tag match with the corresponding tag in the gold standard.

Figure 2 illustrates the per-word tagging accuracy on the test sets for each tagger from the first series of experiments. We arranged the models in ascending order of their joint corpora size. Each line corresponds to the respective output layer, so different lines also correspond to different test sets.

From the results, it is clear that low-resource corpora always benefit from multitask learning scenario when trained jointly with the larger corpora, despite their tagsets and annotation guidelines. The opposite does not hold. However, Taiga and GSD generally benefit from each other, as well as SynTagRus and RNC. We speculate that this might be due to two reasons. The first reason is that these two pairs have comparable corpus size. The second reason is that GSD and Taiga have almost identical tagsets.

Another finding is that despite the single-task learning models show a clear trend “more data — better

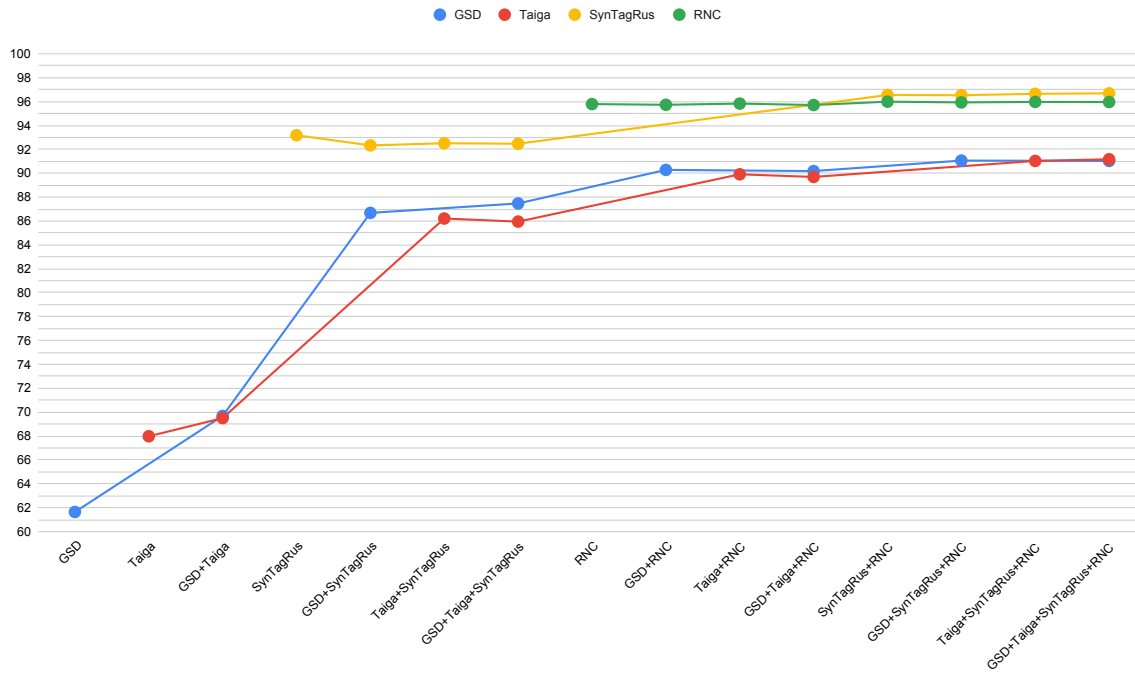


Figure 2: Per-word tagging accuracy (%) on the test sets for each model from the first series of experiments. Each line shows accuracy with respect to its output layer. The models are arranged in ascending order of their joint corpora size.

performance”, the SynTagRus corpus shows the best overall performance. We believe that this is because SynTagRus has relatively small tagset, and it suffers less from the data sparseness problem.

One limitation of our comparison is the fact that we fine-tuned the model architecture’s hyperparameters using the SynTagRus+RNC pair, which might be the reason why these two corpora benefit from each other. At the same time, the best performance for each test set provide the largest or the second-largest model in terms of joint corpora size. This contrasts with the paper (Mishra, 2019): the authors utilised a similar multitask learning approach to do POS tagging of English tweets, but did not improve the results for all corpora compared to a single-task learning approach.

Figure 3 illustrates the performance of the taggers from the second series of the experiments in the same manner as in Figure 2. This chart has two key differences from the previous one. The first difference is that the performance of the model for a given tagset does not depend on the tagset’s corpus size at all: the largest corpus is RNC Open, and it performs poorly compared to the UD SynTagRus and GICR corpora. This appears to be the case of annotation guidelines differences. Since all these models share the same test set, the results show which corpus’ annotation guideline is closer to the test set’s one. This agrees with the fact that according to (Sorokin et al., 2017) the test set is the GICR subcorpus.

The second difference is the fact that here each corpus benefits from all others. This does not contradict the previous findings because all these corpora have comparable size. One exception which is visible on the UD SynTagRus line has already been explained: using data from the GICR corpus leads to better performance.

The best performance of the second series of experiments achieved by the largest model with the GICR tagset prediction layer. We compared our best model with the models provided by the participants of the MorphoRuEval-2017 contest in a closed setup, since we did not use any extra resources. We also added our combined model mentioned in section 3 into comparison. The results are shown in Table 3.

Our best model performance is comparable to the performance of the contest participants’ models,

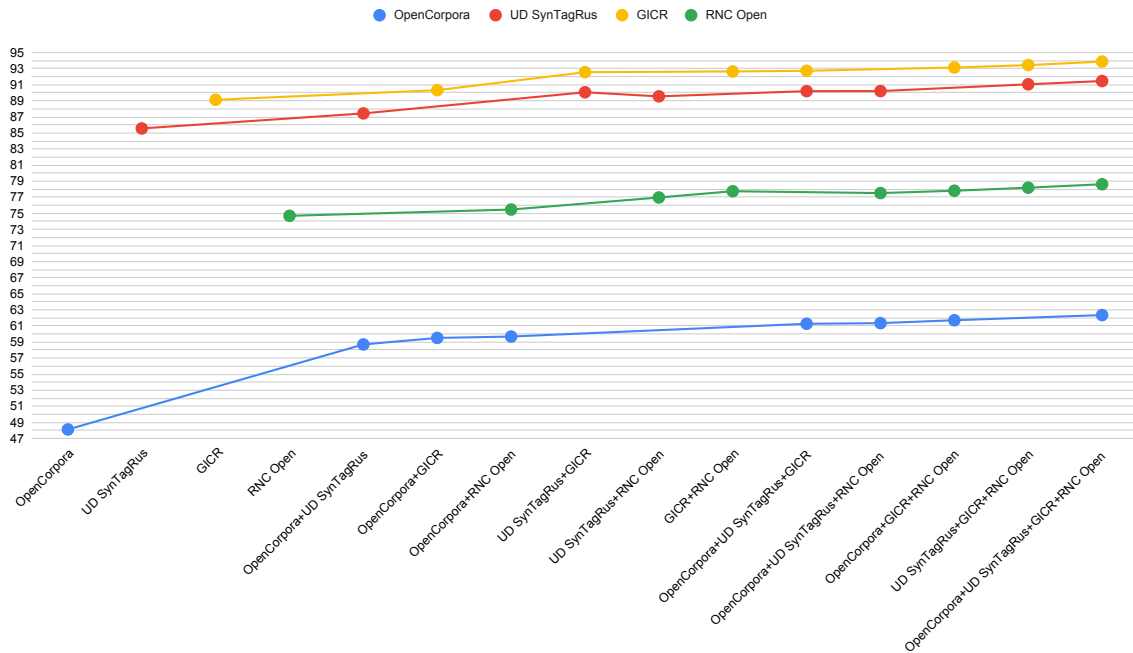


Figure 3: Per-word tagging accuracy (%) on the test sets for each model from the second series of experiments (excluding the combined model). Each line shows accuracy with respect to its output layer. The models are arranged in ascending order of their joint corpus size.

Model name	Per-word accuracy, %	Per-sentence accuracy, %
OpenCorpora+UD SynTagRus+ +GICR+RNC Open (GICR tagset output)	93.88	62.58
Combined	91.25	52.65
MSU-1	93.39	65.29
IQUMEN	93.08	62.71
Sagteam	92.64	58.40
Aspect	92.57	61.01

Table 3: Comparison of our models with the top 4 models provided by the participants of the MorphoRuEval-2017 contest on the test set in a closed setup.

although we did not use any dictionaries or hand-crafted features. We achieved the best per-word accuracy and third best per-sentence accuracy. The comparison with the combined model provides supporting evidence that even corpora with a shared tagset may perform poorly when merged together because of the differences in the annotation guidelines.

5 Conclusion

In this paper, we proposed a multitask learning based approach to Russian neural morphological tagging, which effectively utilises multiple corpora with different tagsets or annotation guidelines. To our knowledge, we for the first time applied the multitask learning technique in terms of predicting tags from different tagsets to the task of morphological tagging of Russian texts.

We showed that the effectiveness of morphological tagging depends on corpora size, tagset size and annotation consistency. Our findings help to better understand how tagset conversion affects performance of NLP tasks.

Our model is able to indirectly make tagset conversion in a scalable way taking into account differences in the morphological annotation guidelines, but full morphologically annotated corpora conversion does not end there. Such corpora often have other differences, including tokenisation and lemmatisation scheme. This may constitute the object of future studies.

Acknowledgements

This research has been partially supported by the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15-2020-793.

References

- Ivan Andrianov and Vladimir Mayorov. 2017. Transfer learning for morphological tagging in Russian. // *2017 Ivannikov ISPRAS Open Conference (ISPRAS)*, P 58–63.
- V. V. Bočarov, S. V. Alekseeva, D. V. Granovskij, E. V. Protopopova, M. E. Stepanova, and A. V. Surikov. 2013. Crowdsourcing morphological annotation. // *Papers from the Annual International Conference “Dialogue” (2013)*, volume 1, P 109–114.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. // *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, P 103–111, Doha, Qatar, October. Association for Computational Linguistics.
- Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. // *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, P 7482–7491.
- 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.
- Kira Droganova and Daniel Zeman. 2016. Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies. Technical report, ÚFAL MFF UK.
- Jirka Hana and Anna Feldman. 2010. A positional tagset for Russian. // *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, May. European Language Resources Association (ELRA).
- Georg Heigold, Guenter Neumann, and Josef van Genabith. 2017. An extensive empirical evaluation of character-based morphological tagging for 14 languages. // *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, P 505–513.
- Evgenija Inšakova, Leonid Iomdin, Leonid Mitjušin, Viktor Sizov, Tat’jana Frolova, and Leonid Cinman. 2019. SynTagRus segodnja [SynTagRus today]. *Trudy russkogo jazyka im. V. V. Vinogradova*, P 14–40.

- Zhenghua Li, Jiayuan Chao, Min Zhang, and Wenliang Chen. 2015. Coupled sequence labeling on heterogeneous annotations: POS tagging as a case study. // *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, P 1783–1792, Beijing, China, July. Association for Computational Linguistics.
- O. N. Ljaševskaja, I. Astaf'eva, A. Bonč-Osmolovskaja, A. Garejšina, Ju. Grišina, V. D'jačkov, M. Ionov, A. Korol'eva, M. Kudrinskij, A. Litjagina, E. Lučina, E. Sidorova, and S. Toldova. 2010. NLP evaluation: Russian morphological parsers. // *Papers from the Annual International Conference "Dialogue" (2010)*, P 318–326.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. // *International Conference on Learning Representations*.
- Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. 2017. Neural network-based question answering over knowledge graphs on word and character level. // *Proceedings of the 26th International Conference on World Wide Web, WWW '17*, P 1211–1220, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Olga Lyashevskaya, Victor Bocharov, Alexey Sorokin, Tatiana Shavrina, Dmitry Granovsky, and Svetlana Alexeeva. 2017. Text collections for evaluation of Russian morphological taggers. *Jazykovedny Casopis*, 68(2):258–267.
- Shubhanshu Mishra. 2019. Multi-dataset-multi-task neural sequence tagging for information extraction from tweets. // *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, P 283–284.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. // *Proceedings of the 12th Language Resources and Evaluation Conference*, P 4034–4043, Marseille, France, May. European Language Resources Association.
- Alexander Piperski, Vladimir Belikov, Nikolay Kopylov, Vladimir Selegey, and Sergey Sharoff. 2013. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. // *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, P 24–29.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. // *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, P 742–751.
- V. A. Plungjan and D. V. Sičinava. 2004. Nacional'nyj korpus ruskogo jazyka: opyt sozdanija korpusa tekstov sovremennogo ruskogo jazyka [Russian National Corpus: Experience in creating a corpus of texts of the modern Russian language]. // *Trudy meždunarodnoj konferencii «Korpusnaja lingvistika-2004»*, P 216–238.
- Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a Russian tagset. // *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).
- Aleksei Sorokin, Tatiana Shavrina, Olga Lyashevskaya, Victor Bocharov, Svetlana Alexeeva, Kira Droganova, Alena Fenogenova, and Dmitry Granovsky. 2017. MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. // *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017)*, volume 1, P 297–313.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. // I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Detecting Auto-generated Texts with Language Model and Attacking the Detector

Mikhail Orzhenovskii
Saint Petersburg, Russia
orzhan057@gmail.com

Abstract

We propose a simple approach to the detection of automatically generated texts. A pre-trained language model, fine-tuned on the shared task's dataset, achieved 3rd place on the binary task leaderboard with 82.6% accuracy. In the multi-task leaderboard, the language model achieved an F1 score of 64.5% after being fine-tuned with the same procedure. In order to investigate the weaknesses of this approach, we explore two possible attacks on the detector: selecting from language model outputs and directed beam search. These attacks reduce the likelihood of detecting the generated texts without significant loss in quality. Both attacks do not require retraining the generative model and are applied at inference time.

Keywords: generated text detection, text classification, pre-trained language models

DOI: 10.28995/2075-7182-2022-21-412-419

Обнаружение сгенерированного текста с помощью языковой модели и атаки на детектор

Орженовский М.В.
Санкт-Петербург, Россия
orzhan057@gmail.com

Аннотация

Мы предлагаем простой подход к обнаружению автоматически сгенерированных текстов. После дообучения на данных соревнования, предобученная языковая модель заняла 3 место в бинарной классификации с точностью 82.6%. В задаче множественной классификации, аналогичная модель достигла 64.5% по метрике F1 (также 3 место). Изучая слабые стороны такого подхода, мы рассматриваем два типа атак: выборка из сгенерированных языковой моделью текстов и направленный лучевой поиск. Эти атаки снижают вероятность обнаружения сгенерированных текстов без существенной потери их качества. Обе атаки не требуют переобучения генеративной модели, внедряются на этапе исполнения.

Ключевые слова: обнаружение сгенерированного текста, классификация текстов, предобученные языковые модели

1 Introduction

Large language models are capable of generating high-quality coherent texts. However, they can be used to generate fake news or product reviews. The shared task (Shamardina et al., 2022) is focused on detecting Russian texts that were created with large language models, distinguishing them from human-written ones. One of the tracks is binary classification: identifying whether a text was written by a human or a language model. Another track (multi classification) is to find which model was used to generate the text.

Humans cannot easily solve this problem. The longer the text is, the easier it is to identify the source. Most language models work with limited context, which causes them to lose coherence if a long text is produced. The dataset in the shared task includes short texts (one sentence long) which makes the task challenging.

Our approach is straightforward: fine-tuning a large pre-trained language model with a sentence classification head. We release the source code for training the models and the models' weights¹.

Additionally, we analyze the weaknesses of the proposed model. We explore two methods of generating texts that are less likely to be detected: selecting one of the language model's outputs with a discriminator model, and adversarial beam search driven by a discriminator model. While the unmodified language model's output is detected in 65% cases, these methods produce texts that are detected in 55% and 42% cases respectively. We are not publishing the code for the attacks.

2 Related work

(Gehrmann et al., 2019) use BERT and GPT models to identify top-k rank of each word in the text; generated texts consist of words with lower ranks, and human-generated texts include a high fraction of high-ranked words.

(Uchendu et al., 2020) apply neural models built on RNN and CNN as well as fine-tuned RoBERTa in the settings that are related to both binary and multi tracks of the shared task.

(Ippolito et al., 2020) show that humans and machines use different approaches to identify generated text. The authors show that common decoding strategies introduce statistical features that can be used by automatic systems.

(Pillutla et al., 2021) introduce a comparison measure to compare the distributions of human-written and machine-generated texts, using divergence frontiers.

(Scialom et al., 2020) use guided decoding with a discriminator to generate human-like texts in abstractive summarization task.

3 Datasets

3.1 Binary track

There are 129,065 samples in the training set, 21,511 samples in the validation set and 64,533 samples in the test set. In the training set, the length of the samples varies from 1 to 376 words, mean length 31 words, 75% of samples have no more than 22 words. 64,535 samples in the training set belong to class H (human-written) and 64,531 samples belong to class M (machine-written). The statistics for both classes are not very different. The dataset is balanced. Examples of the texts are displayed in Table 1.

Таблица 1: Dataset examples

Sentence	Binary class	Multi class
Никто ни разу не навесил меня в больнице	M	OPUS-MT
Под монастырем, на самой верхушке скалы, обнаружил почти 200 древних археологических находок.	M	ruGPT3-Large
На лицо и руки садился тяжелый и липкий туман.	H	Human
Они чем-то кормились на земле и только в случае тревоги взлетали на деревья.	H	Human

3.2 Multi track

As we can see in Table 2, Human class is the largest. The classes of language models are not balanced, for example OPUS-MT has 7 times more samples than ruT5-Large. Average word count also differs between the models: average sample from ruGPT3-small has 71 word, and average M-BART50 sample has only 10 words. Examples of the texts are displayed in Table 1.

¹<https://github.com/orzhan/ruatd>

Table 2: Value counts in the training part of Multi track dataset

Class	Count	Percentage	Average word count
Human	51150	39.63%	35.43
OPUS-MT	12087	9.36%	9.95
M-BART50	11913	9.23%	9.67
M2M-100	10817	8.38%	9.87
ruGPT3-Large	9870	7.64%	53.70
ruGPT3-Medium	7020	5.44%	68.42
ruGPT3-Small	6930	5.37%	71.06
mT5-Large	4860	3.77%	12.49
mT5-Small	2940	2.28%	10.52
ruT5-Base	2640	2.05%	32.05
M-BART	2510	1.94%	29.40
ruGPT2-Large	2370	1.84%	14.19
ruT5-Base-Multitask	2219	1.72%	11.59
ruT5-Large	1740	1.35%	16.21

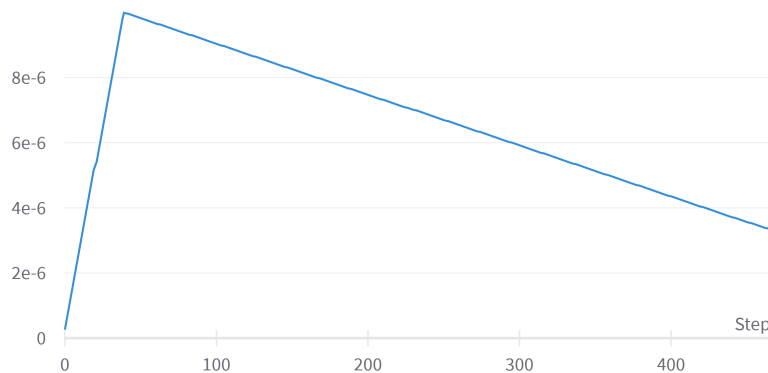


Figure 1: Learning rate of the binary model.

4 Classification model description

We used pre-trained language model sberbank-ai/ruRoberta-large² based on (Liu et al., 2019). This model scores high on the Russian SuperGLUE leaderboard (Shavrina et al., 2020), indicating its strong capabilities in various tasks. We directly fine-tune the model with a classification head on the training part of the dataset without any pre-processing. For the experiments, we set the learning rate to a relatively small value 1×10^{-5} and use weight decay 0.01 and label smoothing factor 0.1. For binary classification, we picked the model from epoch 2, which had the best evaluation accuracy; for multi classification, we also chose the model from epoch 2, with the best evaluation F1 score. For the experiments, we used an implementation by HuggingFace Transformers (Wolf et al., 2019).

The learning rate chart is shown on Figure 1. The accuracy chart is shown on Figure 2.

5 Attacks on the classification model

We explore an imaginary situation when the classification model is used to filter out malicious texts, and the attacker knows about it; however, the attacker does not have access to the model and the data that was used to train the model. The attacker collects another dataset of human-written and machine-generated

²<https://huggingface.co/sberbank-ai/ruRoberta-large>

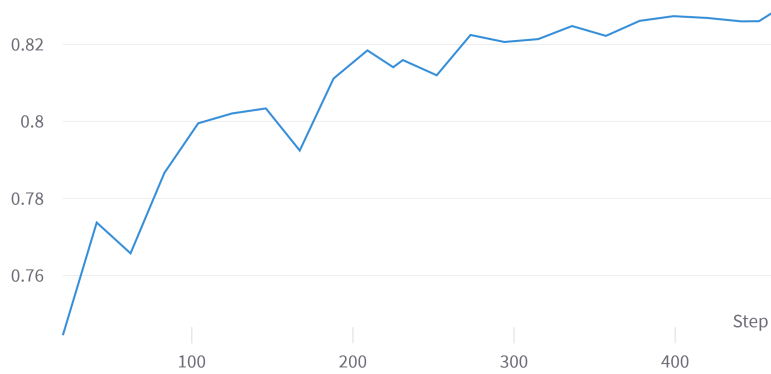


Figure 2: Accuracy of the binary model.

texts. The attacker already has some generative model and is going to use it to produce the texts. To reduce the percentage of rejected texts, the attacker trains some discriminator model and uses it to direct their generative model.

For the purpose of evaluating the attack methods, we used RuSimpleSentEval task (Sakhovskiy et al., 2021) from Dialogue Evaluation 2021. This text-to-text task requires to write a simplified version of a Russian sentence. The target metric is SARI (Xu et al., 2016) (generated text is compared to several human-written reference sentences). We calculate attack success rate as percentage of the generated sentences that were falsely identified as human-written by the main classification model. First, we generate simplifications with fine-tuned ruGPT3 by (Orzhenovskii, 2021) and calculate SARI metric on the public test part of RuSimpleSentEval dataset.

For both attacks we use the same discriminator model, which is based on rubert-tiny2³. To classify unfinished sentences, we augment the dataset with partial sentences, making 160,000 samples out of 20,000 from the validation set. This small discriminator model is further used to score the generative language model’s outputs.

In the first attack (scoring), we produce several outputs from the generative model using nucleus sampling (Holtzman et al., 2019) and calculate the class probabilities with the discriminator model. The output with the highest human-class probability is selected.

In the second attack (beam search), we use beam search with one modification. Following (Scialom et al., 2020), on each step of the beam search algorithm, we add discriminator model’s log-probability $S_{dis}(\hat{y})$ to the generator’s log-probability $S_{gen}(\hat{y})$ of the partial sequence \hat{y} , so that beams that look machine-generated are less likely to be selected.

$$S_{DAS}(\hat{y}) = S_{dis}(\hat{y}) + \alpha \times S_{gen}(\hat{y})$$

where $\alpha > 0$ is a weighting factor.

The process is repeated until end-of-sentence token is generated. This method slow, because we have to run inference of the discriminator model N times per token. In our experiments, we used $\alpha = 1$ and $N = 10$ beams.

6 Results and analysis

6.1 Binary classification model

The binary model scored 0.82629 on the private leaderboard (3rd) which is close to the validation score of 0.83054. These numbers are far behind the results from (Uchendu et al., 2020) where RoBERTa-

³<https://huggingface.co/cointegrated/rubert-tiny2>

tuned scored 0.9702 in the same binary setting; the difference can be attributed to significantly longer texts (average word count was 432 against 31 in RuATD shared task).

Analysis of the validation results indicates positive correlation 0.177 between word count and results. Longer texts are easier to classify, as expected. See Table 3.

Table 3: Binary model: accuracy and text length

Number of words	Number of samples	Accuracy
1 to 9	6208	0.759
10 to 13	4238	0.785
14 to 22	5738	0.811
23 and more	5324	0.952

6.2 Multi classification model

The multi model achieved 0.64573 on the private leaderboard (3rd). This value is close to the validation score of 0.64427.

F1 score is highest for Human (the largest class), ruT5-Base, and ruGPT3-Small (Table 4). In the dataset, the samples from these classes were longer (30+ words), which could help in their identification. Another possibility is that the base model ruROBERTa-large has been trained a dataset that overlaps with the training data of ruT5 and ruGPT3 models, which improves the score.

The confusion matrix (Figure 3) indicates that certain models are more difficult to distinguish. Interestingly, different sizes of the same model (like ruGPT3-small and ruGPT3-medium, or ruT5-base and ruT5-large) that have been trained on the same data, can be separated by the classifier.

Models OPUS-MT, ruGPT3-small, ruGPT3-medium, ruGPT3-large, ruT5-base, M-BART are the easiest to distinguish from human. Models M2M-100 and ruGPT2-large which had the shortest samples in the dataset, were the most difficult to identify.

Table 4: Classification report for Multi model

	precision	recall	f1-score	support
Human	0.81	0.90	0.85	8524
M-BART	0.69	0.53	0.60	418
M-BART50	0.56	0.39	0.46	1986
M2M-100	0.36	0.23	0.28	1804
OPUS-MT	0.36	0.64	0.46	2014
mT5-Large	0.37	0.22	0.27	810
mT5-Small	0.61	0.43	0.50	490
ruGPT2-Large	0.67	0.63	0.65	395
ruGPT3-Large	0.63	0.59	0.61	1645
ruGPT3-Medium	0.72	0.56	0.63	1170
ruGPT3-Small	0.66	0.72	0.69	1155
ruT5-Base	0.66	0.79	0.72	440
ruT5-Base-Multitask	0.48	0.22	0.30	370
ruT5-Large	0.37	0.09	0.14	290

6.3 Attacks

We evaluate our approaches based on two metrics: SARI (the target metric of the text generation task, higher is better) and attack success rate (higher is better).

The unmodified model achieved SARI 39.06, and only 35% of samples were not detected by the binary classification model. Scoring attack using discriminator model with the best parameters achieved SARI

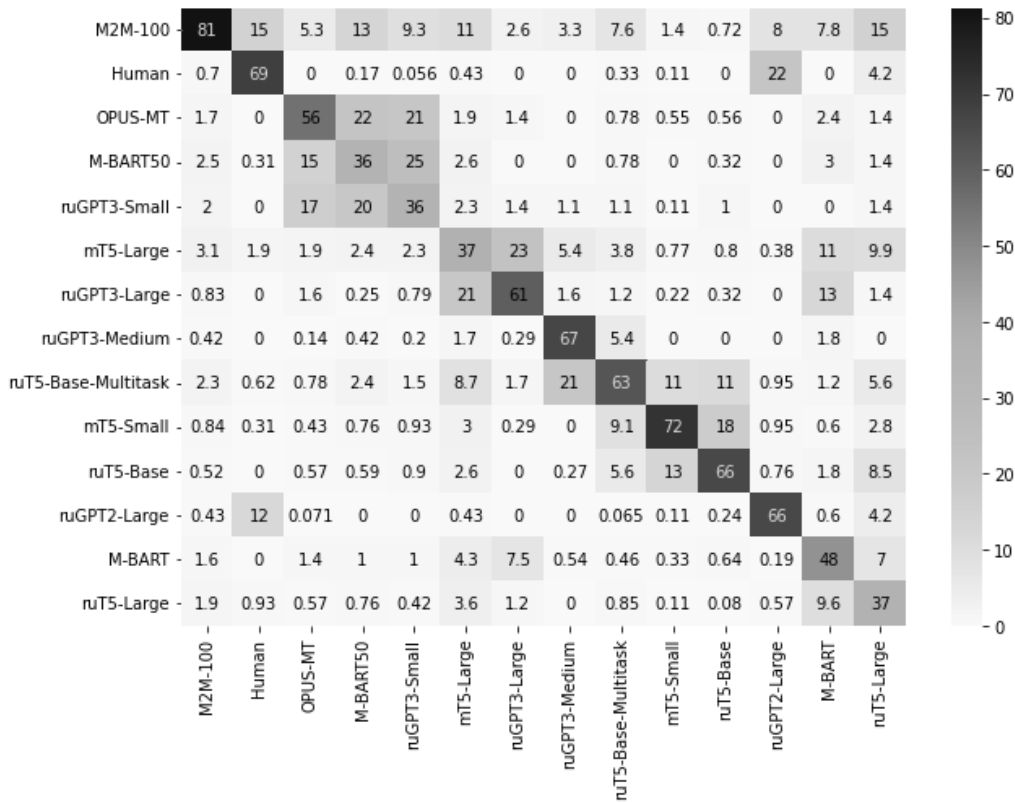


Figure 3: Confusion matrix for the multi model.

38.27, and 46% samples successfully escaped detection. By modifying the number of generated samples before scoring, we can trade-off between increasing SARI and detection rate.

Beam search attack was able to achieve SARI 38.61, and 56% samples were not identified as machine-generated. See Figure 4.

Examples of original and adversarial samples are shown in Table 5. Binary class is predicted by the Binary model. Adversarial beam search sometimes produces longer sentences compared to the original model. This may be the cause of reduced SARI score, however some of the sentences are more fluent and are not detected as machine-generated.

7 Conclusion

With our simple approach, we are able to achieve high accuracy on the leaderboard. However, it does not indicate practical applicability. For shorter texts, the model’s accuracy is as low as 75%. The future development of large language models will make this task even more challenging. Additionally, adversarial generation can fool some models of automatic classification. Ensemble-based models or statistical methods may be more resistant to such attacks.

References

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration.

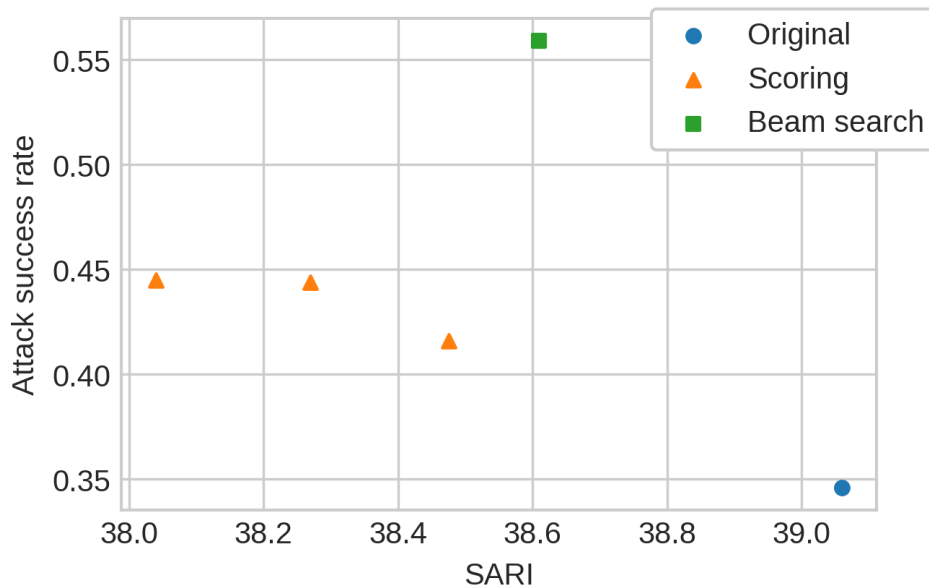


Figure 4: Attack success rate and SARI score for different attack methods.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. P 1808–1822, 01.

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.

Mikhail Orzhenovskii. 2021. Rusimscore: unsupervised scoring function for russian sentence simplification quality. // *Proceedings of the international conference Dialogue 2021*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers.

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. Rusimpleseval-2021 shared task: Evaluating sentence simplification. // *Proceedings of the international conference Dialogue 2021*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Discriminative adversarial search for abstractive summarization.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian. // *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference Dialogue*, volume 21, P xxx–xxx.

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*.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8384–8395, Online, November. Association for Computational Linguistics.

Таблица 5: Generated text examples

Sentence	Model	Binary class
Сенат США постановил, что законы штатов запрещают компаниям платить за пользование интернетом.	Original	M
Сенат США принял закон, запрещающий властям штатов и местным администрациям взимать налоги с пользователей Интернета.	Beam search	H
Английский язык был языком в раннем средневековье.	Original	M
Английский язык - это язык, на котором говорили в раннем средневековье в Британии.	Beam search	M
Додд сказал, что британскую речь можно назвать лицемериями.	Original	M
Додд сказал, что британская речь - пример лицемерия. Он вспомнил события в Ирландии, а также в Индии.	Beam search	M
Анализ грунта, который доставил Аполлон, дал понять, что лунная почва по составу отличается от земной.	Original	H
Анализ лунного грунта показал, что он по составу сильно отличается от земной почвы.	Beam search	H

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2019. Huggingface’s transformers: State-of-the-art natural language processing.

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.

Russian temporal clauses in terms of semantic-syntactic integration (a case study of a coordinating conjunction *kogda* ‘when’)

Pekelis O. E.

Russian State University for the Humanities/Moscow, Russia
opekelis@gmail.com

Abstract

The paper argues that the Russian temporal conjunction *kogda* ‘when’, which is traditionally treated as subordinating, may also be coordinating. Due to a very subtle semantic difference between the coordinating and subordinating instances of *kogda*, it is difficult to distinguish between them using the well-known criteria of coordination and subordination. This difficulty is solved as follows: first, contexts are introduced in which the two *kogda* clearly differ both semantically and syntactically; then the coordinating *kogda* is analyzed in these contexts based on the criteria. In the syntactic literature, temporal clauses are generally assumed to be more closely integrated with the host clause than, for example, causal or concessive clauses. Given the conclusion about the existence of coordinate temporal clauses, this point of view needs reconsideration.

Keywords: coordination, subordination, temporal clauses, adverbial clauses

DOI: 10.28995/2075-7182-2022-21-420-435

Русские временные клаузы на шкале семантико-синтаксической интеграции (на примере сочинительного союза *когда*)

Пекелис О. Е.

Российский государственный гуманитарный университет/Москва, Россия
opekelis@gmail.com

Аннотация

В статье обосновывается существование в русском языке сочинительного временного союза *когда*. Из-за зыбкости семантического различия между сочинительным и подчинительным *когда* их трудно разграничить, используя общепринятые критерии сочинения и подчинения. В работе эта трудность решается следующим образом: сначала выделяются контексты, в которых два *когда* отчетливо различимы на основе семантических и формальных признаков, затем в этих контекстах *когда* анализируется на основе критериев. В синтаксической литературе временные клаузы обычно считаются более тесно интегрированными с главной клаузой, чем, например, причинные и уступительные. Вывод о существовании сочинительных временных союзов требует пересмотра этой точки зрения.

Ключевые слова: сочинение, подчинение, временные клаузы, адвербиальные клаузы

1. Введение

Зависимые клаузы разных типов могут различаться степенью семантико-синтаксической интеграции с главной клаузой. Среди адвербиальных клауз выделяют «центральные», «периферийные» и «неинтегрированные» клаузы, из которых наименее тесно интегрированы последние: предполагается, что они связаны с главной клаузой только семантически, т.е. синтаксическое подчинение как таковое в их случае отсутствует [9: 155 ff.], [5], [6]. Замыкают эту шкалу сочиненные клаузы: степень их интеграции, по-видимому, еще ниже, чем у «неинтегрированных» адвербиальных клауз.

На этом фоне временные клаузы, как считается, тяготеют к центральным, существенно интегрированным клаузам [8], [1]. Однако данные русского языка указывают на то, что временные клаузы могут значительно различаться по степени интеграции. Так, союз *пока* может

употребляться иллокутивно (1), что считается признаком «неинтегрированных» адвербиальных клауз [5], тогда как производный от него союз *до тех пор пока* к иллокутивному употреблению не способен (2).

(1) *Пока помню – в ЭД появится ТТН или нет?* [Google]

(2)^{??}*До тех пор пока помню – в ЭД появится ТТН или нет?*

Предмет настоящей работы – сочинительная разновидность союза *когда*, само существование которой доказывает, что временные клаузы допускают самую разную, в том числе минимальную, степень интеграции. Сочинительное употребление *когда* представлено в примере (3). Как и в подчинительных конструкциях с *когда* [15: 542], в (3) *когда* выражает совпадение двух ситуаций во времени, но, в отличие от подчинения, ни одна из ситуаций не служит временным ориентиром для другой. В самом деле, в подчинительных предложениях с *когда* временным ориентиром чаще всего выступает ситуация, совпадающая с ремой. Так, именно рему, т.е. ситуацию, выраженную в постпозитивной клаузе, уместно расценивать как временной ориентир в примерах (4)–(7).¹ Это представляется очевидным в случаях (4) и (5), где рема совпадает с зависимой клаузой, но и при совпадении ремы с главной клаузой, как в (6) и (7), ориентация осуществляется относительно последней. Например, в (6) сообщается, что ситуация ‘Олег вошел в дом’ совпала по времени с ситуацией ‘зазвонил телефон’, т.е. первая ситуация ориентирована относительно второй, но не наоборот. Показательно, что (6), в отличие от (4), не может служить ответом на вопрос *Когда зазвонил телефон?* Подходящим для (6) был бы вопрос *Что произошло, когда Олег вошел в дом?* (или иначе: *С какой ситуацией совпала во времени ситуация ‘Олег вошел в дом’?*), где ситуация ‘Олег вошел в дом’ входит в известное вопроса – в соответствии с тем, что вопрос направлен на поиск другой ситуации (ориентира), которой первая ситуация должна быть сопоставлена на временной оси. В примере (3), между тем, выраженная в постпозитивной клаузе ситуация ‘Олег поднялся на смотровую площадку’ не может расцениваться как временной ориентир для ситуации ‘было восемь часов’: наступление восьми часов представляет собой ситуацию, которая возникает независимо от внешних обстоятельств и которую с прагматической точки зрения нелепо ориентировать относительно времени другой ситуации (исключение составляет контекст типа (10), где такая конфигурация все же реализуется, но в других синтаксических условиях).²

(3) *Было восемь часов, когда один из матросов, Олег, поднялся на мачтовую смотровую площадку.* (Национальный корпус русского языка, НКРЯ, [13])

(4) *Телефон зазвонил/, когда Олег вошел в дом\.*

(5) *Олег вошел в дом/, когда зазвонил телефон\.*

(6) *Когда Олег вошел в дом/, зазвонил телефон\.*

(7) *Когда зазвонил телефон/, Олег вошел в дом\.*

Параллельные с (3) подчинительные примеры представлены в (8), (9) и (10). В (8) и (9) в качестве временного ориентира выступает ситуация ‘было восемь часов’, выраженная в зависимой клаузе в (8) и в главной в (9), но в обоих случаях совпадающая с ремой. В (10) по аналогии с (6) и (7) ориентиром уместно считать ситуацию ‘Олег поднялся на смотровую площадку’, также входящую в рему. Отметим, что в этом случае в (10) ситуация ‘было восемь часов’ оказывается ориентирована относительно другой ситуации, что выше, в связи с примером (3), мы признали невозможным. Причина того, что в контексте (10), в отличие от (3), такая конфигурация все-таки допустима, видится в следующем. В (3) предполагаемый временной ориентир (при гипотетической подчинительной интерпретации *когда*) помещен в зависимую клаузу, а в (10) – в главную. При этом для зависимой клаузы роль ориентира более естественна, чем для главной (прежде всего, ввиду таких конструкций, как *Олег поднялся на площадку в восемь часов/когда было восемь часов*). В результате восприятие постпозитивной клаузы как ориентира более отчетливо в (3), чем в (10), и негативные последствия роли ориентира в (3) сильнее, чем в (10). Это различие отражается и в поведении вопросов, ассоциируемых с (3) и (10). В (3) при интерпретации постпозитивной клаузы как временного ориентира вопрос

¹В примерах (4)–(7) и далее знаками «\», «/», «\|», «//» обозначаются носители коммуникативно релевантных акцентов в соответствии с нотацией [18: 36–37].

² Прагматически неестественная ориентация ситуации ‘было восемь часов’ относительно другой ситуации может также реализовываться при подчинительном *когда* в условиях контрастного выделения придаточного, но и в этом случае требует контекстной поддержки, ср. *Когда Олег поднялся на смотровую площадку \|, было восемь часов*.

принимает вид *Когда было восемь часов?* и является прагматически неудачным. В (10) вопрос *Что произошло, когда было восемь часов?* прагматически уместен.³

(8) *Олег поднялся на смотровую площадку/, когда было восемь часов*^.

(9) *Когда Олег поднялся на смотровую площадку/, было восемь часов*^.

(10) *Когда было восемь часов/, Олег поднялся на смотровую площадку*^.

Предполагаемое семантическое различие между подчинительным и сочинительным *когда* не является очевидным. Закономерно, поэтому, что в предшествующих работах (см., например, [15: 545], [2]) граница между двумя классами конструкций с *когда* проводится на разных, иногда противоречащих друг другу, основаниях. Рамки статьи не позволяют нам рассмотреть разные подходы подробно (ср., однако, краткий обзор в [12: 221 ff.]), поэтому мы позволим себе сразу приступить к изложению оснований для разграничения двух *когда* в терминах сочинения и подчинения.⁴ Сначала (раздел 2) мы рассмотрим несколько контекстов, для которых сочинительное прочтение *когда* особенно характерно. Затем, опираясь на эти контексты, мы проанализируем сочинительное *когда* с точки зрения известных признаков, ассоциируемых с сочинением и подчинением (раздел 3). В заключение мы коснемся вопроса о существовании в русском языке других сочинительных временных союзов (раздел 4).

2. Сочинительный *когда*: особенности дистрибуции

Сочинительное прочтение *когда* может подчеркиваться лексическим наполнением лицензирующей (т.е. не содержащей союза) клаузы.⁵ Среди таких, характерных для сочинительного *когда*, признаков выделим два: 1) наличие в лицензирующей клаузе выражения, кодирующего нарушение естественного хода развития ситуации, таких как частица *было* (11), глаголы *собираться* (12), *отчаяться* и под.; 2) наличие в лицензирующей клаузе адвербиала со значением контактного предшествования *только что*, *только-только* и под. (примеры см. ниже).

(11) *От скуки и шума дождя я уже начал **было** дремать, когда вдруг зазвонил телефон.* [13]

(12) *Ваксон **собрался** уже свернуть торговлю, когда через шумную толпу к нему пробралась Ралисса.* [13]

Нарушение нормального хода течения ситуации (ср. в (11): ‘начал дремать, но не задремал’, см. о семантике *было* [14: 187]) неестественно характеризовать с точки зрения временной локализации. Поэтому сочинительное прочтение *когда* в таких случаях обычно предпочтительно. Последняя закономерность представляет собой не строгое правило, а только тенденцию: маркеры нарушения естественного хода развития ситуации и подчинительное *когда* не являются принципиально несовместимыми. Так, в (13) ситуация ‘Граве узнал, что сибирское, уральское и самарское правительства уступили власть омской Директории’ служит временным ориентиром для ситуации ‘Граве повеселел’; при этом нарушение нормального хода развития ситуации ‘повеселел’ произошло позже, так что противоречия между подчинительным употреблением *когда* и семантикой *было* не возникает.

(13) *Граве повеселел **было**, когда узнал, что сибирское, уральское и самарское правительства уступили свою власть омской Директории. <...> Но вот совсем неожиданно под ударами красных пали Казань и Симбирск, на волосок от гибели Самара.* [13]

Вместе с тем, по данным НКРЯ (см. таблицу 1) отмеченная тенденция является статистически значимой: в контексте частицы *было* частотность сочинительного *когда* значимо выше, чем в отсутствие *было* и других маркеров нарушения нормального хода развития ситуаций (критерий χ^2 , $P < 0,01$). Поясним, что выборку предложений без *было*, легшую в основу таблицы 1, составили

³ Альтернативная точка зрения могла бы состоять в том, что временным ориентиром при любой коммуникативной структуре предложений с подчинительным *когда* выступает зависимая клауза. Такая трактовка столкнулась бы с трудностями при интерпретации предложений, содержащих указание на время в составе главной клаузы, как в (9). Кроме того, предположение о том, что расположение ориентира зависит от коммуникативной структуры предложения, позволяет объяснить контраст между примерами (16) и (17), с одной стороны, и (18), с другой, обсуждаемый в разделе 2, тогда как трактовка зависимой клаузы как неизменного ориентира оставляет этот контраст без объяснения.

⁴ В [4: 233] на похожих основаниях выделяется класс *narrative clauses* среди предложений, вводимых союзом *when*.

⁵ Термин «лицензирующая клауза» заимствован в [5]. Поскольку речь идет о сочинении, более привычный термин «главная клауза» здесь неуместен.

50 случайных примеров с постпозитивным *когда*. При разделении примеров на сочинительные и подчинительные мы руководствовались приведенными выше семантическими соображениями (временная ориентация одной ситуации относительно другой при подчинительном *когда* и ее отсутствие при сочинительном; нарушение нормального хода развития ситуации следует за наступлением ситуации, вводимой *когда*, или одновременно с ним и др.), обращая внимание на внешний контекст, особенности просодического оформления клауз, манифестирующие коммуникативную структуру, и другие признаки, подробное описание которых не уместилось бы в рамки настоящей работы.

	сочинение	подчинение	всего	доля сочинения
с <i>было</i>	30	10	40	0,8
без <i>было</i>	1	49	50	0,02

Таблица 1: Сопоставительная частотность сочинительного и подчинительного *когда* в контексте частицы *было* и без нее (Подкорпус текстов 1951-2011 гг. Основного корпуса НКРЯ)⁶

Среди признаков, различающих сочинительный и подчинительный *когда* в контексте маркеров нарушения нормального хода развития ситуации, следующий признак выделяется как наиболее отчетливый и последовательный: только сочинительный *когда* в этом контексте заменим на союз *но*, ср. приемлемость замены *когда* на *но* в (11) (ср. (14)), но не в (13) (ср. (15)). Дело в том, что союз *но* выражает т.н. «обманутое ожидание» (см., например, [17: 171 ff.]) и поэтому востребован при необходимости указать причину того, что нормальный ход развития ситуации нарушился. Такую причину имплицитно вводит и сочинительный *когда*, указывая на ситуацию, совпавшую во времени с нарушением нормального хода развития событий. Напротив, ситуация, вводимая подчинительным *когда*, выступает в качестве временного ориентира для ситуации, предшествовавшей нарушению нормального хода развития событий, что не согласуется со значением *но*.⁷

(14) (11) ≈ *От скуки и шума дождя я уже начал было дремать, но вдруг зазвонил телефон.*

(15) (13) ≠ *Граве повеселел было, но узнал, что сибирское, уральское и самарское правительства уступили свою власть омской Директории.*

Обратимся ко второму из вышеупомянутых контекстов, характерных для сочинительного *когда*: контексты с адвербиалом контактного предшествования *только что* (16), *только-только* (17) и под. в лицензирующей клаузе.

(16) *Я сравнительно поздний современник Чуковского — я только что взял в руки перо, когда он был уже заметным критиком и основателем новой детской поэзии.* [13]

(17) *Смена только-только началась, когда на участок раскладки заглянул десятник Мураш.* [13]

К сочинительному прочтению *когда* в этом случае снова располагает то, что подчинительное прочтение является прагматически неестественным: оно предполагает ориентацию момента повествования относительно двух разных точек на временной оси – недавнего прошлого (о семантике *только что*, *только-только* и под. см. [11]) и ситуации зависимой клаузы. В (17), например, подчинительное употребление *когда* означало бы ориентацию момента повествования как следующего вскоре за ситуацией 'смена началась' и одновременного с ситуацией 'на участок заглянул десятник'. При сочинительном прочтении прагматически неестественного смысла не

⁶ Использовались следующие запросы: «V & indic & praet на расстоянии 1 от "было", -amark на расстоянии 1 до 6 от *когда*, asomta; V & indic & praet на расстоянии 1 от "-было", -amark на расстоянии 1 до 6 от *когда*, asomta». В выборку без *было* вошли примеры с *когда*, вводящим постпозитивную клаузу, независимо от грамматического статуса последней (т.е. адвербиальные, относительные, актантные и сочинительные), см. полный список примеров в приложении А. Примеры отфильтрованы вручную.

⁷ Еще один признак, позволяющий отличать сочинительное *когда* от подчинительного, – возможность замены сочинительного *когда* на сочетание *и в это время*, уместной не только в контексте маркеров нарушения нормального хода развития ситуации (где востребован *но*), но и в других контекстах (мы благодарны анонимному рецензенту за это наблюдение):

(i) *Ван запикивал в бак последние остатки мусора, когда (?но, *и в это время) в воротах с улицы появилась ладная фигура полицейского Кэнси Убукаты.* [13]

Возможность замены на *и в это время*, хотя и служит еще одним симптомом сочинительности *когда*, в обсуждаемых нами контекстах не играет специальной роли и поэтому подробнее не рассматривается.

возникает, поскольку нет ориентации ситуации, выраженной одной клаузой, относительно ситуации другой, а есть только указание на совпадение двух ситуаций во времени (в (17): момент повествования характеризуется как контактно следующий за ситуацией 'смена началась', при этом ситуация 'на участок заглянул десятник' представлена как наступившая в то же время).

Обратим внимание, что ассоциируемое с подчинительным *когда* прагматически неестественное прочтение возникает только при рематическом статусе придаточного (который, в свою очередь, больше характерен для постпозитивных клауз с *когда*). Так, в (18), с препозитивным подчинительным *когда* в статусе темы, временным ориентиром для ситуации 'Ланэ добрался до Гарднера' выступает ситуация 'Гарднер только что положил трубку' (в соответствии с общей тенденцией к совпадению ориентира с ремой, см. Введение). Тем самым, в отличие от примеров (16) и (17), здесь не происходит ориентации какой-либо ситуации относительно сразу двух точек на временной оси. Вместо этого ситуация, ориентированная во времени с помощью адвербиала *только что*, сама выступает в качестве временного ориентира.

(18) *Когда Ланэ добрался до Гарднера, тот только что положил трубку и что-то записывал на полоску бумаги.* [13]

Кроме того, и для постпозитивного рематического *когда* сочинительное прочтение в контексте адвербиалов контактного предшествования является только тенденцией, а не жестким правилом. В (19) постпозитивная клауза с *когда* выступает как временной ориентир для ситуации 'заливалась смехом', т.е. является подчиненной. Отмеченного выше прагматического противоречия не возникает здесь потому, что вводимая *когда* ситуация ('«запорожцу» пороли шину') произошла одновременно не с моментом повествования (как в (16) и (17)), а с моментом, выражаемым адвербиалом *только что*, т.е. незадолго до момента повествования. Другими словами, клауза с *когда* в (19) служит уточнением к адвербиалу. Напротив, в (16) и (17) клауза с *когда* семантически соотносится со всей лицензирующей клаузой. Это различие проявляется в том, что в (19) клауза с *когда* может быть вынесена в позицию после адвербиала с сохранением исходного значения, ср. (20), тогда как в (16) и (17) аналогичный вынос неуместен, ср. (21).

(19) *И все эти метаморфозы происходят на глазах его очаровательной спутницы, которая только что заливалась смехом, когда «запорожцу» пороли шины.* [13]

(20) *...которая только что, когда «запорожцу» пороли шины, заливалась смехом.*

(21) *...я только что, когда он был уже заметным критиком, взял в руки перо.*

Тем не менее, как и в предыдущем случае, тенденция к сочинительному прочтению *когда* в контексте адвербиала контактного предшествования является по данным НКРЯ статистически значимой (двусторонний вариант точного критерия Фишера, $P < 0,01$), см. таблицу 2. Подчеркнем, что в контексте адвербиала, как и в контексте маркера нарушения нормального хода развития ситуации, сочинительное и подчинительное прочтение *когда* ассоциируются каждое со своим набором свойств (ср. проиллюстрированную в (20) и (21) разную реакцию на вынос клаузы с *когда* влево).

	сочинение	подчинение	всего	доля сочинения
с <i>только что</i> / <i>только-только</i>	10	6	16	0,6
без <i>только что</i> / <i>только-только</i>	2	48	50	0,04

Таблица 2: Сопоставительная частотность сочинительного и подчинительного *когда* в контексте адвербиалов контактного предшествования и вне этого контекста (Подкорпус текстов 1951-2011 гг. Основного корпуса НКРЯ)⁸

⁸ Использовались следующие запросы: «только на расстоянии 1 от "что" на расстоянии 1 от v & indic & praet, -amark на расстоянии 1 до 8 от когда, асомта»; «только-только на расстоянии 1 от V & indic & praet, -amark на расстоянии 1 до 8 от когда, асомта»; «-только-только & -"что" на расстоянии 1 от v & indic & praet на расстоянии 1 до 8 от когда, асомта». Примеры отфильтрованы вручную. Полный список примеров см. в приложении Б.

3. Сочинительный *когда* по критериям сочинения и подчинения

Опираясь на введенные в предыдущем разделе дистрибутивные различия между сочинительным и подчинительным *когда*, продемонстрируем, что сочинительный *когда* ведет себя в соответствии с ожиданиями по некоторым критериям сочинения и подчинения. Ограничимся двумя критериями: 1) только подчиненная клауза допускает выделение посредством фокусных (иначе, контрастивных) частиц *только*, *именно* и *под*. [10: 15]; 2) только подчинительный союз может занимать в предложении начальную позицию [16: 262]. Последний критерий особенно показателен, поскольку, как считается, позволяет отличить сочиненные клаузы от адвербиальных слабо интегрированных [9: 167].

И в контексте маркеров типа *было* (22), и при адвербиале контактного предшествования (23) сочинительный *когда* не сочетается с контрастивными частицами с сохранением исходного значения. При этом элиминация *было* и *только что*, открывающая возможность для подчинительного прочтения *когда*, одновременно создает условия для контрастивного выделения придаточного, ср. (24) и (25).

(22) *Николай открыл было рот, чтобы сказать какую-то колкость, (??только, ??именно, ??как раз) когда в кухню вернулся Шоринов.* [13]

(23) *Тоня и Надежка только что расплясались, (??именно, ??как раз) когда Лилю крикнули домой.* [13]

(24) ^{OK}*Николай открыл рот, чтобы сказать какую-то колкость, именно (как раз) когда в кухню вернулся Шоринов.*

(25) ^{OK}*Тоня и Надежка расплясались, именно (как раз) когда Лилю крикнули домой.*

На то, что в (22) и (23) в самом деле фигурирует сочинительный *когда*, указывает его заменимость на *но* в (22) и неуместность выноса клаузы с *когда* в позицию после *только что* в (23), ср. (26) и (27) (см. об этих признаках раздел 2).

(26) (22) ≈ *Николай открыл было рот, чтобы сказать какую-то колкость, но в кухню вернулся Шоринов.*

(27) (23) ≠ *Тоня и Надежка только что, когда Лилю крикнули домой, расплясались.*

Начальная позиция в предложении для сочинительного *когда* в контексте маркеров типа *было* и адвербиалов контактного предшествования также недоступна: вынос такого *когда* в начало ведет к исчезновению значения, ассоциируемого с сочинением. Так, в (28) клауза с сочинительным *когда* (ср. заменимость на *но*) имплицитно вводит причину нарушения нормального хода развития ситуации: 'не встал, потому что Хрущев заговорил об «Оттепели»'. При выносе клаузы с *когда* в начало, как в (29), значение меняется: такое предложение требует продолжения, в котором сообщалось бы о причине нарушения нормального развития ситуации, а значит, в клаузе с *когда* эта причина больше не сообщается.

(28) *Деловая сторона разговора заняла четверть часа, и я хотел было встать, когда (^{OK}но) Хрущев заговорил о моей «Оттепели».* [13]

(29) *Когда Хрущев заговорил о моей «Оттепели», я хотел было встать (, но передумал).*

Вынос союза в начало предложения ведет к изменению значения и в контексте адвербиалов контактного предшествования. В (30), в отличие от (17), вводимая *когда* клауза совпадает с темой, так что значения, противоречащего подчинительной интерпретации *когда*, не возникает (ср. также обсуждение примера (18) в разделе 2).

(30) *Когда на участок раскладки заглянул десятник Мураш, смена только-только началась.*

4. Другие сочинительные временные союзы (вместо заключения)

Наряду с *когда*, к сочинительным есть основания относить союзы *как вдруг* и *как* (последний, как и *когда*, бывает сочинительным и подчинительным). Ср. в (31) употребление *как* в контексте частицы *было* и в (32) – употребление при адвербиале *только что*. В обоих случаях конструкция обнаруживает семантические и формальные признаки, ассоциируемые с сочинительным *когда*: в (31) *как* имплицитно вводит причину нарушения нормального хода развития ситуации ('почему не вернулся в кабинет') и заменим на *но*; в (32) неуместен вынос придаточного с *как* в позицию после адвербиала, ср. (33).

(31) *Он вместе с Ламзиной дошел до окошка секретаря и собрался было уже вернуться в кабинет, как (^{OK}но) его перехватил коллега, бывший однокурсник, жутко энергичный и активный по части «собратиться и посидеть».* [13]

- (32) *Утомившись, я только что прилег, как дверь в купе резко отъехала, и появилась Любка...* [13]
 (33) *...я только что, как дверь в купе резко отъехала, прилег.*

Высказанное в настоящей работе предположение о том, что в русском языке есть сочиненные временные клаузы, ставит вопрос о типологических параллелях этого явления. До сих пор в исследованиях по семантико-синтаксической интеграции клауз временные клаузы как будто оставались без пристального внимания (вероятно, ввиду их предполагаемой тесной интеграции) – в отличие, например, от причинных [3], [5] или уступительных [7].

Благодарности

Автор глубоко признателен М.А.Холодиловой за обсуждение некоторых фрагментов этой работы, а также анонимным рецензентам за советы и замечания. Исследование выполнено при поддержке гранта РФФ №22-18-00120.

References

- [1] Badan Linda, Haegeman Liliane. The syntax of peripheral adverbial clauses. – Journal of Linguistics, 2022. – P.1–42. DOI: 10.1017/S0022226721000463
- [2] Belošarkova V. A., Ju Xě Ken (1996), On one specific construction of temporal complex sentences [Ob odnoj specifičeskoj konstrukcii vremennyx složnopolčiněnyx predloženíj], Bulletin of Moscow State University, Philology series [Vestnik MGU, serija Filologija], Vol. 1, pp. 82-94.
- [3] Belyaev Oleg. Cause in Russian and the formal typology of coordination and subordination // Donum semanticum: Opera linguistica et logica in honorem Barbarae Partee a discipulis amicisque Rossicis oblata. – Moscow, 2015. – P. 35–66.
- [4] Declerck Renaat. WHEN-Clauses and Temporal Structure. – London, 1997.
- [5] Frey Werner. On some correlations between formal and interpretative properties of causal clauses // Co- and subordination in German and other languages: Special issue of Linguistische Berichte 21. – Hamburg, Buske, 2016. – P.153–179.
- [6] Frey Werner. On the categorial status of different dependent clauses – Ms., ZAS Leibniz-Centre General Linguistics, Berlin, 2020.
- [7] Frey Werner. German concessives as TPs, JPs and ActPs. – Glossa: A Journal of General Linguistics, 2020. – Vol. 5(1).
- [8] Haegeman Liliane. The syntax of adverbial clauses and its consequences for topicalization // Current Studies in Comparative Romance Linguistics [APiL 107]. – Antwerp, 2004. – P. 61–90.
- [9] Haegeman Liliane. Adverbial clauses, main clause phenomena and the composition of the left periphery: the cartography of syntactic structures. – Oxford Studies in Comparative Syntax, 8. – 2012.
- [10] Haspelmath Martin. The converb as a cross-linguistically valid category // Converbs in cross-linguistic perspective (Empirical Approaches to Language Typology 13). – Berlin, 1995. – P.1–55.
- [11] Letučij A. B. (2017), Russian adverbials with temporal meaning *tol'ko*, *tol'ko-tol'ko*, and *tol'ko čto*: recent past meaning and resultative semantics [Russkie narečija so značeniem vremeni *tol'ko*, *tol'ko-tol'ko*, *tol'ko čto*: značenie nedavnego prošlogo i rezul'tativnoe značenie], Issues of Linguistics [Voprosy Jazykoznanija], Vol. 1, pp. 53-73.
- [12] Pekelis O. E. (2009), Coordination and subordination in the context of causal semantics [Sočinenie i podčinenie v kontekste pričinnoj semantiki], Moscow: Russian State University for the Humanities, Cand. thesis.
- [13] Russian National Corpus [Nacional'nyi korpus russkogo jazyka] (2003–2022), available at: <http://www.ruscorpora.ru>.
- [14] Sičinava D. V. (2013), Particles *bylo* and *byvalo*: Russian “secondary modifiers” in the light of typology and diachrony [Časticy *bylo* i *byvalo*: russkie “vtoričnye modificatory” v svete tipologii i diaxronii], Studies in the typology of Slavic, Baltic and Balkan Languages (with primary reference to language contact) [Issledovanija po tipologii slavjanskix, baltijskix i balkanskix jazykov (preimuščestvenno v svete jazykovyx kontaktov)], Saint-Petersburg, pp. 175-194.
- [15] Švedova N. Yu. (1980), Russian Grammar [Russkaja grammatika], Moscow, Vol. 2.
- [16] Testeleť Y. G. (2001), Introduction to General Syntax [Vvedenie v obščij sintaksis], Moscow.
- [17] Uryson E. V. (2011), An attempt at describing the conjunctions' semantics: linguistic data of consciousness activity [Opyt opisanija semantiki sojuzov: lingvisticheskie dannye o dejatel'nosti soznanija], Moscow.
- [18] Yanko T. E. (2001), Communicative strategies of Russian speech [Kommunikativnye strategii ruskoj reči], Moscow.

Приложение А. Выборка примеров с частицей *было* и без нее (см. таблицу 1).

Примеры с *было*

Сочинение

- 1) *Намерзшись за день на сыром ветру вперемежку с дождем, я уже начал было задремывать, когда раздалось со вздохом:* [Г. Я. Бакланов. Мой генерал // «Знамя», 1999]
- 2) *Он стоял спиной к Насте, и она собралась было уже незаметно выйти из кухни, когда ее остановил голос Барина:* [Александра Маринина. Мужские игры (1997)]
- 3) *Отметив про себя слово «правительство» и записав его в нужном месте на схеме, Настя собралась было включить кипяtilьник, чтобы приготовить кофе, когда в ее кабинет ворвался Юра Коротков, который с утра пораньше отправился «погулять» вокруг официального места жительства Евгения Парыгина.* [Александра Маринина. Мужские игры (1997)]
- 4) *Михаил хотел было подойти к нему, когда дверь реанимобиля открылась и выглянула врач.* [Александра Маринина. Мужские игры (1997)]
- 5) *Он видел, как девушка обрадовалась, и собрался было уже задать следующий вопрос, когда дверь распахнулась и в комнату вошел Василий.* [Александра Маринина. Иллюзия греха (1996)]
- 6) *После «летучки» она вышла из редакции и направилась было к метро, когда к ней подошел импозантный полноватый мужчина в затемненных очках и с длинными кудрявыми волосами.* [Александра Маринина. Не мешайте палачу (1996)]
- 7) *Следователь уже начал было терять терпение, когда поступило сообщение от начальника оперчасти следственного изолятора: контролер, заглянув в камеру, где сидел Асатурян, застал там странную картину.* [Александра Маринина. Не мешайте палачу (1996)]
- 8) *Антон Андреевич Минаев снял костюм, переоделся в спортивные брюки и свитер и собрался было приступить к ужину, когда раздался звонок в дверь.* [Александра Маринина. Не мешайте палачу (1996)]
- 9) *Он расслабленно откинулся на сиденье и уже приготовился было в очередной раз закурить, когда увидел неторопливо бредущих в его сторону троих молодых людей, лица которых были ему хорошо знакомы.* [Александра Маринина. Светлый лик смерти (1996)]
- 10) *Николай открыл было рот, чтобы сказать какую-то колкость, когда в кухню вернулся Шоринов.* [Александра Маринина. За все надо платить (1995)]
- 11) *Спустился вниз, сел в машину и уже собрался было отъезжать, когда увидел, как из гастронома напротив выходит Катя в своей белоснежной длинной куртке, неся в руках две полиэтиленовые сумки с продуктами.* [Александра Маринина. За все надо платить (1995)]
- 12) *Не добившись от нее никакого толку, я уже собрался было попрощаться, когда заметил среди вывешенного на просушку белья две белые футболки, которые по размеру никак не могли принадлежать двадцативосьмилетнему молодому мужчине.* [Александра Маринина. Черный список (1995)]
- 13) *Захаров завел мотор и уже начал было трогаться, когда заметил женщину, медленно бредущую со стороны зоны прилета.* [Александра Маринина. Стечение обстоятельств (1992)]
- 14) *Человек опечалился и хотел было уже уходить, когда я сказал:* [Анатолий Жигулин. Черные камни (1988)]
- 15) *Антипов уже поплелся было к двери, когда навстречу быстрыми шагами влетели Таня и Эллочка.* [Ю. В. Трифонов. Время и место (1980)]
- 16) *Молодой человек был заметно огорчен этим обстоятельством и совсем уже собрался было уходить, когда в голову ему пришла счастливая мысль.* [Аркадий Стругацкий, Борис Стругацкий. Жук в муравейнике (1979)]
- 17) *Звягинцев вышел из машины и собирался было уже захлопнуть дверцу, когда снова раздался голос шофера.* [Александр Чаковский. Блокада (1968)]
- 18) *Деловая сторона разговора заняла четверть часа, и я хотел было встать, когда Хрущев заговорил о моей «Оттепели».* [И. Г. Эренбург. Люди, годы, жизнь. Книга 7 (1960-1965)]
- 19) *От скуки и шума дождя я уже начал было дремать, когда вдруг зазвонил телефон.* [Аркадий Стругацкий, Борис Стругацкий. Понедельник начинается в субботу (1964)]

20) *Маша уже пошла было к двери, когда Шмелев остановил ее.* [Константин Симонов. Живые и мертвые (1955-1959)]

21) *<...> а Петр Горбидоныч поднял руку, прося внимания для одного сверхсрочного заявления. Он напустил было на лицо шутливое выражение, когда слово самовольно перехватила одна, в косынке, провинциальная старушка, тетка именинницы, давно порывавшаяся завести душевный разговор.* [Л. М. Леонов. Вор. Части 1-2 (1927, 1959)]

22) *Наверно, к этому вопросу еще вернутся Савченко собрался было уходить, когда вдруг сказал:* [И. Г. Эренбург. Оттепель (1953-1955)]

23) *В вагоне я стал было писать статью для «Дела народа», когда взгляд упал на анилаг газеты у соседа.* [М. В. Вишняк. Дань прошлому (1953)]

24) *Товарищ Медведев собрался было издать призыв о помощи, когда, повернув лицо в сторону выхода из расщелины, он услышал ружейную стрельбу и увидел зрелище, которое преисполнило его чувством искреннего недоумения.* [И. Л. Солоневич. Две силы (1953)]

25) *Мы только наладились было обедать, когда они вошли.* [Константин Воробьев. Друг мой Момич (1965)]

26) *Ей стало немного легче, и она уже собралась было тоже задремать, когда Старыгин, не открывая глаз, проговорил:* [Наталья Александрова. Последний ученик да Винчи (2010)]

27) *Под этот мерный бубнеж Дима уже начал было снова засыпать, когда толстяк неожиданно приблизил свое круглое лицо прямо к его уху и, дохнув на Диму гнилым фруктовым теплом, тихо скомандовал:* [А. А. Старобинец. Семья (2008)]

28) *Ваксон приготовился было сосчитать и его «отпоры», когда до него дошла некоторая странность.* [Василий Аксенов. Таинственная страсть (2007)]

29) *Пребывая некоторое время в растерзанных чувствах, Ян решил было отправиться в город пострадать одиночеством, когда вдруг увидел, что из такси возле главного трапа «Собесского» выскакивают быстроногие две его жены, бывшая, Нэллочка Аххо, и пока что нынешняя Татка Фалькон, обе в случайной одежде, однако с большим общим чемоданом нарядов.* [Василий Аксенов. Таинственная страсть (2007)]

30) *Справившись с этим делом, главный редактор направился было назад, когда я, собрав последние силы, распахнула пассажирскую дверь его автомобиля.* [Татьяна Сахарова. Добрая фея с острыми зубками (2005)]

Подчинение

1) *Вьетнам повёл было речь о создании металлургического завода, когда его населению есть было нечего.* [«Проблемы Дальнего Востока», 2002.12.30]

2) *Поскольку этим вопросом я также пристально интересовался, то ничуть не удивлялся конкурсу в ВУЗы, рост которого начался было года три назад, когда в ВУЗы начали поступать дети 1980-х годов рождения.* [Интернет-альманах «Лебедь», 2003.06.02]

3) *Первушин назначил было обеденное время, когда в здании никого не оставалось, но тут капитан, моментально перестав быть любезным, сказал, что приедет к десяти, и положил трубку.* [Татьяна Устинова. Большое зло и мелкие пакости (2003)]

4) *Что-то почудилось было, когда я попробовал стихи, дальше пока не пошло.* [М. С. Харитонов. Стенография конца века. Из дневниковых записей (2002)]

5) *Правда, некоторое беспокойство вызвала было старая соперница Берже, когда в 1968 году после прохождения спелеоаквалангистами парижского клуба на Сене донных сифонов пропасти неожиданно подступила почти вплотную: - 1141 метр.* [Константин Серафимов. Экспедиция во мрак (1978-1996)]

6) *Граве повеселел было, когда узнал, что сибирское, уральское и самарское правительства уступили свою власть омской Директории.* [А. И. Алдан-Семенов. Красные и белые (1966-1973)]

7) *По привычке взялся было за очки, когда я вошел.* [Георгий Владимов. Три минуты молчания (1969)]

8) *Первый раз за реку Пелагея отправилась было еще в феврале, когда впервые после долгой метели заледенелые окошки вызолотило красное солнышко.* [Федор Абрамов. Пелагея (1967)]

9) *Ломакин уже решил было бросить затею, когда по четвертому разу прокатился до «Рыбацкого» и обратно до «Приморской».* [Андрей Измайлов. Трюкач (2001)]

10) *Пыльный начал было протестующее движение, когда он ухватил Анатолича левой за твидовый ворот, а правой за брючный ремень — но так и не довел его даже до середины.* [Александр Гаррос, Алексей Евдокимов. [Голово]ломка (2001)]

Примеры без было

Сочинение

1) *Ван закинул в бак последние остатки мусора, когда в воротах с улицы появилась ладная фигура полицейского Кэнси Убукаты.* [Аркадий Стругацкий, Борис Стругацкий. Град обреченный (1972)]

Подчинение

1) *Вениамин напомнил нам словосочетание «Содом и Гоморра», когда мы проезжали мимо состоящей из соли горы Содом.* [Т. Н. Ткаченко. Дневник 1995 г. (Болгария, Греция, Израиль, Египет, Турция) (1995)]

2) *О здешних церквах писал Василий Розанов в очерке «Русский Нил», когда ходил парходом по Волге, что только в Москве есть такие прекрасные церкви, как в Романове-Борисоглебске:* [«Наука и религия», 2010]

3) *Правовая оценка действий Ниязова прозвучала только 9 июня, когда директор департамента консульской службы МИД Владимир Котенев заявил, что Россия не признает односторонний выход Туркмении из соглашения о двойном гражданстве.* [«Профиль», 2003.06.23]

4) *Ей было двадцать восемь лет, когда родился Юра.* [Г. Я. Бакланов. Жизнь, подаренная дважды (1999)]

5) *Переломным стал 1960 год, когда, по словам Лаврова, произошли три знаменательных события:* [«Computerworld», 2004]

6) *У нас в поселке на улицах темно, фонари поставили еще в позапрошлом году, митинг был, когда их зажгли впервые, но теперь приходится экономить электроэнергию.* [И. Меттер. Свободная тема (1979)]

7) *Много лет тому назад мы провели там все жаркие дни лета, когда на побережье плавится асфальт.* [Семен Лунгин. Виденное наяву (1989-1996)]

8) *Впрочем, холодность первой встречи искупила следующая: радостные вопли наполнили квартиру, когда он заглянул в гостиную.* [Семен Данилюк. Бизнес-класс (2003)]

9) *Отец уехал из Гори в Тифлис, когда сыну исполнилось пять лет, и появлялся редко.* [«Звезда», 2003]

10) *Шумиха вокруг «Большого Приза» была сродни той, что потрясла мир в конце 90-х, когда на экраны вышел «Титаник» Джеймса Кэмерона.* [«Формула», 2002.02.15]

11) *Было страшно представить, что произойдет, когда Ахмед узнает истинную причину гибели Джохара.* [Александр Михайлов. Капкан для одинокого волка (2001)]

12) *Все ей стало ясно, когда он спросил про Геннадия.* [В. Ф. Панова. Времена года. Из летописей города Энска (1953)]

13) *Это спасло его, когда одна из дьярв ринулась прямо на рубку.* [Алексей Иванов. Корабли и Галактика (1990-1991)]

14) — *попросила я седоусого, когда тот поставил передо мной пластиковый стаканчик с кофе, солонку, перечницу и блюдо с круглой упаковочкой сливок.* [«Русский репортер», № 1-2 (080-081), 22-29 января 2009]

15) *Опыт применения Т-72 в войне с Ираном свидетельствует, что его «лоб» выдерживал удар таких снарядов на меньшем расстоянии, когда их поражающее действие куда значительнее.* [«Техника - молодежи», 1991]

16) *Кто из нас не сидел с дорогим ему человеком — матерью, женой, мужем, другом — в вечерний час, когда спускались сумерки, когда все затихало вокруг.* [митрополит Антоний (Блум). Таинство любви (1971-1986)]

17) *Было время, когда человек пробегал марафонскую дистанцию за три с половиной часа, и это вызывало всеобщее восхищение.* [В. А. Солоухин. Смех за левым плечом (1989)]

18) *Он оглянулся всего один раз, **когда** за спиной, громко разговаривая, прошли два офицера с серебряными эдельвейсами на беретах.* [Сергей Наумов. На расстоянии крика (1978)]

19) *Верочка питалась вроде бы самостоятельно с Николаем Эрастовичем, **когда** тот приезжал, а если его не было, то вроде бы со свояченицей, теткой Любой.* [Ю. В. Трифонов. Старик (1978)]

20) *Я поставил планку на триста пятьдесят метров и несколько раз выстрелил по этому котелку, **когда** немец лежал в картошке.* [«Новый Мир», 1998]

21) *Узор был очень похож на орден Александра Невского — Михаська видел его у одного раненого в госпитале, **когда** они выступали там с шефским концертом.* [Альберт Лиханов. Чистые камушки (1967)]

22) *Бывали случаи, **когда** люди принимали это за пожар, невесть откуда идущий по воздуху.* [В. А. Мезенцев. Чудеса: Популярная энциклопедия. Том 1 (1991)]

23) *Пережил он и тридцатые годы массовых репрессий, **когда** свобода и сама жизнь находились под угрозой.* [Олег Трояновский. Через годы и расстояния (1997)]

24) *Пока шел отпуск, все считал дни, **когда** он закончится и можно будет наконец окунуться в новую работу.* [Юрий Никулин. Клоуна надо видеть (1979)]

25) *Вот что получил СССР, **когда** уничтожил всю технику, всю авиацию, когда все специалисты были в лагерях.* [В. В. Бибихин. Алексей Федорович Лосев (1975-1977)]

26) *Этот замечательный огонь я ощущал и ощущаю всегда, **когда** мысли осеняют меня или чувство заговорит.* [А. Л. Чижевский. Вся жизнь (1959-1961)]

27) *<...> начальник департамента по связям с общественностью «Сибнефти», а теперь и вице-президент этой компании Владимир Руга решил вспомнить то теперь уже далекое время, когда он был нашим братом журналистом.* [«Известия», 2002.03.04]

28) *И Жан спал с лица, когда Микки впервые стремглав на него помчался.* [Интернет-альманах «Лебедь», 2003.10.26]

29) *Да и не приняты были в те времена, в конце шестидесятых, **когда** я выступала, всякого рода амулеты.* [Виталий Мелик-Карамов, Татьяна Тарасова. Красавица и чудовище (1984-2001)]

30) *Нечто подобное ощущал Вайс в моменты, **когда** ему приходилось быть свидетелем неутраченной борьбы между различными германскими разведывательными службами за полноту власти.* [Вадим Кожевников. Щит и меч. Книга вторая (1968)]

31) *Я был еще юношей, лет двадцати, **когда** впервые встретился с ним, в то время тоже еще молодым человеком и писателем, только что замеченным.* [Н. Д. Телешов. А. П. Чехов (1904-1956)]

32) *И ей отвечали, что она была предупреждена, **когда** записывалась, что все имущество теперь общее, городское, она внесла свой пай, но забрать его права не имеет, а получит, как все уходящие, денежный эквивалент полугодового содержания и два комплекта одежды для себя и для ребенка...* [Андрей Лазарчук. Там вдали, за рекой... (1986)]

33) — *уже выехала откуда-то оттуда, где таинственно обитает, **когда** исчезает из нашего дома.* [Дина Рубина. Я и ты под персиковыми облаками (2001)]

34) — *спросил я, **когда** мы выскочили на Рязанский.* [Андрей Геласимов. Год обмана (2003)]

35) *Пятнадцать лет он жил почти в нищете, и не забылось еще время, **когда** остатки макарон, сваренных на ужин, служили завтраком — если обжарить с луком и яйцом (и весьма, если вспомнить, вкусное блюдо! [«Волга», 2010]*

36) *Сегодня все прибрал, поправил в том лучшем настроении, **когда** гляди, что не перед добром.* [«Знамя», 2005]

37) *Мне оставалось одно: следить, чтобы девочка осторожно обращалась с Яшкой, ведь он был слепой и пугался, **когда** его внезапно хватили маленькие неосторожные ручки.* [«Уральская новь», 2003]

38) *Собственно, первые, настоящие уроки я получила от него, **когда**, скитаясь по чужим квартирам, мы засиживались далеко за полночь, и муж — с энтузиазмом, заставляющим сожалеть о малочисленности аудитории, представленной одним-единственным слушателем, — пускался в долгие рассуждения о русской прозе и поэзии.* [«Звезда», 2002]

39) *Он не оттолкнул Митьку, **когда** тот, опустошенный и отверженный, постучался к нему однажды ночью.* [Л. М. Леонов. Вор. Части 1-2 (1927, 1959)]

40) *Мысль посетила меня, когда я рылся в старой почте, составляя опись ожидаемых посылок.* [коллективный. Форум: Книга жалоб и предложений (2004-2006)]

41) *Саламату я встретил впервые, когда ей было девятнадцать лет.* [Евгений Пермяк. Бабушкины кружева (1955-1965)]

42) *Само происшествие застало Голубева в тот момент, когда он вместе с одноруким счетоводом Волковым проверял бабу Дуню на предмет самогоноварения.* [Владимир Войнович. Жизнь и необычайные приключения солдата Ивана Чонкина (1969-1975)]

43) *Слушал я Вас, читал текст выступления и вспоминал 1956 год, когда я работал еще недалеко от Ровеньков, в Краснодоне, учительствовал в тамошней школе и (параллельно) был сотрудником местной газеты «Социалистическая Родина».* [«Огонек», 1989]

44) *Совершенно естественно и логично было, что те, с кем он с детства оказался в таких простецких отношениях и регулярно перевидывался в следующие два десятилетия горячей гульбы, бань, дач, балтийских и черноморских пляжей, заведенной — еще родителями, но и ими тоже, от себя, по своему вкусу — привлекательной жизни на территории в каких-нибудь полквадратных километра вокруг Чистых прудов, и кто, как и он, никуда оттуда не переехал, не уехал, а сталкивался с ним и друг с другом на улице, сплошь и рядом даже не останавливаясь поговорить, а только перебрасываясь на ходу веселыми словами давно, сообщая, незаметно начатой болтовни, приняли его, когда компания перешла в статус клуба, нигде, понятное дело, не зарегистрированного, никак не объявляемого, клуба метафизического, ощущаемого лишь психологически, если хотите, воображаемого, в действительные члены.* [«Октябрь», 2003]

45) *Нам же давали возможность оглядеться, когда мы еще были юными и роскошными.* [Михаил Анчаров. Теория невероятности (1965)]

46) *«Амазонка» Дурова, сбежавшая из отчего дома с намерением посвятить себя воинскому делу, оказалась в Гродно свидетельницей так называемого вербунка, когда молодых людей песнями и плясками зазывали в компанию улан в корчме и на улице.* [«Родина», 2008]

47) *Историю этого поместья, расположенного в живописном уголке провинции Тоскана, в краю виноградников, садов и оливковых роц — настоящей жемчужины ландшафтной архитектуры, — можно проследить с начала XVI века, когда по распоряжению первых ее владельцев — семейства Буонвизи — была не только осуществлена постройка виллы, но и заложен парк в классическом стиле эпохи итальянского Ренессанса.* [«Ландшафтный дизайн», 2003.01.15]

48) *Я как-никак помнил времена, когда на самолётах мы в Японию ещё не летали, а плыли (или, как говорят моряки) от Владивостока ...* [И. Э. Кио. Иллюзии без иллюзий (1995-1999)]

49) *Так что в этом отношении апостол Павел выражал что-то очень понятное, когда говорил: Кто меня освободит от этого тела тления, тела, которое меня тянет вниз?* [митрополит Антоний (Блум). О смерти (1985-1995)]

Приложение Б. Выборка примеров с адвербиалом контактного предшествования и без него (см. таблицу 2).

Примеры с адвербиалом

Сочинение

- 1) *А. П. только-только начинал после академии службу на Дальнем Востоке, **когда** в Москве на печально знаменитом пленуме ЦК ВКП (б), где руководители наркоматов, областей и республик, с энтузиазмом оперируя трехзначными цифрами, докладывали о количестве выявленных ими и уничтоженных врагов, нарком обороны Ворошилов по недомыслию (неужели от смелости?! или пытался выгородить себя?): «В Красной армии врагов вообще не очень много...» Ах, как его вразумили!* [«Звезда», 2001]
- 2) *Смена только-только началась, **когда** на участок раскладки заглянул десятник Мураш.* [Евгений Лукин. Катали мы ваше солнце (1997)]
- 3) *Жизнь только-только начала входить в нормальную колею, **когда** в последней декаде августа стихия нанесла Сахалину новый удар.* [«Природа и человек», 1983]
- 4) *Артист только что вернулся в Москву из гастрольной поездки, **когда** позвонили из Кремля и сообщили, что с ним хочет говорить Сталин и что за ним уже послана машина.* [Борис Ефимов. Десять десятилетий (2000)]
- 5) *Я сравнительно поздний современник Чуковского — я только что взял в руки перо, **когда** он был уже заметным критиком и основателем новой детской поэзии.* [В. А. Каверин. К. И. Чуковский. Дневник 1901-1921. Предисловие (1988)]
- 6) *С ощущением опасной неопределенности нашей победы я вернулся из гимназии, поужинал и только что собрался приняться за чтение, **когда** в окно моей комнаты постучала Валя.* [В. А. Каверин. Освещенные окна (1974-1976)]
- 7) *Тоня и Надежка только что расплясались, **когда** Лилю крикнули домой.* [Василий Белов. Привычное дело (1967)]
- 8) *Только что наступила темнота в роце, **когда** Коста пришел на отмеченное место, где он заломил ветку на дубе.* [А. П. Ладинский. Последний путь Владимира Мономаха (1960)]
- 9) *Вера только что притащила из овощного огромный арбуз и помыла его, **когда** в дверь позвонили.* [Марина Вишневецкая. Есть ли кофе после смерти? (1999)]
- 10) *Ему только что позвонили по телефону, он успел подать первую реплику, **когда** в дверь ворвался гул мотора.* [А. А. Бек. Талант (Жизнь Бережкова) / Части 4-6 (1940-1956)]

Подчинение

- 1) *Только-только начался отсчет — с утра, **когда** меня разбудил волшебник из Санкт-Петербурга, а фарт идет и идет...* [«Знамя», 2001]
- 2) *И все эти метаморфозы происходят на глазах его очаровательной спутницы, которая только что заливалась смехом, **когда** «запорожцу» поролли шины.* [«Столица», 1997.05.27]
- 3) *Я думала, что умру сама от невозвратной утраты, которую я только что перенесла неделю назад, **когда** умер мой лучший друг!..* [Юрий Безелянский. В садах любви (1993)]
- 4) *Она ушла на кухню, а Лопатин, вывинтив лампочку и на ощупь пройдя к себе в кабинет, свинтил ее в стоявшую на столе черную пластмассовую настольную лампу, которые только что появились в магазинах в тридцать восьмом году, **когда** они вдруг получили эту квартиру.* [Четыре шага (1956-1965)]
- 5) *Только что миновал тысячный год, **когда** погрязшие в грехах народы и за десять веков не успевшие подготовиться к царству небесному, с трепетом ждали конца мира.* [А. П. Ладинский. Анна Ярославна — королева Франции (1960)]
- 6) *Брови сдвинуты; глаза, которые только что сверкали, **когда** он боролся у доски за свою конструкцию, глядят куда-то вниз.* [А. А. Бек. Талант (Жизнь Бережкова) / Части 1-3 (1940-1956)]

Примеры без адвербиала

Сочинение

1) *Предвкушая будущий свой разговор с важными людьми и отбирая для него наиболее убедительные аргументы, В. В. из спокойного состояния переходил в возбужденное, размахивал руками, бормотал что-то себе под нос, когда в груди у него, и не слева, а ровно посредине, возникло и стало нарастать непонятное жжение с одновременной отдачей в локти и одеревенением губ.* [Владимир Войнович. Замысел (1999)]

2) *Словом, налаживалась серьезная и положительная беседа двух мужчин, когда вдруг из квартиры раздался голос матери:* [Лазарь Лагин. Старик Хоттабыч (1955)]

Подчинение

1) *Никто поэтому не удивился, когда в стойбище Улак он получил самую почетную должность — стал заведовать Магазином.* [Юрий Рытхэу. Время таяния снегов (1967)]

2) *Я был поражен, когда вдруг понял, что мы плывем по улицам, — писал американский писатель Генри Миллер, лишь недавно переведенный на русский язык.* [«Вокруг света», 1994]

3) *Стригуций лишай бабушка видела своими глазами у одного мальчика, когда жила в эвакуации.* [М.С. Аромштам. Мохнатый ребенок (2010)]

4) *Ахматова писала, что не верит, когда в воспоминаниях появляется прямая речь.* [А. С. Демидова. Бегущая строка памяти (2000)]

5) *Учитывая ограниченность экранного времени, мы остановились на том времени, когда Кантемир возвратился в Молдавию и стал во главе национально-освободительного движения.* [«Советский экран», 1973]

6) *Возраст свой я стала чувствовать недавно, когда обнаружила, что большинство моих постоянных собеседников моложе меня на двадцать, тридцать, сорок лет.* [Р. М. Фрумкина. О нас – наискосок (1995)]

7) *Глаза Лепсуса положительно горели, как уличный газ, когда он пробормотал себе под нос:* [М. С. Шагинян. Месс-Менд, или Янки в Петрограде (1923-1924 , 1954)]

8) *Как все мои одноклассники и друзья, я привычно отмахивалась, когда в жизнь вторгалась какая бы то ни было сложность.* [Л. Р. Кабо. Ровесники Октября (1964, 1997)]

9) *Первую книгу о театре я прочел, когда мне было тринадцать лет.* [Владлен Давыдов. Театр моей мечты (2004)]

10) *— наконец вырвалось у него, когда я пробурчал что-то недовольное.* [Андрей Курков. Милый друг, товарищ покойника (2001)]

11) *Дядя заранее узнавал о том, когда предполагается фестиваль, и скрывался на свою дачку в Финляндии, в Мецекюля (ныне Роцино).* [Н. И. Ильина. Дороги и судьбы (1957-1985)]

12) *У молодого Сократа кружилась голова, когда приехавший в Афины Парменид развернул перед ним антиномию бытия и небытия, единого и многого.* [В. В. Биbihин. Язык философии (1993)]

13) *По ее убеждению, свою знаменитую истину древний философ мог открыть лишь в детстве, когда босыми ногами бродил по гальке древнегреческого ручейка, а вот она, Поля, сколько ни бродила по лесу, нарочно забираясь в дебри погуще, ничего путного пока не изобрела.* [Л. М. Леонов. Русский лес (1950-1953)]

14) *Из семейных преданий о деду я помню рассказ бабушки, что дед потому и заболел чахоткой, что полез в Куру спасать ее, когда они были с ним где-то на Кавказе.* [Муслим Магомаев. Любовь моя — мелодия (1999)]

15) *Это было в восьмом классе, когда мальчишки из параллельного класса, увидев, как мы переходим дорогу, держась за руки, начали его дразнить.* [Брак фиктивный — дети настоящие // «Истории из жизни», 2004]

16) *Нет, нет, даже к сердечной и наивной Кире, начальнице добрейшей сегодня лезть с такими вот импровизациями тухлыми вне всякого сомненья дело недостойное, саму возможность простой спасительной отмазки заказала лапе нашей высокомерной пара мерзких глаз, что сузились, зажглись, уперлись взглядом в спину вечером вчера, тогда, когда стояла Лера Додд в тумане радужном перед зеркальной стеной кафе с ужасной репутацией <...>* [Сергей Солоух. Клуб одиноких сердец унтера Пришибеева (1991-1995)]

17) *Надав и Авиуд умерли в Синайской пустыне, пред Господом, **когда** принесли для служения Господу чуждый огонь; детей у них не было.* [Библия. Современный русский перевод Книги Священного Писания Ветхого и Нового Завета Канонические. Ветхий завет. Четвертая книга Моисея. Числа (2011)]

18) *Соблазн возник опять и оказался непреодолимым, **когда** я недавно прочел роман А. Бека «Новое назначение» и комментарий к нему доктора экономических наук Г. Попова «С точки зрения экономиста».* [А. Ефимов. Элитные группы, их возникновение и эволюция // «Знание — сила», 1988]

19) *Думается, что изучение мирового опыта не должно сопровождаться забвением своего отечественного, хотя он и был накоплен в советское время, **когда** действовала административно-командная система управления.* [«Финансы и кредит», 2003.02.03]

20) *И я протирала лицо «Облепиховым молочком», которое покупалось в магазине ВТО, **когда** в продаже еще не было ничего, нигде, никогда.* [Яна Зубцова. Алика Смахова без полутонов // «Домовой», 2002.11.04]

21) *То были несуетные времена, **когда** человек и не подозревал, что может настать время, чреватое не библейским, а вполне реальным рукотворным концом света.* [«Семья», 2000.01.19]

22) *Грабор чувствовал, **когда** Лизонька пьяна, а когда не просто пьяна.* [«Урал», 2002]

23) *Мы это обсуждали с Шурой Выгодским, **когда** он с Волей Римским навещал нас с Ниной на Зеленом озере, и пришли к заключению, что определенно наметился поворот к лучшему.* [И. М. Дьяконов. Книга воспоминаний. Глава одиннадцатая (1937-1939) (1995)]

24) *Кстати я заметила за собой, убираюсь чаще, **когда** кто-то дома.. т. е. когда одна, люблю посидеть в тишине, побездельничать в волю.* [Наши дети: Дошколята и младшие школьники (форум) (2005)]

25) — *Я взял эти крайности, — заключил Колдуэлл, — чтобы показать, что писал о той жизни, которая была в то время, **когда** я о ней писал.* [С. А. Дангулов. Колдуэлл (1981)]

26) *Воздух густел и застыл в ожидании близкого заката, фигурка неумолимо приближалась, и наступал тот миг, **когда** Владимир Иванович опускался на песок рядом с подругами.* [Андрей Дмитриев. Штиль (1983)]

27) *Тугошние милицейские тоже пришли к подобному предположению, **когда** сынок Кудимова получил шило по рукоятку в спину.* [Андрей Измайлов. Трюкач (2001)]

28) *У нас друзья только через 10 лет брака венчались, **когда** уже все было проверено и пройдено.* [Женщина + мужчина: Брак (форум) (2004)]

29) *<...>и скучала, **когда** он затевал с Сережей игру в городки.* [Н. М. Гершензон-Чегодаева. Воспоминания дочери (1952-1971)]

30) *Так получалось каждый раз: Незнайка совершал хорошие поступки только тогда, **когда** вспоминал о волшебной палочке.* [Николай Носов. Незнайка в Солнечном городе (1958)]

31) *На душе было то же чувство, что и в августе, **когда** по всем церквам звонили колокола, а на другой день в семьях прощались с сыновьями и внуками, давали на дорогу серебряные рубли и плакали: «Сыночек мой!* [В. Лихоносов. Ненаписанные воспоминания. Наш маленький Париж. Ч. 3-4 (1983)]

32) *Я уловил этот принцип, это организующее начало, **когда** попробовал вспомнить, о чем я думал только что, о чем думал перед этим и перед тем, как думал перед этим.* [Борис Хазанов. Далекое зрелище лесов (1998)]

33) *А Чуковскому в детстве мать сказала, **когда** он потерял ее рубль: «Что ж, подумай, как обрадуется тот, кто его найдет»* (Дн. [М. Л. Гаспаров. Записи и выписки (2001)]

34) *В общем, я даже любила чистить ножи, мне нравилось, **когда** они начинали блестеть.* [В. А. Каверин. Открытая книга (1949-1956)]

35) *С этого лета мы встречались регулярно до того рокового случая, **когда** Ландау перестал быть Ландау.* [Давид Самойлов. Проза поэта (1970-1980)]

36) *Она старалась появляться в те дни, **когда** Ижмайлов отправлялся в свои ответственные командировки.* [Василий Аксенов. Новый сладостный стиль (1997)]

37) — *Я с детства боялся смерти и более-менее свыкся с тем, что она неизбежна, **когда** умерли родители.* [«Дело» (Самара), 2002.06.25]

38) — *я говорил ей эти слова, **когда** ничего другого нельзя было придумать.* [Константин Воробьев. Вот пришел великан (1971)]

- 39) *За это время мы с подругой виделись три или четыре раза, **когда** она приезжала на встречи выпускников.* [«Истории из жизни», 2004]
- 40) *Они стали весьма содержательными, **когда** эти народы стали жить в едином государстве в российском царстве, ставшем империей, а затем в СССР и теперь в Российской Федерации.* [«Жизнь национальностей», 2000.09.29]
- 41) *А пошло, Пармеша, с сегодняшнего утра, **когда** мы с тобой посуду пустую сдавать повезли.* [Василий Белов. Привычное дело (1967)]
- 42) *Это он рассказал в конце ноября, **когда** был у меня в доме.* [Владимир Личутин. Любостай (1987)]
- 43) *Вот и настало время, **когда** некому рассказать про мою маму.* [Александр Терехов. Каменный мост (1997-2008)]
- 44) *Это было вызвано распадом СССР, спадом в экономике страны, **когда** многие предприятия утратили оборотные средства, выросли железнодорожные тарифы, цены на энергоносители, резко сократился платежеспособный спрос на заводах по производству минеральных удобрений.* [«Горная промышленность», 2004]
- 45) *<...>темные двory, наполненные тишиной ночи, обшарпанные парадные, таинственные лестничные площадки <...> вызывали у него то чувство ожигающего прикосновения к загадочной человеческой жизни, какое испытал однажды еще в детстве, **когда** случайно нашел на улице кем-то потерянный кошелек, новенький, сшитый из бордовой кожи, сверкающий золотистым замочком.* [Юрий Бондарев. Берег (1975)]
- 46) *Рассказ о том, что видел, слышал, переживал, понимал автор тогда, **когда** событие происходило, вытесняет объяснение, почему оно случилось.* [М. В. Вишняк. Дань прошлому (1953)]
- 47) *Правила настоящей статьи не распространяются на случаи, **когда** примененные в отношении лица меры процессуального принуждения или постановленный обвинительный приговор отменены или изменены ввиду издания акта об амнистии, истечения сроков давности, недостижения возраста, с которого наступает уголовная ответственность <...>* [Уголовно-процессуальный кодекс Российской Федерации (2001) // 2004]
- 48) *Что он меня пожалел тогда и подобрал, **когда** я вышла из родильного дома Грауэрмана с пустыми руками и двумя гробами.* [Людмила Петрушевская. стакан воды (1980-1990)]

The Pilot Corpus of the English Semantic Sketches

Maria Petrova
ABBY
Moscow, Russia
m.petrova@abby.com

Maria Ponomareva
HSE, ABBY
Moscow, Russia
maria.ponomareva@abby.com

Alexandra Ivoylova
RSUH, MIPT, ABBY
Moscow, Russia
aleksandra.ivoilova@abby.com

Abstract

The paper is devoted to the creation of the semantic sketches for English verbs. The pilot corpus consists of the English-Russian sketch pairs and is aimed to show what kind of contrastive studies the sketches help to conduct. Special attention is paid to the cross-language differences between the sketches with similar semantics. Moreover, we discuss the process of building a semantic sketch, and analyse the mistakes that could give insight to the linguistic nature of sketches.

Keywords: word sketches, semantic sketches, frame semantics, word sense disambiguation, corpus lexicography

DOI: 10.28995/2075-7182-2022-21-436-446

Пилотный корпус английских семантических скетчей

Мария Петрова
ABBY
Москва, Россия
m.petrova@abby.com

Мария Пономарева
ВШЭ, ABBY
Москва, Россия
maria.ponomareva@abby.com

Александра Ивойлова
РГУ, МФТИ, ABBY
Москва, Россия
aleksandra.ivoilova@abby.com

Аннотация

Работа посвящена созданию семантических скетчей для глаголов английского языка. Пилотный корпус состоит из англо-русских пар скетчей, на примере которых демонстрируется, какие сопоставительные исследования скетчи позволяют проводить. Особое внимание уделяется межъязыковым различиям скетчей одного семантического поля в разных языках. Кроме того, в статье обсуждается процесс построения скетча, возможные ошибки и их лингвистическая природа.

Ключевые слова: скетчи слов, семантические скетчи, семантика фреймов, разрешение лексической многозначности, корпусная лексикография

1 Introduction

In the current paper, we present the pilot corpus of the English semantic sketches and compare the English sketches with their Russian counterparts. The semantic sketch is a lexicographical portrait of a verb, which is built on a large dataset of contexts and includes the most frequent dependencies of the verb. The sketches consist of the semantic roles which, in turn, are filled with the most typical representatives of the roles.

The influence of context on word recognition has been well-known for quite a time. Semantic context allows faster word recognition and the inferring of the skipped words while reading. The research in

this area has been conducted in psycholinguistics since the 1970s, with the earliest works by (Tweedy et al., 1977) and (Becker, 1980). Here the focus is on visual word recognition while reading and word recognition by bilingual persons (Assche et al., 2012). Another aspect of the topic is the automatic inferring of the skipped words by context, widely known as a common NLP task today.

The ability to represent the word by its context is the central idea of distributional semantics. It serves as a basis for the bag-of-words task, which is a training objective for static vectors like word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). In the approach, the context has a set length, and the words entering the fixed window are considered equally.

The semantic sketches do not have such disadvantages, as they are based on the result of the semantic parsing and therefore take into account not all the words occurring in the context, but only the words that semantically depend on the given core. That is, we take not the linearly nearest tokens, but the tokens close in the parsing graph, where the type of the links is considered as well.

The BERT (Devlin et al., 2019) contextual embeddings, which followed the static vectors and became a state-of-the-art solution for meaning representation, also rely on the idea of expressing word semantics through its context, using the objective of masked language modeling.

One of the main weaknesses of all vector representations is their interpretation and quality evaluation. The common practice is to consider the vectors as good, if they allow one to get the necessary quality for the down-stream task.

The advantage of the semantic sketches is in their interpretability and clear creation process. The sketches can be regarded as human-interpretable representation of word meanings, which one gets automatically with the help of the statistical methods used on the large text datasets.

The semantic sketches were first demonstrated in (Detkova et al., 2020), where we presented the idea of the semantic sketches itself and analysed the semantic mark-up used for building the sketches. Further, the pilot corpus of the Russian sketches has been created (Ponomareva et al., 2021). Herein, we have continued the work and created the pilot corpus of the English semantic sketches. The corpus is bilingual: each English sketch is accompanied by the Russian analogue with the same semantics, so one can compare the English sketch with the Russian one and analyse the contrastive differences between the sketches. Thereby, the contribution of the current paper is the creation of the English semantic sketches, on the one hand, and the creation of the parallel bilingual sketch corpus – on the other.

The structure of the paper is as follows. First, we briefly characterise the semantic sketches themselves. Second – give a description of the suggested corpus and explain what kind of verbs it contains. After that, we analyze the mistakes one faces when building the sketches, and focus on the cross-language differences between the sketches with similar semantics. In conclusion, we summarize the results.

2 Semantic Sketches

The idea to represent word compatibility in the form of the sketch belongs to Adam Kilgarriff (Kilgarriff et al., 2014) and is currently realized in the Sketch Engine project¹. Verbal dependencies are classified according to their syntactic roles and statistically ranged, which allows one to see all of the most frequent syntactic dependencies of the verb at the same time.

The problem is that the syntactic sketches do not differentiate between various meanings of the verbs and combine all possible meanings in one sketch. To overcome this problem, we suggested the semantic sketches, which take the semantic models into account and classify the dependencies by their semantic relations with the core instead of their surface realizations (Detkova et al., 2020). For instance, see fig. 1 with the sketch of the verb ‘to focus’ in the meaning ‘to concentrate on smth., to pay special attention to smth.’:

¹www.sketchengine.eu

Theme	Object_Situation	Object	Agent	Time	ParentheticalSaliency
upon issues	attention	these chapters	investors	thus far	primarily
on development	his research	the study	investigators	nowadays	mainly
on businesses	its efforts	resources	facilities	in the future	first and foremost
on the activities	discussions	the review	players	recently	principally
on sectors	its activities	the programme	the museum	in the past	mostly
on education	the strategy	the paper	the school	hitherto	predominantly
on managing performance	our energy	the film	the division	typically	above all
on other areas	its analysis	the media	our institutions	henceforth	most of all

Figure 1: The sketch for the verb ‘to focus:TO_FOCUS’

Such sketches are built for each meaning separately, however, it demands a significant text corpus with full semantic mark-up. The authors settled on the Compréno mark-up built by the Compréno parser, which includes not only actant dependencies, but all possible links.

In the Compréno model, all words are presented in the form of a thesaurus-like semantic hierarchy, which consists of the semantic classes (semantic fields), and a set of the semantic roles for the classes (for detail, see (Anisimovich et al., 2012), (Petrova, 2014)). If a verb has several meanings, it enters several semantic classes with its own semantic model each. The semantic class is specified for each sketch.

3 English SemSketches Corpus

The SemSketches pilot corpus consists of 100 English sketches which are manually checked. It means that we have chosen the sketches manually according to their quality. The sketches are built on the corpus of the English texts comprising different genres, such as technical texts, news, fiction, and containing 14 million syntactic verbal links, that is, links which depend on the verbal cores.

Each English sketch is provided with the parallel Russian sketch from the same semantic class, as shown in fig. 2 and 3:

Object_CreationDestruction	Locative	Time	Concurrent_Situative	Locative_Orientation	Time_Situation
a bomb	in a bus	prematurely	wounding palestinians	midway	when overheated
a grenade	in flames	overnight	killing four people	south of the capital	during prayers
a car	on a train	at dawn saturday	injuring two passer-bys	outside general santos	before dawn
the battery	on the island	instantly	causing no injuries	north of the city	in fire
pieces	in the capital	on monday night	blowing out windows	down the ramps	on impact
the rocket	in a dustbin	on christmas eve	toppling a building	inside the city	during the party
a particle	in mid air	early on monday	targeting a gendarmerie vehicle	below the bilge keel	upon contact
artillery shells	on the west bank	shortly after takeoff	damaging a carriage	higher	

Figure 2: The sketch for the verb ‘to explode:TO_BLOW_UP’

Object_CreationDestruction	Time	Locative	Modality	OrderInTimeAndSpace	Agent
бомба	вот-вот	на mine	как будто	в конце концов	террористы
мосты	до августа	на полигоне	неожиданно	наконец	шахид
мина	в любой момент	в метро	словно	опять	талибы
снаряд	в мгновение	в воздухе	вдруг	затем	саперы
граната	время от времени	в нейтральных водах	внезапно	потом	партизаны
самолет	в любую секунду	в автобусе	непрерывно	сначала	чеченцы
дома	мгновенно	в голове	как	вновь	большевики
храм	зимой	в небе	как бы	снова	американцы

Figure 3: The sketch for the verb ‘взорвать’:TO_BLOW_UP’

For 100 English sketches, 84 Russian sketches are used: it means that some Russian sketches correspond to more than one English sketch. Totally, the corpus includes 113 English-Russian sketch pairs.

The choice of the English verbs is based on the Russian corpus which was built in (Ponomareva et al., 2021). The Russian corpus, in turn, includes only polysemantic verbs as an important point is to investigate how good the sketches can differentiate between various meanings of the verbs.

To form the English sample, we have taken the verbs from the same semantic classes and set the threshold of 200 semantic links for each English verb: it means, the verb must have at least 200 links in the English texts corpus. (For comparison, the threshold for Russian verbs was 2000 links, but the Russian sketches were collected on the bigger dataset which includes more than 36 million links.)

After it, 100 English sketches were chosen, which met the above mentioned criteria and seemed to be enough representative to show the ability of the sketches to deal with polysemy, word sense disambiguation (WSD) problem, and asymmetrical compatibility of the verbs with similar semantics in different languages. Of course, the pilot corpus of 100 sketches is not enough for conducting representative contrastive research, however, certain observations seem to be of interest for comparative studies even on the small sample, as it is demonstrated below.

4 What the mistakes in the sketches demonstrate

The sketches are based on (1) the semantic relations the verb has in the text collection; (2) the work of the parser which classifies the relations according to their semantic roles and defines the meanings of the verbs. Therefore, the view of the sketch depends on the number of links the verb has in the corpus, and on the correctness of the parser’s work. Herein the following mistakes are possible, which concern the automatic generation of the sketches.

4.1 ‘Empty’ sketches

The insufficient number of links leads to partly ‘empty’ sketches, where the semantic roles contain very few fillers, up to only one. So when the semantic role column is partly empty, it can mean that the number of the role’s links in the corpus turned out to be poor (for instance, see the [Cause_Actant] slot in the sketch for ‘inflict’ on fig. 4). As the number of texts grows, this problem occurs rarer.

Another reason for the lack of fillers comes from the narrowness of the semantic role filling. That is, slots like [Object] or [Cause] have rather wide filling, while [Locative] and [Time] are more restricted in this respect. In turn, the Comprono parser has a large set of characteristic slots (for size, colour, speed, modality, and so on), so some slots possess rather narrow semantics and include a small set of fillers (like the [StaffOfPossessors] slot in the same sketch on fig. 4).

Object_Situation	Cause_Actant	Experiencer	Agent	Object	StaffOfPossessors
injury	self	on detainees	hijackers	casualties	self
pain	by the iceberg	on the exchequer	by the spouse	the hall	
harm	poor acting	on the attackers	by a mob	a stinging rebuke	
damage	by the war	on both combatants	by an attacker	bruises	
casualties	methamphetamine	on juveniles	the opposition	wounds	
suffering	by a club	on its inhabitants	the responden	dreams	
defeat		on the survivors	by the army	any injuries	
hardship		on the environment	by others	losses	

Figure 4: The sketch for the verb ‘to inflict:TO_BRING_STATE_TO_SMB’

Moreover, there are verbs with narrow compatibility, such as lexical functions. For instance, see the [Object] slot in the sketch of ‘играть:TO_COMMIT’ (fig. 13).

In sketches like these, empty lines in the semantic slots are correct.

4.2 Incorrect semantic roles or incorrect fillers

Other errors concern either the incorrect choice of the semantic slot for the given verb meaning, or the wrong fillers of the slot. As one of the key points is to examine how well the sketches solve the WSD problem, this type of mistakes is important for us.

An illustration for the incorrect semantic slot is the Russian sketch for ‘доставлять:TO_BRING_STATE_TO_SMB’ (fig. 5), parallel for the above shown ‘inflict:TO_BRING_STATE_TO_SMB’. It contains the [Locative_FinalPoint] slot, which must definitely belong to another meaning of the verb – ‘bring to some place’.

Object_Situation	Experiencer	Cause_Actant	Time	Modality	Locative_FinalPoint
удовольствие	окружающим	чтение	немедленно	явно	в отделение
наслаждение	хозяину	работа	минуты	похоже	во дворец
радость	родителям	прогулка	за час	вряд ли	на этаж
много хлопот	царю	процесс	сегодня	видимо	в больницу
неудобства	читателю	письмо	ночи	вероятно	сюда
удовлетворение	зрителям	удовольствие	скоро	действительно	домой
неприятности	врагам	общение	всю жизнь	конечно	в город
массу неудобств		игра	до сих пор	может быть	на место

Figure 5: The sketch for the verb ‘доставлять:TO_BRING_STATE_TO_SMB’

Examples of the wrong fillers have already been shown in (Ponomareva et al., 2021). The reasons are usually bound either with the statistics, or with the work of the parser. At the analysis stage, all possible hypotheses are built for the sentence – with all possible homonyms that can fit. The final structure turns out to be the one with the highest scores. In some cases, hypotheses with more frequent homonyms win due to their higher frequency, in spite of the fact that the whole structure with the wrong homonym gets lower evaluations.

As the text collections for building the sketches grow, the statistics of the proper analysis improves, therefore, we expect that most part of the errors will be corrected with enlarging the corpora. Nevertheless, in case of the improper work of the parser, the opportunity to correct the semantic models that the parser uses exists as well.

4.3 The syntactic homonymy

Key difference between the semantic and the syntactic sketches is that in the former 1 surface realisation can correspond to various semantic roles. For instance, ‘for’-dependency can introduce Time, Purpose, Distance, Motive and a number of other relations.

Usually, the proper semantic role is chosen according to the semantic model of the given verb in *Compreno* – namely, the set of the semantic slots with the necessary surface realisation, the fillers of the semantic slots, and their status (which marks the role as more or less preferable).

When the model or the statistics give improper results, the semantic role of the dependency can be defined incorrectly. For instance, see the [Purpose_Goal] slot of the verb ‘throw:TO_THROW’: the first line contains the nominal group ‘for 408 yards’, which must evidently belong to the [Locative_Distance] slot (fig. 6).

Object	Locative_FinalPoint	Agent	Purpose_Goal	Time	Object_Situation
rocks	overboard	protesters	for 408 yards	after nightfall	a no-hitter
stones	at police	the demonstrators	at the issue	now and then	an interception
a grenade	at soldiers	assailants	not to back off	beforehand	a fastball
bombs	in the fire	a guy	for luck	tonight	a shutout
shoes	into jail	the attackers	for career	rarely	a hissy
the ball	into the sea	youths	to remember	often	changeups
firebombs	into prison	crowds		overnight	chew
weight	in the air	the prisoners		later	passes

Figure 6: The sketch for the verb ‘to throw:TO_THROW’

Another example is the group ‘for this moment’ in the [Time] slot instead of [Motive] in the sketch of ‘to thank’ (fig. 7). Here, on the contrary, [Motive] is definitely more frequent, but ‘moment’ is a very typical [Time] filler, therefore, high statistical evaluation of the correlation ‘moment’-[Time] made the incorrect structure win.

Addressee	Motive	Agent	Addition	Time	Ch_EvaluationOfHumanTemperAndActivity
the rapporteur	for their support	the chairperson	again	in advance	warmly
the senator	for his work	the authors	lastly	meanwhile	sincerely
the commissioner	for his statement	the forum	likewise	for this moment	heartily
the member	for his reports	the chair	inter alia	once again	
the government	for its efforts	the party	finally	to date	
participants	for his briefing	the conference	also	in a few minutes	
the staff	for their contributions	the group		later	
the secretariat	for discussions	the convener			

Figure 7: The sketch for the verb ‘to thank:TO_THANK’

All the mistakes deal with different aspects of the WSD and homonymy problems. Their number does not seem significant, nevertheless, their statistical estimation must be made when creating a larger sketch corpus.

5 Cross-language differences between the sketches with similar semantics

The one-language sketch corpora suggest good lexicographic portraits of the verbs, showing their most frequent semantic links sorted according to the semantic roles of the dependencies. Moreover, apart from

purely lexicographic tasks, the sketches allow one to solve various problems bound both with the context usage of the verbs and with their polysemy.

Another purpose of the sketches deals with contrastive studies. Parallel sketches from different languages give perfect representation of the correlation between similar verbs, therefore, parallel sketch corpora would be helpful in this respect.

Evidently, each sketch can correspond to more than one sketch in another language. To get a full set of all possible counterparts, one should take the necessary sketch in one language and the sketches for all the semantic equivalents in the same semantic class in another language. After it, one can range the counterparts according to their affinity with the primary verb. We have not made such full sets in the pilot corpus, however, adding this option is included in further plans.

At the current stage of the project, the correlations between the English and the Russian sketches do not include all possible correlations for each verb, so the sketch pairs are just a subset of the possible variants.

Some pairs look similar: both English and Russian sketches include the same set of semantic roles, and the semantic roles contain either fillers with close semantics, or just a wide range of fillers with no special semantic restrictions on them.

At the same time, many sketches demonstrate significant differences between the English and Russian equivalents. Most of them concern the following situations:

- (a) some semantic slot is present in the sketch of one language and is absent – in the corresponding sketch in another language;
- (b) equivalent sketches contain the same sets of the roles in both languages, but the fillers of some role differ significantly;
- (c) the semantic field where the considered verbs belong is structured differently in different languages.

5.1 Different semantic roles in the equivalent sketches from different languages

Frequently, the semantic role sets in the parallel sketches do not coincide completely. It concerns both the actant roles and the circumstantial ones. The reasons can be different. First, the semantics of one verb may be wider than the semantics of the other, therefore, the model of the former can include additional roles which are absent in the model of the latter. Second, the model of both verbs can include the same sets of roles, however, the frequency of some roles may differ for various verbs, which can be motivated both by the verb's semantics and by the representativeness and contents of the corpora for building the sketches.

An example of the first case is the correlation between the semantic derivatives in different languages. For instance, Russian verb 'трясти' 'to shake' does not attach the initial point dependency in contexts like (1) and (2), while the English 'shake' does:

- (1) A sound they couldn't shake [from their Locative_InitialPoint: heads] – Звук, который им никак не удавалось вытряхнуть [из Locative_InitialPoint: головы];
- (2) I saw immediately that my few belongings had been disturbed—collars not refolded, one of my chemises balled up and pushed into a corner, the tortoiseshell comb shaken [from its Locative_InitialPoint: handkerchief]. – И сразу увидела, что в моих вещах кто-то рылся — воротники были сложены неаккуратно, одна из моих рубашек скомкана и засунута в угол, черепаховый гребень вытряхнут [из носового Locative_InitialPoint: платка].

In Russian, the semantic derivative 'вытряхнуть' 'shake out' is used when the initial point role is expressed in a sentence. Therefore, the sketches can show that 'shake' usually corresponds to the Russian 'трясти' (which does not mark the 'direction of shaking'), but can also correspond to 'вытрясти' (which denotes the 'from' direction) with the dependency of the initial point.

Nevertheless, there can be occasional variations depending on the contents of the corpora, especially as far as less frequent verbs are considered. The more the corpora are, the more stable are the results. Thus we permanently enlarge the size of the dataset for building the sketches.

As an instance of such statistical oscillations, see the sketches for “find:TO_SEEK_FIND” and “найти:TO_SEEK_FIND”. The first five roles coincide, but the sixth one is different – it is [Metaphoric_Locative] for the English ‘find’ and [Modality] for the Russian ‘найти’ (fig. 8, 9):

Object	Object_Situation	Locative	Possessor	Time	MetaphoricLocative
the body	a solution	here	police	recently	at the website
a place	ways	where	one	so far	in chapter
information	a job	there	scientists	in the winter	on pages
a partner	employment	elsewhere	archaeologists	often	in memory
a buyer	further details	at home	people	soon	in appendix
no evidence	the answer	in a car	the reader	commonly	in table
remains	something	in the river	the researchers	today	in literature
refuge	inspiration	in the area	the men	occasionally	in the book

Figure 8: The sketch for the verb ‘to find:TO_SEEK_FIND’

Object	Object_Situation	Locative	Time	Possessor	Modality
места	работу	здесь	сразу	читатель	обязательно
общий язык	отражение	в кармане	ново	люди	вряд ли
слово	выход	в лесу	утром	автор	непрерменно
человека	ответ	в городе	недавно	мама	едва ли
время	применение	езде	до сих пор	отец	наверняка
выражение	силы	на помойке	или поздно	поэт	безошибочно
приют	отклик	в шкафу	сейчас	герой	может быть
своего читателя	способ		скоро	жена	может

Figure 9: The sketch for the verb ‘найти:TO_SEEK_FIND’

Both roles – [Metaphoric_Locative] and [Modality] – can be frequently used with both verbs. In this case, the difference does not seem meaningful.

5.2 Different fillers of the semantic roles

Let us consider some sketches for the descendants of the semantic class “TO_COMMIT”: the English verbs ‘do’, ‘play’ and the Russian verbs ‘делать’, ‘играть’. “TO_COMMIT” is a kind of lexical function, where the verbs have rather narrow compatibility in the [Object] role (place trust/hope vs pay a visit vs play a joke/trick vs take a look/try/walk/etc., and so on).

As fig. 10, 11, 12 and 13 demonstrate, the compatibility of the verbs ‘do’ and ‘делать’ is rather wide, while ‘играть’ combines with only four Object fillers.

Ch_Relation_Coincidence	Object_Situation	Agent	Object	Time	Agent_Metaphoric
so	business	the government	everything possible	before	by hand
differently	things	people	something	so far	the economy
the same	the job	men	whatever	in the future	the system
otherwise	our best	a parent	exercises	in my life	life
thus	the work	everybody	a favor	now	process
unevenly	good	a woman	homework	normally	management
different	research	somebody	any act	in the past	
alike	no harm	a person	the rest	in history	

Figure 10: The sketch for the verb ‘to do:TO_COMMIT’

Object_Situation	Time	Agent	Object	Modality	Locative
шаг	теперь	люди	выводы	правильно	здесь
дело	сейчас	автор	дело	неправильно	в стране
выбор	в жизни	власть	паузу	невозможно	на моём месте
операцию	тогда	писатель	замечание	собственно	в больнице
укол	раньше	отец	предложение	вроде бы	в мире
попытку	вовремя	поэт	снимок	непрерывно	дома
движение	сегодня	женщина	доклад	конечно	
ставки	обычно	врач	фильм		

Figure 11: The sketch for the verb ‘делать:TO_COMMIT’

Object_Situation	Agent_Metaphoric	Agent	Addition	Sphere	Time
a role	factors	organizations	last	in the development	today
a part	the sector	society	incidentally	in promoting	during peacetime
the sport	the proteins	media	besides	in the progress	in world history
the tournament	our role	women	also	in the revolution	over the decade
a function	variables	the city	moreover	in diseases	hitherto
a trick	your feedback	the european union	as well	in addressing the problem	in the past
hustle	politics	your community	too	where	ever since
	religion		again	in relations	during his lifetime

Figure 12: The sketch for the verb ‘to play:TO_COMMIT’

Object_Situation	Agent_Metaphoric	Time	Locative	Modality	Agent
роль	фактор	в истории	в государстве	несомненно	организации
злую шутку	литература	в жизни	в обществе	безусловно	интеллигенция
значение	обстоятельство	в дальнейшем	здесь	по-видимому	церковь
свадьбу	деньги	в период	в картине	возможно	актеры
	религия	в эпоху	в мире	бесспорно	государство
	идеи	впоследствии		вероятно	журналы
	понятие	подчас		по всей видимости	театр

Figure 13: The sketch for the verb ‘играть:TO_COMMIT’

Besides, the four verbs differ in the sets of the semantic roles as well. [Agent], [Object_Situation] and [Time] are present in all four sketches. [Object] is absent in the sketches of ‘играть’ and ‘play’ as their compatibility does not include the corresponding fillers.

‘Do’ and ‘играть/play’ include [Agent_Metaphoric] slot, while ‘делать’ does not include it. The reason seems to be in the semantics of the fillers of the [Object] and [Object_Situation] slots: the most frequent Object_Class fillers are ‘шаг’ ‘step’, ‘выбор’ ‘choice’, ‘операция’ ‘operation’, ‘снимок’ ‘picture’ and so on, which are more often combined with active human-like agent rather than inanimate agents like ‘economy, system, process’ and alike.

As far as the circumstantial dependencies are concerned, both Russian sketches include the semantic roles of [Modality] and [Locative] while the English ‘do’ includes [Ch_Relation_Coincidence] slot (in the Compreno model, it characterizes objects or situations according to their similarity) and ‘play’ – [Addition] and [Sphere]. At first sight, these differences do not seem meaningful, however, it would be interesting to regard the sketches of the whole semantic class TO_COMMIT to examine how regular such correlations are.

Another example concerns verbs with wider compatibility, where the restrictions on the Object role are not purely lexicalized, but concern a wider range of fillers with common semantic features. For instance, let us take the semantic field “TO_POUR” (something liquid or friable). English and Russian structure it differently as far as the core verbs’ compatibility is concerned. Namely, the English verb ‘to pour’ attaches objects which are liquid (water, wine), friable (sand, sugar), or consist of many small pieces (crystals, euros, diced meat, and so on). In Russian, the verb ‘лить’ is used with liquid objects only and the verb ‘сыпать’ – only with friable objects and objects consisting of many small pieces. Therefore, the Object slot fillers differ correspondingly in the sketches (fig. 14).

Объект (pour:TO_POUR)	Объект (лить:TO_POUR)	Объект (сыпать:TO_POUR)
tea	слезы	соль
beverage	воду	песок
wine	вино	зерно
petrol	чаю	пепел
sand	кипяток	снег
salt	водки	труха
crystals	коньяку	штукатурка
million euros	стакан водки	порошок

Figure 14: The fragment of the sketches for the verbs ‘to pour:TO_POUR’, ‘лить:TO_POUR, and ‘сыпать:TO_POUR’

Nonetheless, the amount of eight most frequent fillers which is usually shown in the sketches is not always enough to demonstrate such differences, as the most frequent objects can bear the same semantic features.

As one can see, the sketches suggest a wide range of comparative data in the field of semantics and demonstrate the semantical differences between the verbs of the same semantic class both in different languages and within one language as well.

6 Conclusion

In the given paper, we have presented the pilot corpus of the English semantic sketches.

As the sketches are provided with their semantic parallels in Russian, we have also illustrated what kind of comparative studies the sketches allow to conduct, especially as far as the differences in the

semantic roles and their typical fillers are concerned. An important point is the ability of the sketches to deal with polysemy and to differentiate between various homonyms.

We have also discussed common types of mistakes occurring while building the sketches and speculated about their linguistic and technical nature.

Our further plans are to improve the sketches by obtaining them on a bigger dataset, to enlarge the sketch corpus and build the sketches for each verb from the dataset, to provide the corpus with some additional features, such as the opportunity to show more semantic slots and more fillers of the slots when necessary, and to see the correlations between all the verbs of the same semantic class. After it, the work on adding other languages to the sketch corpus will be started.

At the same time, we work on the open corpus of the Compreno semantic mark-up which will include a detailed description of the mark-up principles and the semantic roles used in the mark-up, which will facilitate the understanding of the roles used in the sketches.

The current corpus is available at [github²](https://github.com/dialogue-evaluation/SemSketches/tree/main/data/task_2). Besides, we continue the work on integrating the semantic sketches in the General Internet-Corpus of Russian (GICR).

We hope the corpus would contribute to different NLP areas, especially to solving the WSD problem.

References

- 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.
- Eva Van Assche, Wouter Duyck, and Robert J Hartsuiker. 2012. Bilingual word recognition in a sentence context. *Frontiers in psychology*, 3:174.
- Curtis A Becker. 1980. Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & cognition*, 8(6):493–512.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Julia Detkova, Valeriy Novitskiy, Maria Petrova, and Vladimir Selegey. 2020. Differential semantic sketches for russian internet-corpora. // *Proc Dialogue, Russian International Conference on Computational Linguistics*, Moscow.
- 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.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. // *Neural and Information Processing System (NIPS)*.
- MA Petrova. 2014. The compreno semantic model: the universality problem. *International Journal of Lexicography*, 27(2):105–129.
- Maria Ponomareva, Maria Petrova, Julia Detkova, Oleg Serikov, and Maria Yarova. 2021. Semsketches-2021: experimenting with the machine processing of the pilot corpus. // *Proc Dialogue, Russian International Conference on Computational Linguistics*, Moscow.
- James R Tweedy, Robert H Lapinski, and Roger W Schvaneveldt. 1977. Semantic-context effects on word recognition: Influence of varying the proportion of items presented in an appropriate context. *Memory & Cognition*, 5(1):84–89.

²https://github.com/dialogue-evaluation/SemSketches/tree/main/data/task_2

Between Denoising and Translation: Experiments in Text Detoxification

Sergey Pletenev

National Research University Higher School of Economics
Moscow, Russian Federation
alex010rey@gmail.com

Abstract

This paper describes a solution for the RUSSE Detoxification competition held as part of the Dialogue 2022 conference. The paper presents experiments based on autoregressive and non-autoregressive models. The following approaches are described in this paper: 1) Detoxification as a special case of the text style-transfer problem and the use of modern approaches to solve this task in Russian. 2) Using the Automatic Post-Editing algorithm as a task of translation from toxic to normative Russian text. The article provides an analysis of the listed models, their results in detoxification of sentences, as well an analysis of errors and reasons why the models gave such a diverse result.

Keywords: Sentence detoxification, pretrained language models, Non-autoregressive models, Russian language

DOI: 10.28995/2075-7182-2022-21-447-455

Между восстановлением текста и переводом: эксперименты по детоксикации текста

Плетенев Сергей

НИУ ВШЭ

Москва, Россия

alex010rey@gmail.com

Аннотация

В данной статье описано решение для соревнования по детоксикации предложений RUSSE Detoxification, проводящегося в рамках конференции Диалог 2022. В работе представлены эксперименты на основе авторегрессионных и неавторегрессионных моделей. В данной статье описываются следующие подходы: 1) Переопределение задачи детоксикации как частного случая задачи переноса стиля текста (style-transfer) и использование современных подходов для решения данной задачи на русском языке. 2) Использование алгоритма автоматического пост-редактирования текста (Automatic Post-Editing) в качестве задачи перевода из токсичного в нормативный русский текст. В статье дан анализ перечисленных моделей, их результатов в детоксикации предложений, а также анализ ошибок и причин, по которым модели дали столь разнообразный результат.

Ключевые слова: Детоксикация предложений, Предобученные языковые модели, Неавторегрессионные модели, Русский язык

1 Introduction

With the widespread development of chats, social networks, and various forums, the need to classify and filter offensive content has emerged. There is a large class of articles (Wang et al., 2021) (Georgakopoulos et al., 2018) dealing with identifying and classifying offensive content sentences. But in addition to categorizing toxic sentences, there may be a requirement to detoxify sentences, i.e. to bring the text into a neutral, readable form. The task of detoxifying sentences seeks to reduce the offensiveness of the original sentence, but at the same time preserve the meaning and message of the text. The text detoxification problem can be reformulated as a subclass of the text style-transfer problem since the style transfer problem is a widely discussed and researched area of natural language processing.

The existing methods of text detoxification and style transfer are mostly made for the English language, which makes it difficult to transfer to other languages. For this purpose, the RUSSE Detoxification corpus (Dementieva et al., 2022) was developed to solve the detoxification problem in the Russian language. This paper describes the general problem statement and proposes a detoxification method based on RuT5 and describes in detail experiments with autoregressive (AR) and non-autoregressive models (NAR) for style transfer. We compare the capabilities of RuT5 (Raffel et al., 2019) models according to the baseline models of the competition, and explore different word alignment methods, combining different inference strategies and text preprocessing. The method was ranked 4th in the Automatic Evaluation and 1st in the Private Human Evaluation Leaderboard between models. To clarify, first place was awarded to "Human References", so the model received 2nd place overall. The article itself is organized as follows: Section 2 briefly describes previous research in style transfer; Section 3 describes the data used in the experiments; Section 4 describes the experiments; in Section 5 we discuss the results and provide an analysis of the proposed method and the generated the best model capabilities, In Section 6 we discuss the possible errors in the datasets and models that led to the disagreement of the scores; Section 7 concludes the article.

Our contribution is as follows:

1. We propose our method to detoxification using Seq2Seq models.
2. We adopt state-of-the-art style transfer models and evaluate them for the Russian language.
3. We publish experiments and our models for future research.¹

2 Related works

We can categorize style transfer models into three types. The first type is the editing-based method (Li et al., 2018) (Shen et al., 2017), which edits the source sentence with several simple operations. The operations themselves consist of simply removing, replacing, or adding words to a sentence. The operations are usually trained separately and then constitute a pipeline. These methods are highly explainable and can be interpreted, but they usually need to locate and replace the stylist words, which hardly applies to complex tasks that require changes in sentence structures. The second type is sequence-to-sequence model because the detoxification task is similar to text generation tasks such as machine translation, summarization, or paraphrase generation. In this case, the model completely translates the text into a hidden representation of the model, and using a decoder generates new text sequentially, or autoregressive. This approach has shown good results in style transfer (John et al., 2019) and detoxification (Dale et al., 2021) tasks. But the main problem of such models is to preserve the original context, especially for long texts, which is a difficult task for seq2seq models. The third type of model (Huang et al., 2021) (Luo et al., 2019) combines the two previous approaches: on the one side, they create or learn a set of word alignments, and on the other hand, the sentence is generated end-to-end fashion. In addition, we adapted the Automatic Post-Editing (APE) method with Levenshtein transformer (Gu et al., 2019) to detoxification task. The APE consists of two steps: autoregressive(AR) generation using the seq2seq model and post-editing using an additional non-autoregressive(NAR) model.

3 Data

The organizers of the RUSSE Detoxification shared task has introduced a parallel detoxification dataset. The source sentences are Russian toxic messages from Odnoklassniki, Pikabu, and Twitter platforms. The target part of the dataset is the same messages which were manually rewritten by crowd workers to eliminate toxicity. Some toxic sentences contain multiple (up to 3) variants of detoxification. The dataset is divided into train, development, and test sets.

Dataset statistics:

- train: 3,539 toxic sentences with 1-3 detoxified versions;
- development: 800 toxic sentences with 1-3 detoxified versions;
- test: 1,474 toxic sentences with 1-3 detoxified versions.

¹<https://github.com/AlexRey/DenoiseOrTranslation>

GYAFC Dataset	
Informal	Formal
Even the day after would be okay. well all ur missing is a million dollar smile..:)	Even the following day would be alright. Well, all you are missing is a smile worth a million dollars.
And I hear ya Fountain...same thing happened to me. altho, i dont really like girls all that much.	I hear you, Fountain, something similar happened to me. Although, I do not really like girls all that much.
RUSSE Detox Dataset	
Toxic	Neutral
это не наглость. это подлость! мерзавец папаколи это твари а не люди <ThumbsDownEmoji> пошла на хер со своим гарантом дура ты зостовляиш миня стратать	Это не наглость, это подлость Это плохие люди. Оставьте всех в покое с гарантом. ты заставляешь меня страдать

Figure 1: Examples from datasets

Due to fact that the amount of data is sufficient to train large pre-trained models but completely unsuitable for training models from scratch, we used an additional dataset (Dale, 2021) of paraphrases in Russian to train part of NAR models. This dataset contains 1 million sentences from the Russian part of Leipzig Corpora Collection 2019. Our choice of paraphrase is related to the success of such models in detoxification (Krishna et al., 2020).

Following (Kim and Rush, 2016), knowledge distillation from an autoregressive model can effectively simplify the training data distribution, which is essential for NAT-based models to learn good results. The easiest way of performing distillation is to follow the instructions of training a standard transformer model on the same data, and then decode the training set to produce a distillation dataset for NAT.

We use a pipeline to clean-up and prepare our datasets:

1. Remove punctuation marks from the text, except for the end of the sentence;
2. Remove emojis from text;
3. Remove unnecessary spaces in the sentence.

4 Experimental setup

The organizers of the competition provided several baselines:

1. **BertDelete** As a simple baseline this model is based on RuBert(Kuratov and Arkhipov,) model and toxic word dataset. Using lemmatized search on text, this algorithm mask-out all toxic words and then generate it with masked language model.
2. **RuT5** This baseline was simply trained on detoxification dataset with RuT5-base. No additional prefixes there used.
3. **ruPrompts** This baseline was trained on RuT5-base with model tuning via automatic prompt search.

For our experiments, we used NAST(Huang et al., 2021) model that performed best on the GYAFC (Rao and Tetreault, 2018) dataset. Grammarly’s Yahoo Answers Formality Corpus (GYAFC) is the largest dataset for any style containing a total of 110K informal / formal sentence pairs in english. Differences between GYAFC and RUSSE Detox corpuses shown in Table 1. As we can see, both datasets have similar characteristics: missing or incorrect punctuation, emoji, grammatical errors, and toxic words.

NAST(Huang et al., 2021) is a non-autoregressive generator based on the observation of the word alignment: in style transfer tasks, most generated words can be aligned with the source words, where each pair of the aligned words is either identical or highly relevant. Additionally, we redesigned two non-autoregressive models to handle the detoxification task: LevT and DisCo.

text	Шустрая бурая лисица прыгает через ленивого пса																		
src	Ш_ устра_ я бу_ рая ли_ си_ ца пры_ гает через лени_ вого п_ са																		
0	Ш_	у_	стра_	я	бу_	рая	ли_	си_	ца	пры_	га_	ет	че_	рез	лени_	во_	го	п_	са
1	Ш_	у_	стра_	я	бу_	рая	ли_	си_	ца	пры_	га_	ет	че_	рез	лени_	во_	го	п_	са
2	Ш_	у_	стра_	рый	бе_	рый	ли_	с	ца	пры_	га_	ет	че_	рез	лени_	во_	го	п_	са
3	Ш_	у_	стр_	рый	бу_	рый	ли_	с	пры_	га_	ет	че_	рез	лени_	во_	го	п_	са	са
4	Ш_	у_	стр_	рый	бу_	рый	ли_	с	пры_	га_	ет	по	лени_	вы_	м	п_	сам	[pad]	

text	Шустрая бурая лисица прыгает через ленивого пса																			
src	None																			
0	Ш_	[mask]	я	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	п_	са
1	Ш_	[mask]	я	[mask]	ли_	с	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	п_	са
2	Ш_	[mask]	я	[mask]	ли_	с	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	п_	са
...	...																			
12	Ш_	гур_	мо_	вая	ли_	су_	шка	прибе_	га_	ет	к	не_	лов_	кому	п_	су				

Figure 2: The difference between NAR paraphraser with and without src initialization

Disentangled Context (DisCo) transformer (Kasai et al., 2020) is a non-autoregressive sequence-to-sequence model. But unlike classical NAR architectures, where the model can only predict masked words, DisCo can predict all tokens simultaneously, which gives faster inference and improved model quality. It also uses a *easy-first algorithm*, in which each word is predicted by the words the model is most confident about. This decoding algorithm allows different contexts to be predicted in each iteration for all available tokens, allowing the decoding to stop when the model gets a good prediction. In this work we also use *mask-predict algorithm*, in which the number of iterations is always specified by some constant T .

The Levenshtein Transformer (LevT) (Gu et al., 2019) is a type of transformer that aims to transform text by sequentially adding, replacing, and deleting words. Hence, LevT is proposed to break the standardized decoding mechanism and replacing it with two basic operations — insertion and deletion. LevT is trained using imitation learning. The resulted model contains two policies and they are executed in an alternate manner. The authors argue that with this model decoding becomes more flexible. For example, when the decoder is given an empty token, it falls back to a normal sequence generation model. On the other hand, the decoder acts as a refinement model when the initial state is a low-quality generated sequence.

These models were initially trained on distilled russian paraphrase data, then fine-tuned on the detoxification corpus. In our case, as Teacher model we used for distillation `rut5-paraphrase2` which was evaluated on the paraphrase training dataset. A beam-search with size 3 was used as an additional parameter.

Because of the iterative generation of NAR models, it is possible to initialize them with any text. In experiments, we initialize them in three ways: **blank** – generation without initialization; **src** – generation with duplication of input text; **RuT5** – initialization with text obtained from autoregressive model RuT5. Similar methodology is used in LevT to solve APE task.

Evaluation is based on various metrics (Dale et al., 2021): **Style accuracy** (ACC) is based on pretrained toxicity classifier. **Content preservation** (SIM) is evaluated as the similarity of sentence-level embeddings of the original and transformed texts computed by the model. **Fluency** (FL) measured with the classifier of linguistic acceptability trained on the CoLA dataset. And **J** which is the multiplication of sentence-level style accuracy, content preservation, and fluency.

5 Results

For generation in all autoregressive and non-autoregressive models we use beam-search 12, with no limit on generation length. For non-autoregressive models we use the number of iterations equal to 16. For

²<https://huggingface.co/cointegrated/rut5-base-paraphraser>

model	Accuracy	Similarity	Fluency	J
RuT5-large	0.9475	0.8191	0.9107	0.7094
RuT5-baseline	0.796	0.827	0.837	0.560
RuT5-prompts	0.811	0.793	0.804	0.528
rubert-delete	0.558	0.887	0.852	0.406
NAST	0.8339	0.4983	0.7298	0.3074
LevT-blank	0.7327	0.0386	0.4135	0.0122
Disco-easy-first	0.1734	0.9321	0.9502	0.1435
Disco-mask-predict	0.3639	0.6685	0.7029	0.1707

Table 1: Test result without initialization

model	Accuracy	Similarity	Fluency	J
LevT-src	0.8262	0.4969	0.7071	0.2938
Disco-easy-first	0.4470	0.6611	0.7489	0.2255
Disco-mask-predict	0.3639	0.6685	0.7029	0.1707

Table 2: Test result with source initialization

generation with DisCo we use two different decoding methods: mask-predict and easy-first.

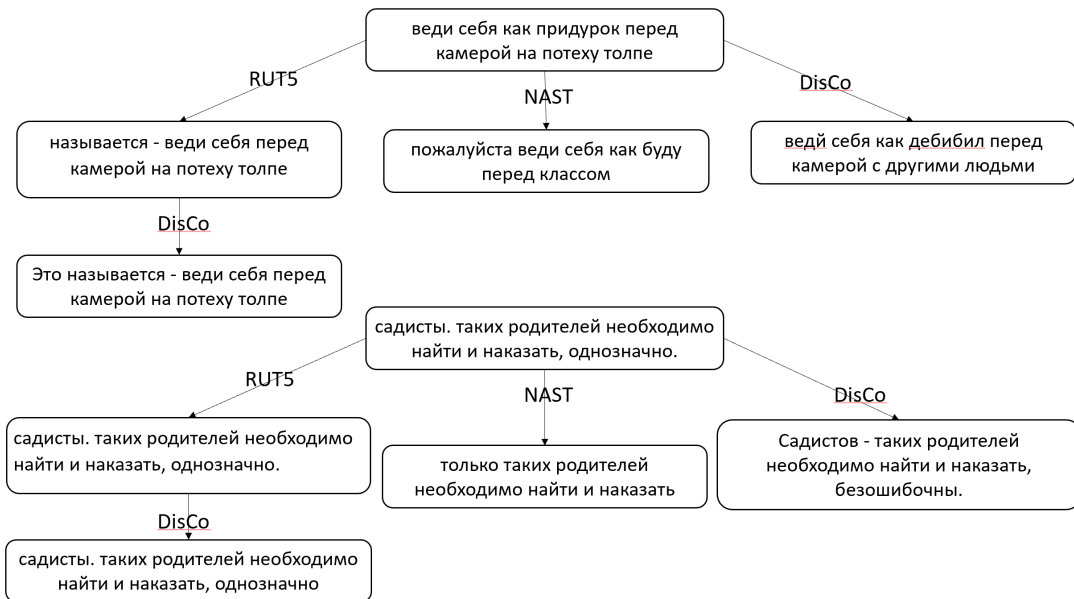


Figure 3: Visualisation of detoxification

We use HuggingFace Transformers³ for RuT5 training and prediction. Each model is trained with following parameters: encoder length 256, decoder length 256, batch size 3, 3 epochs, learning rate $5e-05$, after each 1000 steps we evaluate our models with beam size 12. For NAR and paraphrase models training we use modified FairSeq⁴ library.

As a strong baseline, we trained RuT5-large on the cleaned dataset. Table 1 shows the results of model generation without any additional initialization. The APE model showed the worst result, which is due to

³<https://huggingface.co/transformers>

⁴<https://github.com/jungokasai/deep-shallow>

model	Accuracy	Similarity	Fluency	J
RuT5-on-RuT5	0.9559	0.8043	0.9048	0.7007
RuT5-on-RuT5-beam1	0.9612	0.7979	0.9029	0.6982
LevT-mt	0.7881	0.6196	0.6195	0.3311
Disco-easy-first	0.9398	0.7645	0.8658	0.6314
Disco-mask-predict	0.8320	0.4789	0.3963	0.1570

Table 3: Test result with RuT5-large initialization

model	Accuracy	Similarity	Fluency	J
Human References	0.888	0.824	0.894	0.653
RuT5 (our)	0.794	0.872	0.903	0.633
RuT5 (baseline)	0.791	0.822	0.925	0.606
Ruprompts (baseline)	0.803	0.703	0.866	0.493
Delete (baseline)	0.387	0.705	0.726	0.162

Table 4: Private test with human evaluation

the fact that the model was originally trained not to generate a sentence from scratch, but only to rewrite the already prepared text. The best result was shown by the RuT5 model. On the other hand, none of these methods, even NAST, which showed state-of-the-art results on GYAFC dataset, could even beat the weakest baseline (bert-delete).

The table 2 shows results for non-autoregressive models with source text initialization. Compared to the previous table, all models except mask-predict showed a significant increase in quality. Probably, the reason for the low performance of mask-predict algorithm is connected to the fact, that large value of iterations for NAR model is set (16 iteration for each text), but this algorithm does not have an early-stopping mechanism. At the same time, easy-first has such an algorithm, so it gives us better results.

Next table 3 shows results for all of our models, which we initialized using the best results with RuT5-large. Additionally, we used the same model RuT5 again on the data from RuT5, with beam-search 1 and 12, but did not get any increase in the quality of the answers. The Disco-easy-first model also did not show any improvement in results. The model refused to complete most of the sentences, considering them as already good, while those sentences it stopped at we got a worse result than the original one. The levT model, although it got an additional increase, did not show high results. The Disco with mask-predict algorithm, like last time, showed a decrease in quality.

Since the automatic metrics (both reference-less classifiers and reference-based metrics) cannot reliably identify the best-performing model, competition organizers also conduct the manual evaluation of the private test set. Our best model (RuT5-large) was additionally tested with human evaluation. The result of this evaluation shown on table 4. Our model got first place, losing only to human evaluation.

6 Error Analysis

The first issue that could affect the quality of the models is the lack of data for paraphrasing. Although the training set has 1 million sentences, the encoder and decoder have the same dictionary, and the model is trained on the distilled data instead of the original data, the quality of the models indicates a lack of training. The second problem has to do with the language itself. The Russian language has a strong morphology, which can ruin the ability to link words to produce toxic-not-toxic pairs. This is the reason why the NAST model score is so low. The NAST model is very related to the generation of such pairs (Figure 4). The third problem is related to the dataset: in GYAFC dataset is much easier to extract individual words, while in RUSSE dataset there are some noisy data, which causes the tokenization to replace from 5% to 10% of all data with <unk>. As can be seen from the examples presented in

src	:-d:-d:-d ой бляя во даёт блоханосец
src tokenized	: _ - _ d _ _ : _ - _ d _ _ : _ - _ d _ ой б _ ляя во даёт бло _ ха _ но _ се _ ц
tgt	Ой, во даёт
tgt tokenized	О _ й _ , во даёт
src	Эй, слющай, чо оскорбляешь срязю?
src tokenized	Эй _ , с _ лю _ щ _ яй _ , чо оскорб _ ля _ ешь с _ ря _ зю _ ?
tgt	Послушай, почему сразу оскорбляешь ?
tgt tokenized	Пос _ лу _ шай _ , почему сразу оскорб _ ля _ е _ ш _ ь ?

Figure 4: BPE tokenization for src and tgt sentences

Appendix A.1 when text is written only in capital letters BPE encoder generates a very large number of <unk> tokens, which prevents the NAR models from generating any meaningful text.

7 Conclusion

This paper presents our approach to detoxification in Russian language, which got J 0.633 and placed 2nd place (1st without Human References) on the private leaderboard. Our paper shows the effectiveness of the method. The error analysis also shows that the proposed approaches can perform better with additional datasets. But there is still the question of why larger models handle the task better. As a simple answer, let’s try to reformulate the problem: detoxification as the sub-task of denoising. The input to the model is noisy text: it includes emoji, text with grammatical errors, toxic words, incorrect syntax, etc. Large models treat toxic words as noisy data, so the bert-delete baseline is relatively successful and does not generate a toxic word from the masked data again, but rather picks up a more neutral word. Yet larger models, such as RuT5, do this with all words. In addition, for such large models, there is no problem with tokenization so texts can be translated into tokenized form fully. This probably explains the high quality of sentence detoxification with the RuT5 model. In the future, we plan to adopt more NAR models for the style-transfer task.

References

- David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. *CoRR*, abs/2109.08914.
- David Dale. 2021. Paraphrasing russian texts: corpus, models, metrics, June. [Online; posted 28-June-2021].
- Daryna Dementieva, Irina Nikishina, Varvara Logacheva, Alena Fenogenova, David Dale, Irina Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. Russe-2022: Findings of the first russian detoxification task based on parallel corpora.
- Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional neural networks for toxic comment classification.
- Jiatao Gu, Changhan Wang, and Jake Zhao. 2019. Levenshtein transformer. *CoRR*, abs/1905.11006.
- Fei Huang, Zikai Chen, Chen Henry Wu, Qihan Guo, Xiaoyan Zhu, and Minlie Huang. 2021. NAST: A non-autoregressive generator with word alignment for unsupervised text style transfer. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: Findings*.
- Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 424–434, Florence, Italy, July. Association for Computational Linguistics.

- Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. 2020. Non-autoregressive machine translation with disentangled context transformer. // *Proc. of ICML*.
- Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 737–762, Online, November. Association for Computational Linguistics.
- Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language.
- Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. // *North American Association for Computational Linguistics (NAACL)*.
- Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. // *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*.
- Colin Raffel, Noam 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. *CoRR*, abs/1910.10683.
- Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 129–140, New Orleans, Louisiana, June. Association for Computational Linguistics.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *CoRR*, abs/1705.09655.
- Kehan Wang, Jiayi Yang, and Hongjun Wu. 2021. A survey of toxic comment classification methods.

A Appendix

A.1 Generated examples

src	В такой ситуации виноваты пидарасы из Ростелекома у которых даже кошка может купить фильм с пульта. Дебилы блять.
t5	в такой ситуации виноваты люди из ростелекома, у которых даже кошка может купить фильм с пульта.
NAST	в такой ситуации виноваты люди из мешай у которых даже кошка может купить фильм с ещё
APE	в такой ситуации виноваты у которых даже кошка может купить фильм с пульта.
DisCo (easy-first)	В этой ситуации виноваты пидарасы из < unk>< unk>Ростелекома< unk>< unk>даже могут купить фильм с пульта
DisCo (mask-predict)	В этой ситуации у кокошки, есть есть кошки из стелекома, у которых даже есть возможность возможность кошка
src	и как земля таких уродов носит?
t5	и как земля их носит?
NAST	и как земля их носит как
APE	Ужие и как земля носит?
DisCo (easy-first)	И как земля, которые носят такие уроды?
DisCo (mask-predict)	И как их переносит, , как земля их носит?

src	а где ты денег столько взяла,уродина?
t5	а где ты денег столько взяла?
NAST	а где ты денег столько взяла
APE	УМИ: а где ты денег столько взяла?
DisCo (easy-first)	А где вы взяли столько денег,
DisCo (mask-predict)	Так где же вы взяли так много а г , что взяла?

src	козел!!!!!! ему бы там пожить. продумано до мелочей,как угро-бить людей за 10 жизни.
t5	ему бы там пожить. продумано до мелочей , как угробить лю-дей за 10 жизни.
NAST	ему бы там не не до не не как не людей за не жизни
APE	ему бы там пожить. продумано до мело, как угробить людей за 10 жизней
DisCo (easy-first)	козел<unk><unk><unk><unk><unk><unk>! ему бы там по-жить. продумано до мело <unk>,как угробить людей за 10 жизни.
DisCo (mask-predict)	Он бы бы там дожить до мельсамых разных блюдней вплоть до мелоза то, что что было продумано до мело<unk>, как оскор-бить людей за 10 лет лет до конца

src	ПРОПУСТИЛА ГОДОВЩИНУ ПОБЕДЫ ЛИТЛ МИКС НА ИКС ФАКТОРЕ:((((СКА ТОЛЬКО Я ТАК МОГ
t5	пропустила слов нет, одни эмоции
NAST	пропустила прям хлеб литл всю на выходные я всю биологию только я так мог
APE	В ДЕке-: : : : —————ВИДЕК<unk>
DisCo (easy-first)	<unk>ПРОПУЧИНЫ <unk><unk>ЕЛЬ МИКС НА ИКС ФАКТОР<unk><unk><unk><unk><unk><unk>СК СКАЛЬКО Я ТАК МО<unk><unk><unk><unk>
DisCo (mask-predict)	<unk><unk><unk><unk>Нет, <unk><unk>я так мог

“Because no one reads prose any more”: grammar and prosody of insubordinated reason clauses

Podlesskaya V. I.
Russian State University for the
Humanities, Moscow, Russia
vi_podlesskaya@il-rggu.ru

Abstract

Based on data from the multimedia subcorpus of the Russian National Corpus, the paper addresses syntactic, pragmatic and prosodic features of insubordinated adverbial clauses introduced by the adverbial subordinator *potomu čto* ‘because’. The quantitative analysis showed that more than 30% of reason clauses in spoken discourse appear to be insubordinated. Qualitatively, we observed symptoms of insubordination at various levels. (1) Prosodically, insubordinated clauses are placed after discourse fragments that are articulated with falling pitch projecting no continuation and are separated from them by the prosodic break. (2) Pragmatically, they can have independent illocutionary force and can form separate turns in dialogues. (3) Grammatically, they allow right dislocation of the adverbial subordinator – otherwise blocked in adverbial clauses.

Keywords: insubordination, adverbial clauses, adverbial subordinators, prosody

DOI: 10.28995/2075-7182-2022-21-456-469

«Потому что больше никто не читает прозу»: грамматика и просодия автономных причинных придаточных по корпусным данным

Подлесская В. И.
РГГУ, Москва, Россия
vi_podlesskaya@il-rggu.ru

Аннотация

На материале мультимедийного подкорпуса НКРЯ рассматриваются синтаксические, прагматические и просодические свойства придаточных с союзом *потому что* в автономных употреблениях. Количественный анализ показал, что в устной речи автономные придаточные, т.е. придаточные, формирующие отдельное высказывание, составляют больше 30% всех вхождений клауз с союзом *потому что*. Качественный анализ показал, что такие употребления (1) располагаются после фрагмента, реализованного с интонацией завершенности, и отделены просодическим швом; (2) могут иметь иллокутивную силу, не совпадающую с иллокутивной силой смежных дискурсивных фрагментов, и формировать самостоятельную реплику в диалоге; и (3) допускают дислокацию союза вправо, что в неавтономных употреблениях оказывается невозможным.

Ключевые слова: расподчинение, обстоятельственные придаточные, подчинительные союзы, просодия

1. Постановка вопроса

Потому что так чудно под ветром вспухает штора
И в широкую щель пробивается звездный свет,
Потому что мы, кажется, сможем проверить скоро,
Рухнет мир без романов и вымысла или нет?

Александр Кушнер
«Потому что больше никто не читает прозу»

В работе рассматриваются иллокутивно автономные употребления причинных придаточных с союзом *потому что*, т.е. такие употребления, в которых придаточное формирует отдельное высказывание. Минимальная степень автономизации наблюдается при парцелляции постпозитивного придаточного – главное предложение произносится с интонацией завершенности, не формирующей у слушающего ожидание продолжения; на письме такая парцелляция передается точкой (восклицательным или вопросительным знаком) после главного предложения, и придаточное начинается с заглавной буквы:

(1) *У Пети по утрам голова болит. Потому что у него бессонница.*

Возможна дальнейшая автономизация – главное и придаточное могут формировать реплики, принадлежащие разным говорящим, причем с помощью придаточного говорящий может или самостоятельно выдвинуть соображение о причинах события, описанного собеседником, см. (2а), или ответить на вопрос о причине, заданный собеседником, см. (2б), или даже задать вопрос о правильности предполагаемой причины, см. (2в):

(2а) *А: У Пети по утрам голова болит. Б: Потому что у него бессонница.*

(2б) *А: Почему у Пети по утрам голова болит? Б: Потому что у него бессонница.*

(2в) *А: У Пети по утрам голова болит. Б: Потому что у него бессонница?*

Далее, клауза типа *Потому что у него бессонница* может следовать не после единичной клаузы, а после целого эпизода, описывающего целый комплекс неприятных явлений, вызванных причиной, обозначенной в придаточном, данном случае – бессонницей:

(3) *У Пети плохой аппетит. Вчера он отказался идти на работу. Все жалуется на его раздражительность и все такое... Потому что у него бессонница.*

Наконец, в принципе, возможна такая ситуация, когда главная клауза или семантически эквивалентный ей фрагмент дискурса вовсе отсутствует. В таких случаях союз утрачивает свойства коннектора и превращается в дискурсивный маркер с тем или иным, обычно модальным, значением. Например, условный союз *если* в сочетании с частицей *бы* в составе автономной клаузы приобретает значение желательности, ср. *Вот если бы удалось выспаться!* (в значении «хорошо бы выспаться»). В таких случаях говорят о явлении полного расподчинения (insubordination), Evans 2007, Добрушина 2019, Maschler 2020. Для союза *потому что* такая крайняя степень автономизации, по-видимому, не характерна, однако более мягкие варианты автономизации, продемонстрированные в примерах (1) - (3) наблюдаются массово. Они и станут объектом исследования в данной работе. При описании грамматических свойств таких конструкций мы будем опираться на данные основного корпуса НКРЯ, при описании их просодических свойств – на данные мультимедийного подкорпуса (МУРКО).

Необходимо оговориться, что мы обсуждаем только так называемый нерасчлененный вариант союза. За пределами работы остается его расчлененный вариант, при котором компонент *потому* и компонент *что* размещаются в разных коммуникативно-просодических составляющих, причем компонент *потому* может сдвигаться внутрь главной клаузы и становиться акцентоносителем:

(4) *У Пети потому по утрам голова болит, что у него бессонница.*

Конструкции такого рода заслуживают в перспективе отдельного исследования.

Дальнейшее изложение будет строиться следующим образом. В разделе 2 мы обсудим частотность автономизации постпозитивного придаточного с союзом *потому что* в сопоставлении с другими средствами выражения причинно-следственной связи между клаузами. Раздел 3 посвящен качественному анализу иллокутивно автономных причинных придаточных с *потому что* – мы продемонстрируем грамматические и просодические контексты, в которых эти конструкции наблюдаются. В разделе 4 обсуждается одно частное свойство автономных клауз с *потому что* – они допускают дислокацию союза в правую периферию придаточного (*У него бессонница потому что*). В разделе 5 будут подведены итоги.

2. Частотность иллокутивно автономных причинных придаточных

Придаточные с *потому что* обладают следующей важной особенностью: как было показано в работе Апресян, Пекелис 2012, они вводят информацию о причинной связи как новую для слушающего, и потому придаточное тяготеет к рематическому статусу и заключительной позиции. Так, предложение (5) с препозицией придаточного значительно менее естественно, чем приведенный выше вариант с постпозицией (4):

(5) ?*Потому что у Пети бессонница, у него по утрам голова болит.*

Это ограничение не является абсолютным. При определенной поддержке контекста препозиция придаточного с *потому что* становится возможной. В частности, ритмика и стилистика поэтической речи (которая обычно более лояльна к отклонениям от стандарта, чем проза) может облегчить препозицию. Так в стихотворении И.Бродского «Песня невинности, она же – опыта» препозиция придаточного с *потому что* используется несколько раз, как повторяющийся художественный прием:

(6)

*Потому что у куклы лицо в улыбке,
мы, смеясь, свои совершим ошибки.*

....

*Потому что душа существует в теле,
жизнь будет лучше, чем мы хотели.*

Препозиция придаточного с *потому что* облегчается, по-видимому, если оно попадает в сферу действия вопросительного оператора (частиц *ли, или, что ли*), ср. строки стихотворения С.Есенина:

(7)

*Шаганэ ты моя, Шаганэ!
Потому что я с севера, что ли,
Я готов рассказать тебе поле,
Про волнистую рожь при луне.*

Примеры, типа приведенных выше, допустимы, но не составляют ядра употреблений придаточных с *потому что*, ядром являются предложения с постпозицией придаточного. Это делает возможным использовать в качестве инструмента оценки частотности автономных употреблений придаточного такую метрику, как частотность употребления *потому что* с заглавной буквы – поскольку основным массивом употреблений «заглавного» *потому что* оказываются случаи иллокутивно автономной клаузы. Это могут быть парцелированные причинные придаточные, причинные придаточные во вставке, самостоятельные реплики в диалоге, клаузы, вводящие причину не непосредственно к предшествующей клаузе, а к более крупному дискурсивному фрагменту (см. подробнее раздел 3).

Ниже мы сравним общую частотность *потому что* и частотность заглавного *потому что* в основном корпусе и в МУРКО. Далее, чтобы понять, насколько обнаруженные соотношения характеризуют индивидуальные свойства союза, а насколько – в целом, семантический тип придаточных, мы сравним частотности строчного и заглавного написания для некоторых других коннекторов причинно-следственной зоны.

К сожалению, выбранная нами метрика не годится для оценки склонности к автономизации клауз с наиболее близкими к *потому что* союзами *так как* и *поскольку*: как показано в (Апресян, Пекелис 2012), союзы типа *так как* или *поскольку* могут вводить уже известную причину, потому соответствующая клауза оказывается уместной и в препозиции к главной ср. гораздо более уместные, чем с *потому что*, варианты примера (5) с *так как* или *поскольку*:

(8) *Поскольку/ Так как / Потому что у Пети бессонница, у него по утрам голова болит*

В результате количество этих союзов, написанных с заглавной буквы, не коррелирует с иллокутивной автономизацией.

Более показательное сравнение с союзом следствия *так что*. Этот союз традиционно квалифицируется в русских грамматиках, как подчинительный, однако, как было показано в Пекелис 2015а,б, по реакциям на стандартные синтаксические тесты он является сочинительным. И как таковой не может вводить первую клаузу в предложении. Там же в Пекелис 2015а,б показано, что синтаксически сочинительным является и причинный союз *ибо*, поэтому, естественно, вводимая им причинная клауза всегда расположена справа от клаузы, обозначающей следствие. Анафорическое союзное наречие следствия *поэтому* тоже вводит клаузу, которая всегда располагается правее фрагмента дискурса, к которому отсылает. Таким образом, для коннекторов *так что*, *ибо* и *поэтому* – также как и для *потому что* – написание с заглавной буквы в большинстве случаев говорит о том, что мы имеем дело с вводимой этими коннекторами иллокутивно автономной клаузой.

В таблице 1 сведены количественные данные о частотности этих единиц в целом и частотности их написания с заглавной буквы в основном корпусе и в МУРКО. Для каждой из них приведено (а) общее число вхождений; (б) число вхождений, приведенное к объему соответствующего корпуса (число вхождений на один миллион слов, ipm), (в) число тех вхождений, где единица записана с заглавной буквы (признак “capital” в запросе); и (г) доля «заглавных» вхождений от общего числа вхождений. Союзы *потому что* и *так что*, учитывались только в нерасчлененных вариантах, для чего в запросе первые части союза (*потому* и *так*) снабжались ограничением «не перед знаком препинания» (признак “-bmark” в запросе), а вторая часть (*что*) снабжались ограничением «не после знака препинания» (признак “-amark” в запросе):

	Основной корпус	МУРКО
Всего слов в корпусе на момент запроса	337 025 184	5 449 075
<i>потому</i> , -bmark на расстоянии 1 от <i>что</i> , -amark	147602	6682
ipm <i>потому что</i>	437.96	1226.26
<i>потому</i> , -bmark & capital на расстоянии 1 от <i>что</i> , -amark	20246	2382
Доля <i>потому что</i> с заглавной буквы от общего числа вхождений <i>потому что</i>	13.72%	35.65%
<i>так</i> , -bmark на расстоянии 1 от <i>что</i> , -amark	60 251	1626
ipm <i>так что</i>	178.77	295.69
<i>так</i> , -bmark & capital на расстоянии 1 от <i>что</i> , -amark	23 633	1200
Доля <i>так что</i> с заглавной буквы от общего числа вхождений <i>так что</i>	39.22%	74.46%
<i>ибо</i>	72 170	167
ipm <i>ибо</i>	214.14	30.65
<i>ибо</i> , capital	10605	67
Доля <i>ибо</i> с заглавной буквы от общего числа вхождений <i>ибо</i>	14.69%	40.12%
<i>поэтому</i>	100 384	3016
ipm <i>поэтому</i>	297.85	553.49
<i>поэтому</i> , capital	51 329	1 539
Доля <i>поэтому</i> с заглавной буквы от общего числа вхождений <i>поэтому</i>	51.13%	51.03%

Таблица 1: Частотность союзов

Первое важное наблюдение, которое следует из приведенных в Таблице 1 данных: в устной речи (по крайней мере, в тех ее регистрах, которые представлены в МУРКО) очень высока доля иллокутивно автономных клауз, открывающихся маркерами причинно-следственной связи. Для союза *потому что* она составляет 35.65%, для трех других маркеров она еще выше, достигая максимума в 74.46% для союза *так что*.

Второе наблюдение – доля автономных причинных клауз в два-три раза выше в устном дискурсе, чем в письменном: 35.65% против 13.72% для *потому что*, 74.46% против 39.22% для *так что*, 40.12% против 14.69% для *ибо*. Исключением считается союзное наречие *поэтому*: возглавляемые им клаузы столь же часто автономизируются в письменном дискурсе, как и в устном – примерно в половине всех вхождений. По-видимому, причина в том, что клауза с *поэтому* связывается с контекстом только семантически (дискурсивно), это – не сильная грамматическая связь. Это не подчинение – об этом свидетельствует, например, категорический запрет на препозицию клаузы с *поэтому*, но это и не сочинение – об этом свидетельствует допустимость сочетания *поэтому* с сочинительными союзами (*и поэтому*), тогда как для прототипических сочинительных союзов такое сочетание запрещено (**и а*, **и но*). Кроме того, *поэтому*, в отличие прототипических сочинительных союзов, не обязано занимать крайнюю левую позицию в «своей» клаузе, ср. *у него поэтому по утрам голова болит*, *у него по утрам поэтому голова болит*.

В следующем разделе мы рассмотрим актуальные употребления автономных клауз с *потому что* в устном дискурсе по данным МУРКО.

3. Грамматика и просодия иллокутивно автономных клауз с *потому что*

В общем случае, если постпозитивное придаточное хорошо интегрировано в сложное предложение, то можно наблюдать следующую просодическую картину: (а) в главной клаузе слово, которое несет фразовый (рематический) акцент, реализуется с резким и существенным подъемом тона по типу ИКЗ в терминах интонационных конструкций (Брызгунова 1982),

который является стандартным маркером просодической незавершенности, т.е. сигналом для слушающего ожидать продолжения; и (б) между главным и придаточным нет выраженной просодической границы (нет так называемого ресета, т.е. нет возвращения тона на средний уровень для данного говорящего), а также нет длительной паузы. В качестве иллюстрации приведем пример (9). 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 акцентов¹ и интонограмма в формате анализатора PRAAT:

(9) *И с точки зрения эволюции вида это хорошо/ потому что нет дополнительного травмирования.*
[Вячеслав Дубынин. Мозг и агрессия (2017) // <https://postnauka.ru/video/82023>]

И с точки зрения эволюции /вида,
это /хорошо,
потому что /нет дополнительного \травмирования.

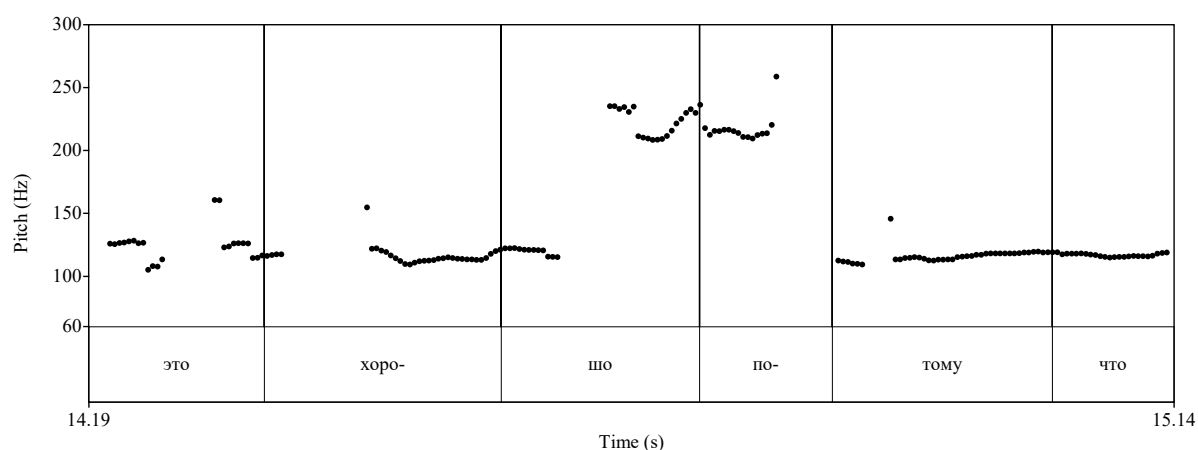


Рисунок 1: Интонограмма к примеру (9)

Как видим, носитель фразового акцента главной клаузы, слово *хорошо*, реализуется с резким подъемом на ударном слоге, союз *потому что* начинается непосредственно с того же высокого уровня частоты основного тона, и между *хорошо* и *потому что* нет даже минимальной паузы. Все это вместе обеспечивает просодическую целостность сложного предложения.

Если же придаточное парцеллируется, т.е. становится иллокутивно автономным, то мы наблюдаем обратную просодическую картину:

(а) и в главной клаузе, и в зависимой клаузе слова, несущие фразовый (рематический) акцент, реализуются с просодией завершенности; это может быть падение тона по типу ИК1 в терминах интонационных конструкций (Брызгунова 1982), которое является стандартным маркером просодической завершенности, т.е. сигналом для слушающего, что продолжение не ожидается, или падение по типу ИК2 (Янко 2001: 92-97) – интенсивное падение, маркирующее контраст либо особые иллокутивные значения, в первую очередь, значение причины/мотивации;

(б) между главным и придаточным есть выраженная просодическая граница (есть ресет), а также нередко имеется пауза. В качестве иллюстрации приведем пример (10):

¹ 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, Подлеская (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.

(10) Но всё-таки из всех мастеров для примера я выбрал себя, потому что себя самого я знаю лучше других. [М.М. Пришвин. Мои тетрадки (1953) // ImWerden/prishvin_moi_tetradki.mp3]

Но /всё-таки,

из всех /мастеров для /примера я выбрал \себя.

(0.35)

Потому что /себя самого (0.41) я знаю лучше \других.

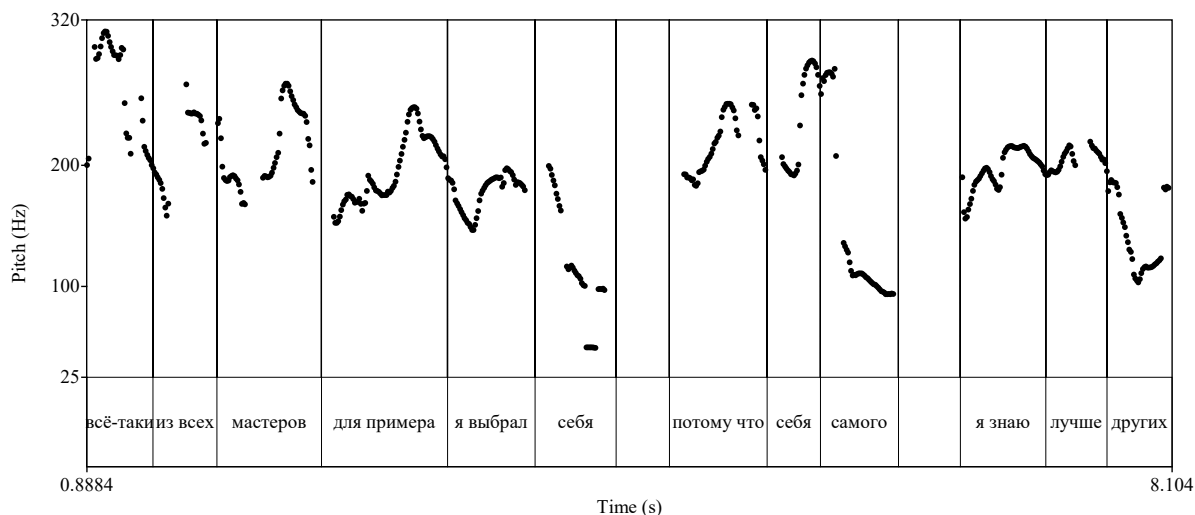


Рисунок 2: Интонограмма к примеру (10)

В (10) носитель фразового акцента главной клаузы слово *себя* и носитель фразового акцента причинной клаузы слово *других* реализуются с резким падением на ударном слове по типу ИК2, союз *потому что* начинается не с того уровня тона, где завершается главная клауза: он возвращается на средний для данного говорящего уровень, типичный для начала высказывания, т.е. происходит ресет. Кроме того, перед *потому что* имеется пауза. Все это вместе обеспечивает иллюкутивную автономность причинного придаточного.

Заметим также, что в примере (10) имеется любопытная для нашего исследования особенность: несмотря на очевидные просодические признаки автономизации, в графической версии МУРКО союз записан не с заглавной, а со строчной буквы. Причина по видимому в том, что перед нами не запись устного текста транскрайбером, а воспроизведение письменного оригинала – М.М.Пришвин на этой записи читает вслух собственный текст. На письме, как мы уже видели, написание союза с большой буквы встречается заметно реже – по-видимому, работает стереотипная пунктуационная схема «запятая перед союзом», даже если пишущий ощущает ослабление дискурсивной связи между клаузами. Но при чтении появляется дополнительный ресурс – просодия, и ослабление связи маркируется просодией. При чтении вслух этот феномен наблюдается довольно часто. Таковы, например, случаи (11) и (12). В (11) Ф.А.Искандер читает авторский текст с выраженным падением на слове *мог*; в (12) Б.Н.Ливанов читает из «Героя нашего времени» с выраженным падением на слове *моя*, т.е. причинные придаточные просодически автономизируются, несмотря и запятую в оригинале:

(11) Но и оскорбить его в самом застолье не мог, потому что не был уверен, что и другие заметили эти взгляды. [Ф.А. Искандер. Пастух и козуля (1990-2004) // ImWerden]

(12) Я нанял нашу духаницу: она знает по-татарски, будет ходить за нею и приучать её к мысли, что она моя, потому что она никому не будет принадлежать, кроме меня [М.Ю. Лермонтов. Бэла (исп. Б.Н. Ливанов) (1953) // «Старое радио» <http://www.staroradio.ru/audio/10187>]

Не отраженная в транскрипте автономизация придаточного с *потому что* обнаруживается в МУРКО не только при чтении вслух письменного источника. Иногда транскрайберы просто не замечают просодических симптомов автономизации и по инерции используют стереотипную пунктуационную схему «запятая перед союзом», как в примере (13):

(13) Я выбил шпагу из рук этого мерзавца. Не добил, потому что разняли. [Евгений Никишов, Карен Шахназаров, Сергей Рокотов. *Исчезнувшая империя*, к/ф (2008)]

Не ^добил!

(0.16)

Потому что (0.16) ^разняли!

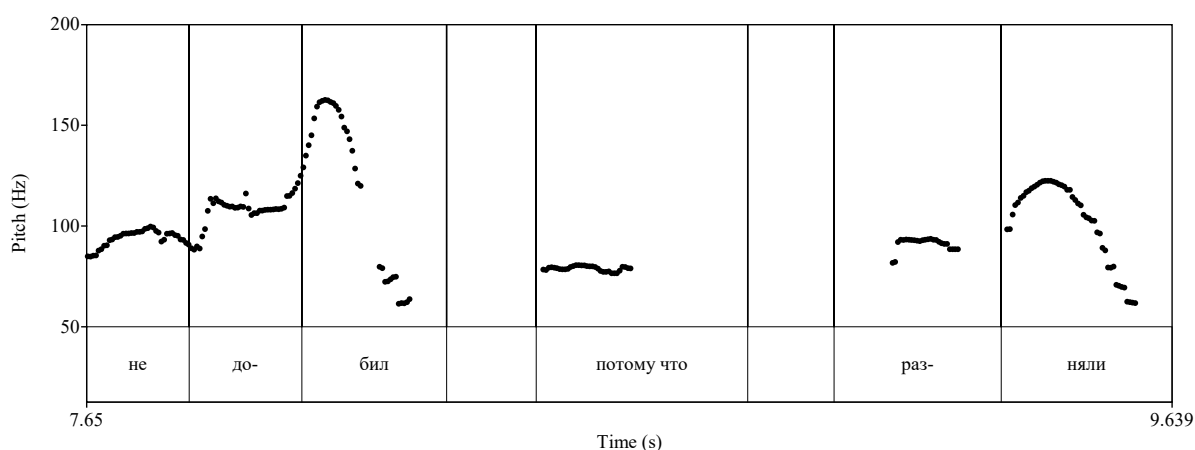


Рисунок 3: Интонограмма к примеру (13)

В (13) акцентоносители и в главном, и в придаточном (*добил* и *разняли*) реализуются с резким падением с предварительным заходом вверх по типу ИК2. Кроме того, придаточное начинается с ресета, и перед ним есть пауза. Таким образом, налицо основные просодические симптомы автономизации, но в транскрипте это не отражено. Т.е. фактически, автономные придаточные представлены в МУРКО еще чаще, чем можно судить по доле написаний союза с заглавной буквы.

Просодические признаки автономии придаточного могут дополняться признаками дискурсивного разрыва на границе придаточного. Наиболее частая причина дискурсивного разрыва в монологе – смена типа иллокуции, в диалоге – смена говорящих (о возможной иллокутивной неоднородности компонентов полипредикативной конструкции, см. Кобозева 1999). Приведем примеры.

В (14) автономное придаточное появляется после встроенной в текущий нарратив прямой речи, формирующей отдельную цепочку иллокуций. Прямая речь просодически не проецирует продолжения, после нее – длительная пауза, придаточное начинается с ресета частоты основного тона. Углублению дискурсивного разрыва способствует и то, что причинная клауза связывается не пропозициональным отношением, а так называемым «эпистемическим» в терминах Sweetser 1990 или «метатекстовым» в терминах Падучева 2009 – клауза предьявляет не причину положения дел, а обоснование вывода говорящего (я считаю, что она притворилась, на том основании, что она сказала то-то и то-то):

(14) *Значит/ она притворилась/ когда сказала/ «Что ты/ Павел/ я не собираюсь замуж!».* Потому что она мне самому говорила/ «Привези мне какого-нибудь пожилрого». [Василий Шукиин. *Живет такой парень*, к/ф (1964)]

\Значит,
 /она \притворилась,
 /когда \сказала:
 ..(0.34)
 «/Что ты \Павел!
 ..(0.24)
 Я не \собираюсь замуж.»
 ... (1.20)
 Потому что она мне \самому говорила:
 ... (0.59)
 «/Привези мне какого-нибудь пожилото́го!».

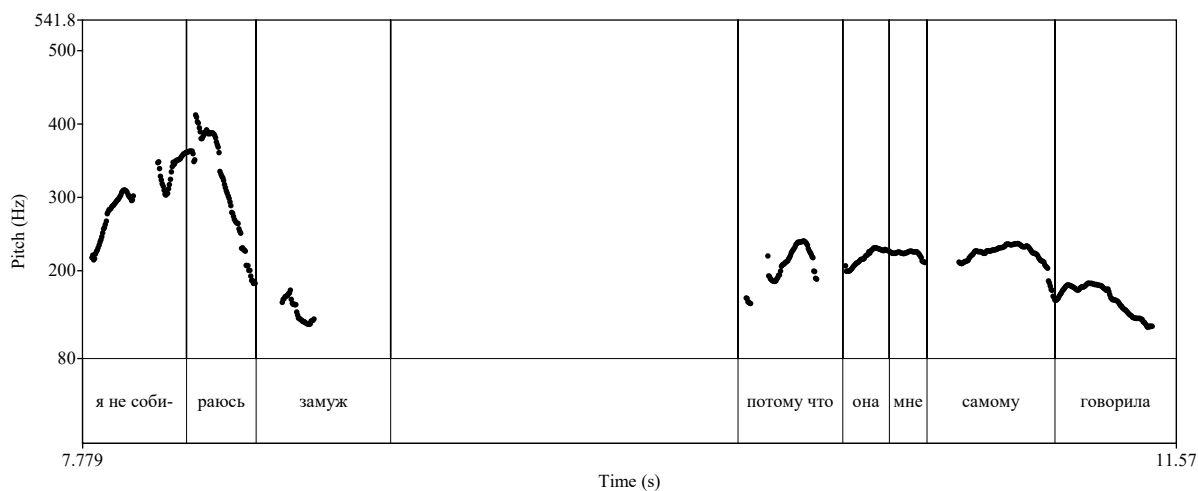


Рисунок 4: Интонаграмма к примеру (14)

В примере (15) – говорящий задает вопрос, затем уточняет его. Уточнение реализуется с интонацией завершенности, не предвещающей продолжения. Но уже *post hoc* у говорящего возникла необходимость объяснить, почему он задет вопрос – возникает автономное придаточное, в котором вводится «речеактная причина» в терминах Sweetser 1990, т.е. здесь мы тоже имеем дело не с базовым, пропозициональным причинным значением, отсюда – углубление дискурсивного разрыва:

(15) Скажите/ будут ли новые проекты какие-нить в этой сфере/ ну и в частности/ развитие медицинского туризма. Пту что не секрет/ что половина Москвы едет в Рязань и зубы делать/ и глаза лечить. [Пресс-конференция Николая Любимова (2017)]

\Скажите;
 будут ли новые /проекты какие-нибудь в этой сфере?
 Ну и в /частности-и ..(0.37) /развитие-е медицинского \туризма.
 ..(0.1)
Потому что не /секрет,
 что половина /Москвы едет в Рязань и /зубы делать,
 и \глаза лечить.

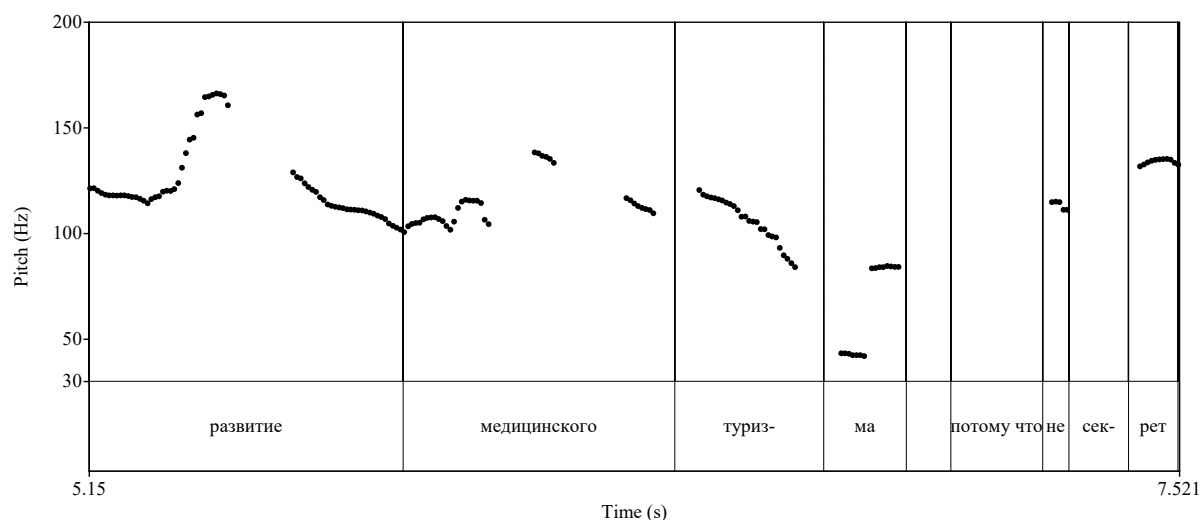


Рисунок 5: Интонограмма к примеру (15)

Наконец, в примерах (16)-(19) придаточное причины является иллокутивно автономным ответом на вопрос – заданный говорящим самому себе с риторической целью, как в (16), (17) или заданный собеседником, как в (18), (19):

(16) *Почему мы выбрали/ скажем/ религию в виде православия Византийского? Да потому что контакты были с Византией близкие/ постоянные...* [Андрей Сахаров. *Россия как часть мирового цивилизационного процесса. Проект Academia (ГТРК Культура) (2010)*]

(17) *Почему? Да потому что лингвист-любитель с полной серьёзностью может вам сообщить/ что он прочёл этрусскую надпись по-русски...* [Андрей Зализняк. *Что такое любительская лингвистика? Лекции Полит.ру (2010)*]

(18) [№ 2, муж, лингвист] *Ещё один маленький вопрос: а всегда ли есть эта семантическая разница/ например вот в парах типа «усталый» и «уставший»? Я её не вполне вижу мм... разницы.* [Г.И. Кустова, жен, 52, 1960, лингвист] *Ну/ потому что «усталый» – это/ строго говоря/ тоже как бы причастие.* [Г.И. Кустова. *Ответы на вопросы после доклада на конференции Диалог 2012 (2012) // Из коллекции НКРЯ*]

(19) [Таня (Ирина Розанова), жен, 45, 1961] *Видишь ли/ плохого тут ничего нет/ просто я знаю тебя/ твоего Никиту/ твою жизнь. Я боюсь/ что ты/ сама того не замечая/ отдалиться от всего этого/ старое разрушишь/ а нового не построишь.* [Нина (Анна Михалкова), жен, 32, 1974] *Это почему?* [Таня (Ирина Розанова), жен, 45, 1961] *Да потому что он не нашего круга.* [Авдотья Смирнова. *Связь, к/ф (2006)*]

Заметим дополнительно, что в автономных ответах на вопрос союзу *потому что* регулярно предшествуют частицы (*ну, да* и др.); в этом случае автономное придаточное также не попадает в выборку, полученную по запросу «союз с заглавной буквы». Не попадает в выборку и еще один очень важный класс автономных клауз с *потому что*, который будет рассмотрен в следующем разделе – конструкции с дислокацией союза вправо (*У него бессонница потому что*)

4. Дислокация союза в иллокутивно автономном придаточном

Союз *потому что* может располагаться не только в своей прототипической начальной позиции в придаточном, но и сдвигаться вправо. Такого рода дислокация допустима далеко не для всех союзов. Так, функционально наиболее близкие к *потому что* причинные союзы *так как*, *оттого что*, *из-за того что*, *поскольку* не допускают передвижения:

(20) *У него часто голова болит. Бессонница потому что / *так как, *оттого что, *из-за того что, *поскольку.*

Допустимость передвижения *потому что* определяется иллокутивной автономностью придаточного. В основном корпусе НКРЯ по запросу «*потому*, -amark на расстоянии 1 от *что*, bdot»², т.е. *потому* не после знака препинания плюс *что* перед точкой, выдается 90 примеров. Их сплошная проверка показывает, что в абсолютном большинстве случаев передвижение происходит внутри придаточного, следующего после точки:

(21) *Муж выслал из Москвы за сто первый километр, так она теперь здесь развлекается... — Как это — «выслал»? — Да очень просто. Довела потому что.* [Сергей Шикера. Выбор натуры // «Волга», 2014]

(22) *Он сам хотел, — отрезает Пол, — ты ж помнишь, его не отговорить было. Эгоист потому что.* [Александра Тайц. Настоящее лето // «Сибирские огни», 2012]

Кроме того, придаточные с передвижением массово используются в качестве отдельной реплики, в которой вводится причина или обоснование, обычно — в качестве ответа на прямой или подразумеваемый вопрос:

(23) *Почему ты смеешься? — озадаченно спросил он. — Хорошо потому что.* [Петр Акимов. Плата за страх (2000)]

(24) *Толкую своему приятелю, что меньше, мол, все-таки пить-то стали, и он, являясь зятем отравившегося зельем механизатора, говорит: — Боятся потому что...* [Виктор Астафьев. Затеся // «Новый Мир», 1999]

На тот же запрос применительно к МУРКО выдается 17 примеров. Сплошная проверка показывает, что во всех этих случаях передвижение происходит в иллокутивно автономных клаузах. В примере ниже две клаузы, связанные причинным отношением, не интегрированы, они разделены заметной паузой и произносятся с одинаковым просодическим рисунком — с резким падением в высоком регистре по типу ИК2 на ударном слоге эмфатически выделенного слова, несущего главный фразовый акцент (*равно* и *разговор*, соответственно):

² Разумеется, такой запрос показывает не все случаи линейного передвижения союза, «за бортом» остаются (немногочисленные!) примеры, где союз передвигается не в самую крайнюю правую позицию. Запрос

«-(CONJ | PART), -bmark & -amark & -capital на расстоянии 1 от *потому*, -bmark & -amark & -capital на расстоянии 1 от *что*, -bmark», т.е. *потому* не после союза или частицы и не после и не перед знаком препинания плюс *что* не перед знаком препинания, позволяет обнаружить примеры типа (i), (ii), но подсчет их доли требует ручной работы, потому что они «тонут» в море ложной выдачи (прежде всего, случаев, где ввиду пунктуационной ошибки отсутствует какой бы то ни было знак препинания перед союзом):

(i) *Я всю жизнь работаю, я всю войну девчонкой в госпиталях, и ничего. У меня потому что всегда на первом плане был долг, семья.* [М. М. Роцин. Валентин и Валентина (1970)] [омонимия не снята]

(ii) *В космос, получается, нас запустили люди неместные и ненормальные. Не нашлось потому что здоровых скептиков, чтоб переключить менталитет на земное, чтоб сделать им иужинское «срезал».* [Игорь Мартынов. Шовинист // «Столица», 1997.07.29]

(25) [Валентин (Игорь Лифанов), муж, 41, 1965] Чё-т не больно рады нам здесь.
 [Лаврентий (Юсуп Бахшиев), муж, 41, 1965] Мы всё равно присядем. Разговор имеем потому что.
 [Денис Нейманд, Константин Мурзенко, Юсуп Бахшиев. Жесть, к/ф (2006)]

Мы всё **равно** присядем!
 ..(0.49)
 Разговор имеем потому что!

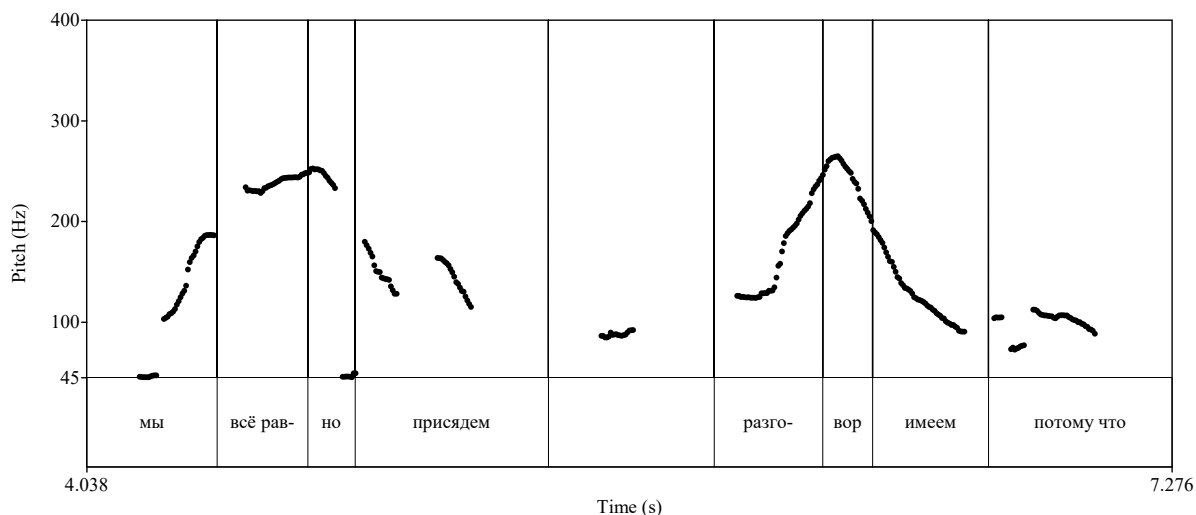


Рисунок 6: Интонаграмма к примеру (25)

Таким образом, корпусные данные убедительно показывают, что у причинных клауз с *потому что* просодическая автономность, может подкрепляться особыми ограничениями на порядок слов.

5. Итоги и некоторые дополнительные штрихи к портрету союза *потому что*

Мы постарались показать, что и в письменном, и – в особенности – в устном дискурсе автономизация причинного придаточного это не маргинальное явление. Независимые клаузы с союзом *потому что* в устном дискурсе составляют около 35% от общего числа вхождений. Количественная оценка была произведена по метрике «количество написаний союза с заглавной буквы», но анализ конкретного материала показал, что, фактически, доля независимых употреблений еще выше. Автономизация проявляется на разных языковых уровнях: на уровне просодии – автономная клауза появляется после интонационно завершенного фрагмента и просодического шва; на уровне дискурса – автономная клауза может иметь иную иллокутивную силу, чем соседние дискурсивные фрагменты, и может формировать отдельную реплику в диалоге; на уровне грамматики – автономная клауза имеет особые ограничения на порядок слов, допуская дислокацию союза вправо.

Упомянем еще один заслуживающий внимания факт. Союз *потому что* может выступать и как изолированная ответная реплика *Потому что!* в ситуации, когда говорящий отказывается ответить на вопрос с вопросительным словом *почему*, (ср. притказку *Потому что потому, что кончается на «У»!*), как в (26), или временно откладывает ответ на этот вопрос, разъясняя причину в следующем дискурсивном шаге, как в (27):

(26) [Феликс (Александр Ширвиндт), муж, 34, 1934] *Итак/ правда первая. Вы не возьмёте меня обратно в отдел?*

[Евдокимов (Александр Лазарев), муж, 30, 1938] *Нет.*

[Феликс (Александр Ширвиндт), муж, 34, 1934] *Почему?*

[Евдокимов (Александр Лазарев), муж, 30, 1938] *Потому что. [Георгий Натансон, Эдвард Радзинский. Еще раз про любовь, к/ф (1968)]*

(27) [Аня (Марина Неелова), жен, 33, 1947] *Ну почему пошутил?*

[Марина (Татьяна Божок), жен, 23, 1957] *Ну потому что. Посмотри на себя и посмотри на него.*

Какой он жених? Толстый/ лысый... И... воображает из себя... [Александр Бородинский, Иван Киасашивили, Карен Шахназаров, Лев Славин. Дамы приглашают кавалеров, к/ф (1980)]

Это уникальное свойство данного союза. В сходных ситуациях с другими вопросительными словами в ответной грубой реплике с отказом выступают рифмованные местоименные наречия, а не союзы, ср. *Когда? – Тогда!, Куда? – Туда!, Как? – Так!*. Возможно, причина кроется в том, что соответствующие адвербиальные клаузы вводятся союзами (*куда, когда, так*), которые омонимичны вопросительным словам, и игровой эффект рифмы пропадает. В целом, способность союза формировать изолированную реплику – это его индивидуальное словарное свойство. Так, например, среди сочинительных союзов союз *но* может формировать отдельную восклицательную реплику (*Но!*), союз *и*, особенно с удлинением и нисходяще-восходящей интонаций по типу ИК4, может формировать вопросительную реплику (*И-ии?*), в то же время союз *а* в качестве отдельной реплики выступать не способен (не путать с омонимичным междометием переспроса).

Таким образом, реальный узус, проявленный в корпусных данных, и особенно, в данных устной речи позволяет постепенно заменить монохромный скетч словарного описания союза на живописный портрет.

Благодарности

Работа поддержана грантом РФФИ № № 22-18-00120.

Я чрезвычайно признательна рецензентам за ценные замечания и проницательные комментарии; критику постаралась максимально учесть.

Литература

Апресян В.Ю., Пекелис О.Е. Подчинительные союзы. Материалы для проекта корпусного описания русской грамматики (<http://rusgram.ru>). На правах рукописи. М., 2012.

Aprėsjan V.J., Pekelis O.E. Podčinitel'nye sojuzy [Complementizers and adverbial subordinators]. Materials for the Russian corpus grammar project (<http://rusgram.ru>). Ms. Moscow, 2012.

Брызгунова Е. А. Интонация, Русская грамматика, том 1, М.: Наука, 1982. — С. 98–118.

Bryzgunova E. A. Intonation [Intonatsija], Russian Grammar [Russkaja grammatika]. Vol. 1, Moscow: Nauka, 1982. — P. 98–118.

Добрушина Н. Р. [Рец. на:] Evans N., Watanabe H. (eds.), Insubordination. Amsterdam: John Benjamins, 2016, xii, 435 pp. / Typological Studies in Language, 115 // Вопросы языкознания. — 2019. — 5. — С.154–159.

Кибрик А.А., Подлеская В.И. (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 2009.

Kibrik A. A., Podlesskaya V. I. [Eds.] Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki Slavjanskix Kul'tur, 2009.

Кобозева И.М. О критериях иллюкутивной самостоятельности частей сложного предложения // Труды международного семинара по компьютерной лингвистике и ее приложениям “Диалог-99”. Таруса, 1999. — С. 133-137.

- Kobozeva I. M. O kriterijah illokutivnoj samostojatel'nosti chastej slozhnogo predlozhenija [On illocutionary independence of clauses in complex sentences] // Trudy mezhdunarodnogo seminara po kompjuternoj lingvistike i ejo prilozhenijam «Dialog-99» [Proceedings of the International Seminar on computational Linguistics and its applications "Dialog-99"]. Tarusa, 1999. — P. 133–137.
- Падучева Е.В. Модальность сквозь призму дейксиса // Е.В.Падучева. Статьи разных лет. М.: Языки славянских культур, 2009. — С. 463-476.
- Padučeva E.V. Modal'nost' skvoz' prizmu dejksis [Modality through the lens of deixis] // E.V.Padučeva. Stat'i raznyx let. Moscow: Jazyki Slavjanskix Kul'tur, 2009. — P. 463-476.
- Пекелис О.Е. Сочинение и подчинение. Материалы для проекта корпусного описания русской грамматики (<http://rusgram.ru>). На правах рукописи. М., 2015a
- Pekelis O.E. Sočinenie i podčinenie [Coordination and subordination]. Materials for the Russian corpus grammar project (<http://rusgram.ru>). Ms. Moscow, 2015a.
- Пекелис О.Е. Причинные придаточные. Материалы для проекта корпусного описания русской грамматики (<http://rusgram.ru>). На правах рукописи. М., 2015b.
- Pekelis O.E. Pričinnye pridatočnye [Reason clauses]. Materials for the Russian corpus grammar project (<http://rusgram.ru>). Ms. Moscow, 2015b.
- Янко Т. Е. Коммуникативные стратегии русской речи. Москва: Языки славянских культур, 2001.
- Janko T. E. Kommunikativnye strategii russskoj reči [Communicative strategies in spoken Russian]. Moskva: Jazyki Slavjanskix Kul'tur, 2001.
- Evans N. Insubordination and its uses // Irina Nikolaeva (ed.), Finiteness: Theoretical and Empirical Foundations. Oxford: OUP, 2007. — P. 366–431.
- Maschler Y. The insubordinate – subordinate continuum: Prosody, embodied action, and the emergence of Hebrew complex syntax // Maschler Yael, Simona Pekarek Doehler, Jan Lindström and Leelo Keevallik (Eds.) Emergent Syntax for Conversation. Clausal patterns and the organization of action. John Benjamins, 2020. — P. 87-125.
- Sweetser, E. E. From etymology to pragmatics. Metaphorical and cultural aspects of semantic structure. Cambridge: Cambridge University Press, 1990.

Artificial text detection in Russian language: a BERT-based Approach

Posokhov P. A.
ITMO University,
Saint Petersburg, Russia
paposokhov@itmo.ru

Skrylnikov S. S.
ITMO University,
Saint Petersburg, Russia
skrylnikovs@itmo.ru

Makhnytkina O. V.
ITMO University,
Saint Petersburg, Russia
makhnytkina@itmo.ru

Abstract

This paper describes our solution for the RuATD (Russian Artificial Text Detection) competition held within the Dialog 2022 conference. Our approach is based on the idea of transfer learning, using pre-trained RuRoBERTa, RuBERT, RuGPT3, RuGPT2 models. The final solution included Byte-level Byte-Pair Encoding tokenization, and a fine-tuned model RuRoBERTa model. The system got Accuracy metric value of 0.65 and took first place in the multi-class classification task.

Keywords: artificial text detection, transfer learning

DOI: 10.28995/2075-7182-2022-21-470-476

Распознавание сгенерированных русскоязычных текстов на основе моделей BERT

Посохов П. А.
Университет ИТМО,
Санкт-Петербург, Россия
paposokhov@itmo.ru

Скрыльников С. С.
Университет ИТМО,
Санкт-Петербург, Россия
skrylnikovs@itmo.ru

Махныткина О. В.
Университет ИТМО,
Санкт-Петербург, Россия
makhnytkina@itmo.ru

Аннотация

В данной статье описано наше решение для соревнования по распознаванию сгенерированных текстов RuATD (Russian Artificial Text Detection), проводящегося в рамках конференции Диалог 2021. Наш подход был основан на идее трансферного обучения, использовались предобученные модели RuRoBERTa, RuBERT, RuGPT3, RuGPT2. Итоговое решение включало токенизацию Byte-level Byte-Pair Encoding, и дообученную модель RuRoBERTa. Система получила значение метрики Accuracy 0,65 и заняла первое место в задаче мультиклассовой классификации соревнования.

Ключевые слова: распознавание сгенерированного текста; перенос обучения

1 Introduction

Artificial text detection systems are being developed for a long time now, first of those were based on the logical linguistic approach and were usually rule-based. The development of such systems was a time-consuming process, besides, the generated texts had the same type, because they used certain

patterns, the generated texts though were meaningful and syntactically correct [9]. Recently, some researchers still use the rules as components of automatic text generation systems [17]. Later development of text generation systems was based on statistical approaches such as Markov chains [14]. However, the result for such models can be unpredictable, semantic connections can be lost and sentences can be grammatically incorrect. The active development of neural networks gave automatic text generation a new life. The generation of meaningful texts that are grammatically correct and close to human-written texts became possible after the creation of neural network architectures based on transformers [16]. These models show impressive results, as Clark et al. [4] suggest the ability of non-specialists to distinguish between human and machine text (GPT2 and GPT3) in three areas (stories, news articles and recipes), and found out that without training evaluators can distinguish GPT3 generated text from human-written text purely by chance.

However, such models can also be used with different aims, for example, to create fake news [15,18], product and service reviews [1, 2]. For example, [18] shows that people rate model-generated disinformation as credible, even more than human-written disinformation. That is why the artificial text detection task is very relevant nowadays. Researchers have already made attempts to develop detection systems for artificial texts. The main approaches are: 1) training models from scratch, using the bag of words model and classical machine learning methods, such as logistic regression [6, 15]; 2) the use of pre-trained models based on transformers [1, 13, 18]. The second approach shows the best results in the artificial text detection. At the same time, it is worth noting, that research in this field was mainly carried out on datasets in English and Chinese languages. In recent years, the automatic text generation in Russian language has also reached a high quality, especially due to the emergence of pre-trained models ruGPT3, ruT5 [8], but the task of artificial text detection has not been given due attention. The article proposes a solution to this task and determines the model used to generate the text. The code is publicly available at <https://github.com/Anpopaicoconat/dialog2022>.

2 Task

The task set on RuATD (Russian Artificial Text Detection) is the multi-class classification of generated texts with generator model determination or assignment of Human class for cases when text is written by a person [11]. The list of response classes for this task contains the following:

- Human – text is written by a person;
- OPUS-MT – text is generated with machine translation model OPUS;
- ruGPT2-Large – text is generated with ruGPT2-Large model;
- ruGPT3-Large – text is generated with ruGPT3-Large model;
- ruGPT3-Medium – text is generated with ruGPT3-Medium model;
- ruGPT3-Small – text is generated with ruGPT3-Small model;
- M-BART – text is generated with Text2Text model M-BART;
- M-BART50 – text is generated with Text2Text model M-BART50;
- M2M-100 – text is generated with Text2Text model M2M-100;
- mT5-Large – text is generated with Text2Text model mT5-Large;
- mT5-Small – text is generated with Text2Text model mT5-Small;
- ruT5-Base – text is generated with Text2Text model ruT5-Base;
- ruT5-Base-Multitask – text is generated with Text2Text model ruT5-Base-Multitask;
- ruT5-Large – text is generated with Text2Text model ruT5-Large.

Initially, the task was to implement a multi-class classification. The input in this case is a text example, with an output being one of the 14 tags, containing the source of the text, being either title of the generation model or human.

Evaluation metric used for this task is accuracy, which is a standard metrics for classifier evaluation. It is the fraction of predictions the model got right.

3 Dataset

The dataset provided for the task contains 215,110 text «text»:«source», examples divided into training(129,066), test(64,533) and validation(21,511) sets. Training examples contain text, representing the statement from dialogue or chat (see Table 1).

Text	Class
Власти планируют закончить строительство аэропорта Сочи к 2018 году [Authorities plan to finish construction of Sochi airport by 2018].	ruGPT3-Large
Путин подписал указ об открытии музея Михаила Ивановича на Моховой улице в Москве[Putin signed a decree on the opening of the Mikhail Ivanovich Museum on Mokhovaya Street in Moscow].	mT5-Large
Мерелбеке — это муниципалитет, расположенный в бельгийской провинции Восточная Фландрия [Merelbeke is a municipality in the Belgian province of East Flanders]	Human
Вторая попытка привела к тому же результату [The second attempt had the same result].	OPUS-MT

Table 1. Examples from the training set are shown in the table

The data distribution by classes is shown in Figure 1. The classes in the data are unbalanced, however this split is explained with the ratio between model generated and human written texts. Therefore, 50% of the provided dataset contain examples of human written texts, the other half is accounted for text generation models, despite the dataset being unbalanced for 14 classes this split of samples is reasonable.

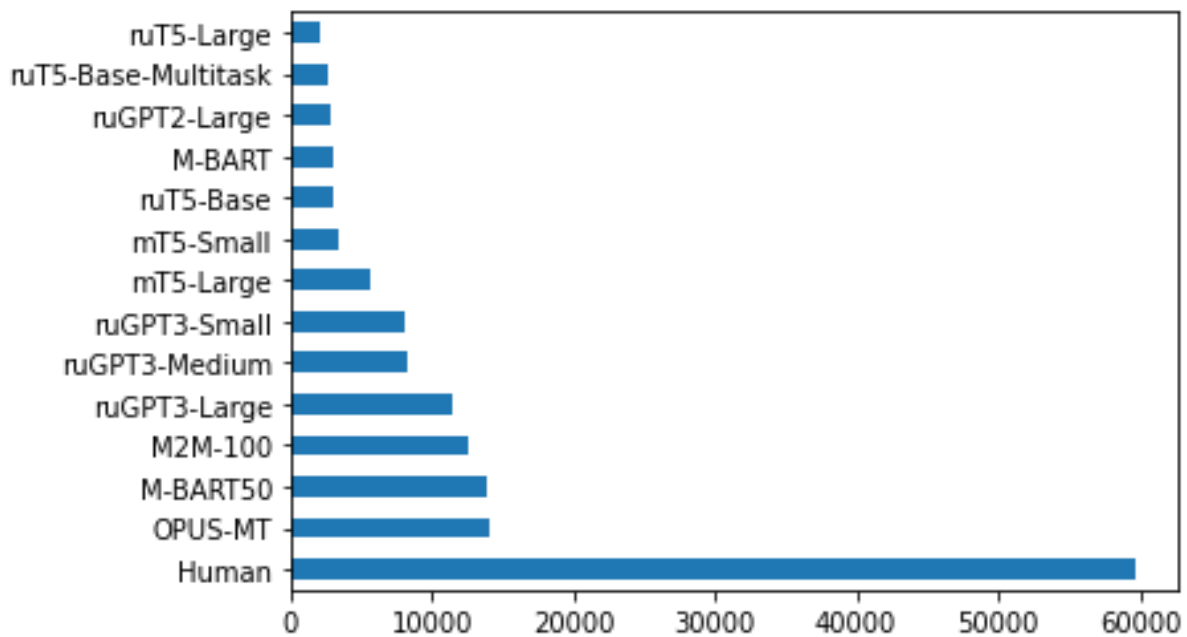


Figure 1. Data distribution by classes

4 Methods

Review of most recent works on this task allowed us to determine the most suitable models for its' solution in Russian language. The most commonly used technologies in natural language processing nowadays are transformer models. That is the reason why, the models chosen for solving this particular task also belong to this group of models. When implementing such models, usually the transfer learning

is used. This is an approach in machine learning, when network knowledges from one task is transferred to solve another, related task. Text processing by such models is based on the process of tokenization. Word tokens are available from the network dictionary, and they represent words or their parts if full word is missing. The tokens themselves are encoded with embeddings, which are their vector representations, that are processed by the network in parallel, but they also save the information about the location of words in the sentence. Initially, such networks are trained on large datasets, after which they are pre-tuned for a specific task, which makes these models quite flexible. For solution of this task the following models were considered:

1. GPT2 - generative pre-trained transformer model often used in natural language processing tasks. It is based on the use of attention mechanisms, which allow the model to segment the input data and selectively focus on the most relevant one. This model surpasses the previous ones based on recurrent or convolutional neural networks, as it parallelizes computations much better. [10]. For this particular task, the `rugpt2large` model was chosen, pre-trained to work with the Russian language and available on the HuggingFace hub¹, with standard configuration and the number of un-freeze layers equal to 8. This model was trained on 170 gigabytes of data, representing 1024 long sequences. Dictionary size is 50257. The number of neurons in the output layer is 14, according to classes given.

2. GPT3. Unlike its' predecessor, it has more than 100 times more parameters [3]. The exact chosen model was `rugpt3large_based_on_gpt2`. This is the Russian language model pre-trained by SberDevices². It has been trained on sequences of the same dimension. 80 billion tokens were used in training the first three epochs, after which the model was tuned to work with sequences of length 2048 for one more epoch. The output layer remained unchanged, however, the number of unfreeze layers was reduced to 4.

3. BERT (Bidirectional Encoder Representations from Transformers) this network was first introduced by Google and provided state-of-art results in many nlp tasks [5]. This model is primarily aimed at solving tasks that use the whole utterance such as sequence classification, token classification or question answering and requires fine-tuning for each specific use. For this task, the `rubert-base-cased` model was used, pre-trained by DeepPavlov³ with 180 million parameters and a dictionary size of 119547. This model is the closest one to the model provided in the baseline solution of the task that is why it was chosen as a metric reference point for other models. The model standard parameters were chosen the output layer had 14 neurons, according to the number of possible response classes.

4. RoBERTa. Is a transformer model pre-trained on a large corpus of the raw texts only, without any labels with the MLM (Masked Language Modeling) objective. This model trains on masked sentences which is rather different from is different from traditional recurrent neural network (RNN) approaches that usually see the words in set order, or from other transformer models like GPT, which internally mask the future tokens. Roberta's approach allows the model to learn a bidirectional representation of the sentence. For this task we used `RuRoBERTa-large` model with the following parameters: number of epochs equals 4, batch size equals 1, learning rate equals $2e-5$. The used optimizer was AdamW, based on it a linear scheduler with a warmup period was also used during which the learning rate increases linearly from zero to the initial one, and after that linearly decreases from the initial one set in the optimizer to 0. The warmup process is used in models with attention mechanisms to avoid the loss of weights the model learned during pre-training. Used model is an encoder, which was trained for the Russian language by the SberDevices team and available on the HuggingFace hub⁴. Its initial task is mask filling. In tokenization, Byte-level Byte-Pair Encoding is used [12]. This method allows model to have a smaller dictionary, with a larger number of options. In this case, the dictionary consists of 50,265 examples, the number of parameters is 335 million. When text is processed by the model, the first `cls` token is used for aggregation, which is followed by a dropout layer, with a probability of 0.1, this parameter is required to avoid model overfitting. The train data is fed to the input layer being fully connected with a dimension of 1024 neurons. Here the hyperbolic tangent (\tanh) activation function is used, followed by a dropout again. The output layer contains 14 neurons, corresponding to the number of classes in the task. The use of `RuRoBERTa-large`, pre-trained by sberbank-ai, for this task is justified by the fact that it shows higher accuracy metrics when working with Russian text data in comparison with other models.

¹ <https://huggingface.co/sberbank-ai/rugpt2large>

² https://huggingface.co/sberbank-ai/rugpt3small_based_on_gpt2

³ <https://huggingface.co/DeepPavlov/rubert-base-cased>

⁴ <https://huggingface.co/sberbank-ai/ruRoberta-large>

5 Experiments

The competition rules included providing the baseline solutions for the artificial text detection task. The first one is based on the use of the "bag of words" method with the tf-idf measure and logistic regression, the second is based on the application of the BERT model after fine-tuning, more specifically, it uses pre-trained DeepPavlov rubert-base-cased available on the HuggingFace hub. The approaches used in the study were also based on the concept of transfer learning, such pre-trained models as RuRoBERTa, RuBERT, RuGPT3, RuGPT2 were used. The Table 2 below shows the results of artificial text detection.

Model	Accuracy
RuRoBERTa	0.65035
RuBERT	0.59817
Baseline BERT	0.59813
RuGPT3	0.54574
RuGPT2	0.47258
Baseline tf-idf	0.44280

Table 2. Models results on artificial text detection

RuGPT2 did not prove to be better than the basic solution in this task, therefore the decision was made to change the model. The RuGPT3 model showed higher accuracy than its predecessor, however, still insufficient in terms of the model applicability for the task. The RuBERT model predictably turned out to be on par with the fine-tuned model provided in the baseline. Still BERT based model showed better results according to GPT ones, as it considers both contexts of the word, whilst GPT models are based on the use of left context only. The best results were obtained with use of the RuRoBERTa model, it resulted in accuracy equal to 0.65035, on the test data of the competition, thus taking first place in the multi-class classification. This model's architecture and hyperparameters are optimized for best efficiency and further modifications would lower the efficiency of the model, without proper pre-training. For comparison, the accuracy of the baseline of the BERT and tf-idf solutions is 0.59813 and 0.44280, respectively.

The figure 2 shows confusion matrix heatmap normalized by number of examples in every class. The model classifies human-generated texts best of all classes, being accurate at 89% of examples. The confusion in classification of artificially generated text messages was mostly noticed among those generated by models designed primarily for machine translation, them being OPUS-MT, M-BART50 and M2M-100. Another notable remark is that the model also confuses the messages generated by the same architecture of different sizes. For example, 28% of the messages generated by mT5-Small are recognized as generated by mT5-Large; the confusion between the ruGPT3-Large, ruGPT3-Medium, ruGPT3-Small models varies from 7% to 15%. Significant error in classification occurs due to recognition of artificial texts as human class. The largest percent of wrongly classified messages were generated by mT5-Large, ruT5-Large models. This can be explained by the quality of text generation for these models.

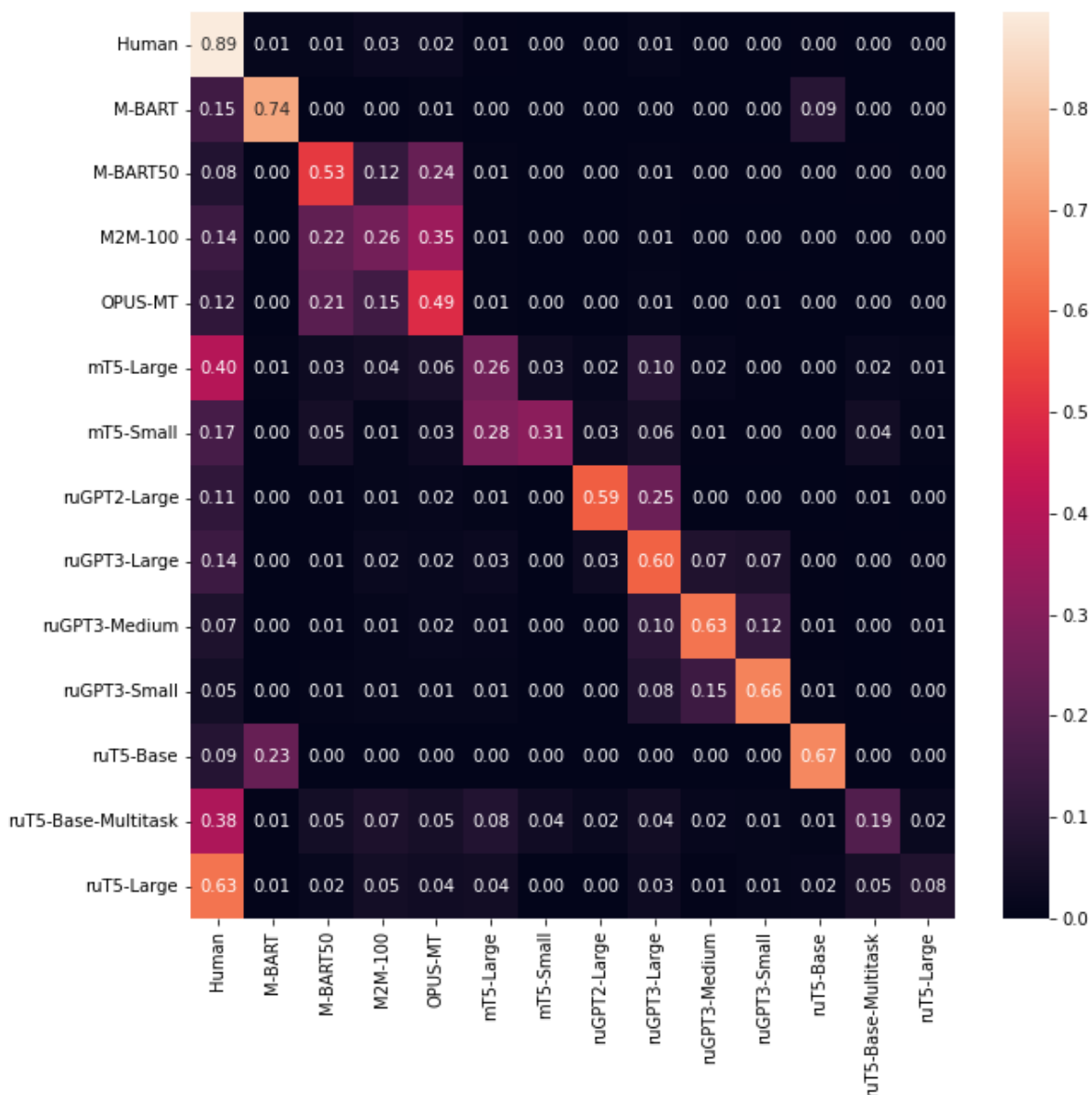


Figure 2. Heatmap of confusion matrix by classes

6 Conclusion

This paper presents the study of transfer learning usage for the artificial text detection task. Based on the analysis of modern research, it was concluded that for datasets in English and Chinese, this approach shows the best results this paper validates this statement also being true for Russian language datasets. A comparative analysis of the pre-trained Russian language models' usage showed the advantages of the RuRoBERTa model. The model presented in the article has several ways to be used, on the one hand, it can be used for its intended purpose to determine artificially generated messages, and to determine the exact model. On the other hand, it can be used to improve the quality of text generation, for example, as a filter for generated messages based on similarity between generated text and text written by a person.

Further development of the model, could possibly include training of separate classifiers corresponding to each applied model and use of output vectors built-in aggregation, thereby combining the models into an ensemble.

References

- [1] Adelani, D., Mai H., Fang F., Nguyen H., Yamagishi J., Echizen I. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection// The 34-th International Conference on Advanced Information Networking and Applications. — Caserta, Italy, 2020. — 1341–1354.
- [2] Bekmanova G., Sharipbay A., Omarbekova A., Yelibayeva G., Yergesh B. Adequate assessment of the customers actual reviews through comparing them with the fake ones // Proceedings of 2017 International Conference on Engineering and Technology — Antalya, Turkey, 2018. — P. 1-4
- [3] Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T.J., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei, D. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. — 2020 — Vol. 33. — P. 1877—1901.
- [4] Clark E., August T., Serrano S., Haduong N., Gururangan S., Smith N.A. All that's 'human' is not gold: Evaluating human evaluation of generated text. // 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference — 2021. — P. 7282-7296.
- [5] Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Minneapolis, MN, 2019. — Vol. arXiv:1810.04805. — version 2. Access mode: <https://arxiv.org/abs/1810.04805>.
- [6] Ippolito D., Duckworth D., Callison-Burch C., Eck D. Automatic Detection of Generated Text is Easiest when Humans are Fooled // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — P. 1808–1822.
- [7] Jawahar, G., Abdul-Mageed, M., Lakshmanan, L. Automatic detection of machine generated text: A critical survey// Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain, 2020. — P.2296–2309.
- [8] Matveev A., Makhnytkina O., Matveev Y., Svischev A., Korobova P., Rybin A., Akulov A. Virtual Dialogue Assistant for Remote Exams // Mathematics — 2021. — Vol. 9, No. 18, 2229. — Access mode: <https://doi.org/10.3390/math9182229>.
- [9] Mauldin M.L. Semantic rule based text generation//10th International Conference on Computational Linguistics, COLING 1984 and 22nd Annual Meeting of the Association for Computational Linguistics. — Stanford, CA , 1984. — P. 376–380.
- [10] Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language Models are Unsupervised Multitask Learners // OpenAI Blog — 2019. — Access mode: https://d4mucfpksyww.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [11] Shamardina T., Mikhailov V., Chernianskii D., Fenogenova A., Saidov M., Valeeva A., Shavrina T., Smurov I., Tutubalina E., Artemova E. Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian//Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue. – 2022. – Vol.21.
- [12] Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. — Berlin, Germany, 2016. — Vol. 1. — P. 1715–1725.
- [13] Solaiman I., Brundage M., Clark J., Askell A., Herbert-Voss A., Wu A., Radford A., Krueger G., Kim J.W., Kreps S., McCain M., Newhouse A., Blazakis J., McGuffie K., Wang J. Release Strategies and the Social Impacts of Language Models — 2019. — arXiv:1908.09203. — version 2. Access mode: <https://arxiv.org/abs/1908.09203>
- [14] Szymanski G., Ciota Z. On-line text generation using Markov Models// Proceedings of the International Conference Modern Problems of Radio Engineering, Telecommunications and Computer Science. — Lviv-Slavsko, Ukraine, 2004. — P. 339–341.
- [15] Uchendu, A., Le, T., Shu, K., Lee D. Authorship attribution for neural text generation//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2020. — P. 8384–8395.
- [16] Vaswani A., Shazeer N.M., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is All you Need // Advances in Neural Information Processing Systems. — Long Beach, CA, 2017. — P. 5999–6009.
- [17] Vodolazova T., Lloret E. The impact of rule-based text generation on the quality of abstractive summaries // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — Varna, Bulgaria, 2019. — P. 1275–1284.
- [18] Zellers R., Holtzman A., Rashkin H., Yonatan Bisk Y., Farhadi A., Roesner F., Choi Y. Defending Against Neural Fake News// Proceedings of the 33rd International Conference on Neural Information Processing Systems. — Vancouver, Canada, 2019. — P. 9054–9065.

Spoken corpora of spontaneous speech as a source to study polar question intonation in Russian dialects

Margje Post

Department of Foreign Languages

University of Bergen

Norway

Margje.Post@uib.no

Abstract

The emergence of several online spoken corpora of Russian regional speech opens new possibilities for the study of regional Russian intonation. The Russian dialect corpora of the Linguistic Convergence Laboratory [32; 1–10] were used to study the intonation of polar (yes/no) questions in regional rural speech. Although using spontaneous speech to study intonation is a challenge, the corpora are large enough to show general tendencies. The typical rising-falling pitch accent of most polar questions in Central Standard Russian is predominant in the regional corpora as well, but with possible variation in phonetic implementation and in the association of the fall. This accent is the most common even in the majority of question utterances with lowered questionhood, and dominates even in the regions known for rising accents in questions. The corpora show that tag questions are frequent in these interview data, unlike the question particles *li*, *ti* and *či*. Not only the dialectal particles *ti* and *či*, but also the Standard Russian question particle *li* shows a varying regional distribution.

Keywords: intonation; spoken corpora; polar questions; dialects; regional variation; Russian

DOI: 10.28995/2075-7182-2022-21-477-487

Устные корпуса спонтанной речи как источник для изучения интонации общих вопросов в русских говорах

Мархье Пост

Институт иностранных языков

Университет г. Бергена

Норвегия

Margje.Post@uib.no

Аннотация

Появление онлайн-корпусов спонтанной диалектной русской речи открыло новые возможности для изучения русской региональной интонации. Для изучения интонации общих вопросов в региональной сельской речи использовались русские диалектные корпуса Лаборатории лингвистической конвергенции [32; 1–10]. Хотя использование спонтанной речи в исследованиях интонации представляет собой сложную задачу, корпуса достаточно объемны, чтобы показать общие тенденции в вопросительной интонации. Типичное восходяще-нисходящее тональное оформление общих вопросов литературной произносительной нормы преобладает и в диалектных корпусах, но с возможным варьированием в фонетической реализации акцента и ассоциации тонального падения. Оно преобладает и в большинстве вопросительных высказываний с пониженной степенью вопросительности, и даже в тех регионах, которые известны восходящим акцентом в общих вопросах. Корпуса показывают, что в этих данных, взятых из интервью, часто встречаются вопросы-теги, в отличие от вопросительных частиц *ли*, *ти* и *чи*. Не только диалектные, югозападные частицы *ти* и *чи*, но и вопросительная частица литературного языка *ли* имеет неодинаковую распространенность по регионам.

Ключевые слова: интонация; устный корпус; общие вопросы; диалектология; региональное варьирование; русский язык

1 Introduction

Little empirical research has been done on the distribution and form of pitch accents in regional varieties of Russian. The emergence of several online spoken corpora of Russian regional speech opens new possibilities for the study of regional Russian intonation. In the study presented here, the Russian dialect corpora available on the website of the Linguistic Convergence Laboratory (LCL) [32] were used to study the intonation of polar (yes/no) questions in regional rural speech.

Almost all empirical intonation studies are based on laboratory speech. Spontaneous speech has the advantage of representing ecologically valid speech data [42], but it is notoriously difficult to study its intonation, since the pitch contour is influenced by a large number of uncontrolled parameters, not only by purely linguistic variables (e.g. sentence type, position of the accents), but also by numerous pragmatic, paralinguistic and sociolinguistic factors. The utterances from spontaneous speech represented in the corpora do therefore not allow direct comparisons of their intonation in all details, but even small corpora show general tendencies, giving evidence for the usage of certain tonal configurations. They also show which tunes are the most frequent in the dialects and even what is their usual alignment.

In Central Standard Russian (CSR), some polar (yes/no) questions are marked lexically with the question particle *li*, but most are marked as questions by intonation only. In the wide definition of questions that will be used here – as all utterances marked by a question mark – polar questions cover utterances with a large variety of pragmatic functions and with a varying degree of questionhood: far from all prompt the listener to give an explicit reply. They range from neutral information-seeking questions through echo questions expressing surprise to rhetorical questions. This variety of forms and functions enables many different tonal configurations for polar questions; cf. [20; 25; 26; 27; 28].

Nevertheless, most polar questions – 87% in the CoRuSS corpus of spontaneous speech [22]¹ – are formed with the same pitch accent: a high rise on the nuclear syllable, followed by a steep fall to low level in the next syllables [39; 46; 51]. After this nuclear pitch accent the pitch level remains low until the end of the utterance (cf. Fig. 1 for a typical example of a polar question in Russian). This rising-falling tonal configuration is known as (the neutral realisation of) the intonational construction IK-3 in Bryzgunova’s well-known model [16] or as H*L in the latest version of ToRI [39].² The accent is truncated to only a rise when the nuclear stress falls on the final syllable of the utterance and no postnuclear syllables are available [37; 47].

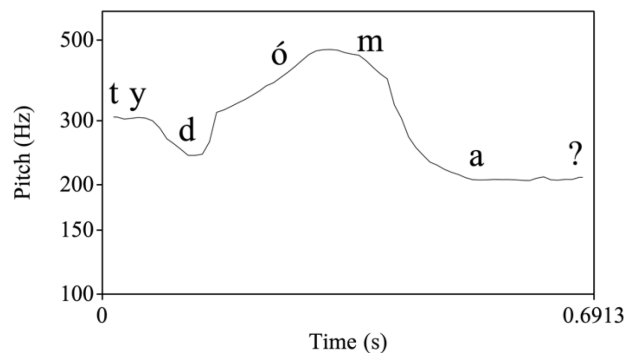


Figure 1. F₀ curve, made in Praat [14], of the Russian polar question *Ты дома?* “Are you at home?” (Khislavichi corpus, http://lingconlab.ru/khislavichi/OUT/2019_stajki_nnm1955_2_3-794300-797183.wav), with the typical rising-falling accent. The acute accent on the vowel indicates the nuclear stress

¹ I counted 539 out of 623 (87%) polar questions with a rising-falling pitch accent in the data in [22] from the speech corpus CoRuSS (Corpus of Russian Spontaneous Speech). The corpus itself is, to my knowledge, not publicly available. CoRuSS consists of connected communicative speech recorded from 60 Russian male and female speakers of different age groups in St. Petersburg [22]. The corpus is too small and too little balanced to give solid statistical information about modern CSR spontaneous speech, but the percentages give at least some cues as to the relative frequency of pitch accents.

² In an earlier version of ToRI [38], the pitch accent was called LH*L, which captures the obligatory rise preceding the high peak. Notice that this tune is called L*H L% in [46], whereas ToRI [39] uses the label L*H for an entirely different accent, which has a nuclear fall and a postnuclear rise (cf. Bryzgunova’s IK-4). To my knowledge, Rathcke did not study or label this falling-rising pitch accent. For example sound files, see the ToRI website [36].

An alternative to the rising-falling pitch accent is a low-rising, or falling-rising accent (Bryzgunova's IK-4; L*H in ToRI; 12% (n=73) in CoRuSS), which is typical for non-first elliptic questions starting with the conjunction *A* [39]. Certain expectations and speaker attitudes can lead to the choice of yet other pitch accents. Kodzasov [26: 163] notices some echo questions with falling intonation in cases with a low degree of questionhood, and a few polar questions have a rise, but no fall (less than 2% (n=11) in CoRuSS), the exact conditions for which remain to be described.³

The pitch contour can be modified by emotions, such as surprise, doubt, distrust and irony [50], affecting timing (alignment of the rise, the pitch peak and/or fall), duration of the stressed syllable and pitch levels (excursion size and/or pitch register); cf. [16; 26; 27; 28; 32; 50].

The scarce literature on intonation in regional varieties of Russian suggests that rising-falling pitch accents are predominant in polar questions elsewhere in Russia as well, but with possible variation in, among others, scaling (excursion size), timing of the pitch peak and association and alignment⁴ of the fall; cf. [43; 49; 51].

Alternatives to the rise-fall are attested. In Southwestern Russian dialects, polar questions can be formed with a nuclear rise continuing on the postnuclear syllables [21; 22], a pattern that is also observed in Polish [24: 144–149], in Ukrainian [24; 35] and in Odessa Russian [21]. Absence of a fall has also been recorded in the north [23].

Dialects bordering to Belarus and Ukraine use the question particle *ti* or *č'i*⁵ (DARJa III [12; 34], map 11, available online at <https://da.ruslang.ru/>). The standard Russian particle *li* is an enclitic, but *ti* and *č'i* can take the first position in the utterance. The dialect atlas [11; 12] gives no information about their intonation.

The main goal of the study was to find out whether the rise-fall predominates in the dialect corpora as well, and which alternatives they contain to express polar questions.

2 Data and methodology

2.1 The Russian dialect corpora of the Linguistic Convergence Laboratory

The 10 Russian dialect corpora that are currently⁶ available on the website of the Linguistic Convergence Laboratory (LCL) represent various regions of European Russia [32; 1–10; see Fig. 2].⁷ They contain biographical dialectological and folkloristic interviews that have been transcribed in Standard Russian, lemmatized and grammatically annotated. The Laboratory's spoken corpora also contain sociolinguistic metadata about the speakers, including information about their age, sex, education level and place of residency. The corpora can be queried via a web-based interface that provides the user with access to the original sound recordings on a per-utterance level [53]. The grammatical and sociolinguistic annotation of the corpus data enables variationist studies of linguistic features; e.g. [18].

Most speakers are typical participants in Russian dialectological fieldwork: the best available representatives of the traditional local dialect (elderly, mainly female speakers with little education who have not lived outside the region), but men have been interviewed as well, and some corpora also include younger speakers with a higher education level, especially the Zvenigorod corpus, which is said to represent Standard Russian [3].

³ [16], [22], [27] [41] and [51] mention rising accents in various meanings, among them in self-reminders, in emotional questions expressing disbelief or perplexity and in a subtype of echo questions. Neither of these usages of rising accents appear to be frequent.

⁴ *Association* relates to which word or syllable the pitch movements are associated to on a structural level; *alignment* is the phonetic property of the relative timing of events in the F₀ contour and events in the segmental string [30].

⁵ Russian words are transliterated following [17].

⁶ In February 2022. In April, an 11th corpus was added.

⁷ The LCL contains not only dialect corpora, but also corpora of bilingual speakers and dictionaries of languages spoken in the Russian Federation. Other online Russian dialect corpora – [48; 13] – were not used in the current study. The former is much smaller (as of February 2022), the latter is currently less suited to do quick analyses of larger data sets.



Figure 2. Map of the 10 Russian dialect corpora of the Linguistic Convergence Laboratory⁸

The LCL corpora are not directly comparable to the CoRuSS corpus [22]. One can expect a different distribution of question types, due to the nature of the dialect corpus data, which stem from dialectological interviews. The local speakers answer questions rather than pose them themselves. A majority of the utterances ending in a question mark in the corpora are not pragmatically neutral, information-seeking questions to which an answer is expected, but utterances with a lowered degree of questionhood. Their main pragmatic function may be to express speaker attitudes (uncertainty, surprise etc.) or to engage the interlocutors in a narrative; cf. [15]. The interviewees often use confirmation-seeking echo questions – repeating an earlier question or other discourse from the interlocutor(s), but very few non-first elliptic questions with A.

2.2 Methodology

Interrogative utterances by dialect speakers were found by searching for question marks. The 10 corpora, which contain from 68324 (Zvenigorod corpus, [3]) up to 959782 (Ustja corpus, [1]) tokens,⁹ contain between 381 and 12333 tokens of question marks. Since in some corpora, question marks were used to mark unintelligible speech or uncertain transcriptions, the number of questions is lower, ranging between 377 to 5728 (see Table 1 in section 3.3).

More specific results were obtained by combined queries, e.g. for *li* + question mark. The query results – the utterance transcriptions and their annotations – were downloaded as csv- or tsv-files and further annotated in Excel for features relevant for their intonation. The intonation of relevant utterances was studied by a combination of auditory analysis and visual inspection and annotation of the pitch contours in *Praat* [14].

Queries for question words in one of the corpora, the Rogovodka corpus [10], suggest that approx. half of the utterances ending in a question mark are polar questions, the other half being wh-questions.¹⁰

⁸ I added the names of the corpora to the map published on the LCL website (<http://lingconlab.ru>), which was made by Timofej Muxin, using Leaflet, Map tiles by Stamen Design (CC BY 3.0) and Map data © OpenStreetMap contributors.

⁹ As of February 2022.

¹⁰ The exclusion of most question words left max. 393 out of 814 question utterances being polar questions in the Rogovodka corpus [10]. This includes those with a low degree of questionhood.

3 Results

3.1 The predominance of the rising-falling accent

The corpus data confirm that the rise-fall is predominant in polar questions in regional Russian as well, at least in the corpora studied.¹¹ It is found in utterances with different pragmatic functions, even in most questions with reduced questionhood, such as in echo questions and in rhetorical questions. The rise-fall is even predominant in tag questions, which are abundant (see section 3.2). This does by no means imply that other tonal configurations are not possible, since the interviews in the corpora probably reflect a limited variety of pragmatic functions, but we can conclude that they are infrequent.¹²

As expected, contours with a rise continuing after the nuclear syllable (Fig. 3) are attested in the southwestern corpora, but they appear to be rare.

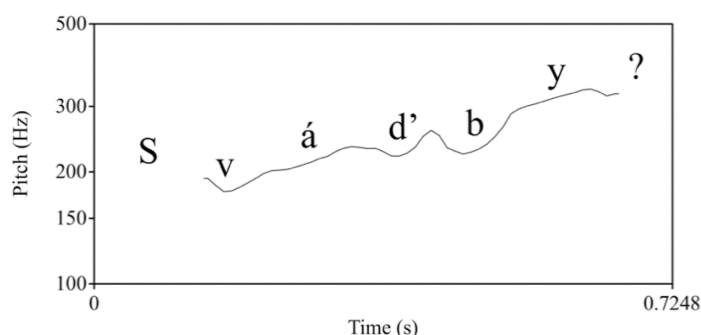


Figure 3. F₀ curve of the echo question *Свадьбы?* “Weddings?” (Khislavichi corpus, http://lingconlab.ru/khislavichi/OUT/2019_stajki_vnz1948_2-445849-446685.wav)

The form of the rise-fall shows regional variation, but mainly in details. The high turning point is typically aligned to the nuclear syllable, unlike in the speech of young Petersburg speakers, where alignment to the first postnuclear syllable has become common [52], and the fall tends to start shortly after, except in the Northern Russian dialect that is represented in the LCL corpora. Here, the fall in the rising-falling pattern is more often than not situated close to the end of the utterance, either on a late postnuclear word, or, in the absence of postnuclear words, on a late postnuclear syllable (Fig. 4; [43; 45; cf. 41; 49]). Further studies are needed to reveal the phonological association and status of these late falls.

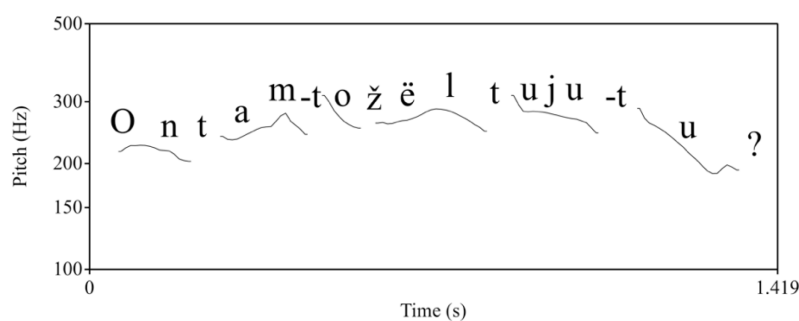


Figure 4. F₀ curve of the question *Он там-то жёлтую-ту?* “Is it the yellow one over there?” (Ustja corpus, <http://www.parasolcorpus.org/Pushkino/OUT/20140625b-gvp-1-943325-948162.wav>) with a typical Northern Russian late fall, here on the final syllable *-tu*, long after the nuclear syllable *žël-* [ʒɐl-] of *žěltuju* “yellow”

¹¹ This conclusion was drawn after listening through a large number of utterances from each corpus. I did not study the considerable amount of single syllable-questions (“*A?*”; “*Da?*” etc.) and other utterances with the nuclear accent on the final syllable, because they cannot show whether the pitch accent is rising-falling or only rising, due to the truncation of accents in (Standard) Russian. The absence or presence of truncation in regional varieties of Russian is a topic for future studies.

¹² No accurate numbers are given, since I did not consider it meaningful to actually count occurrences of pitch accents. The corpora data are far from representative for speech in general and they probably reflect only a subset of pragmatic contexts.

3.2 Tag questions with “*da?*”

Many utterances in the corpora contain question tags. The tags consist of a particle, such as *da?* “yes?” (see Table 1), or of another, often idiomatic, expression, such as *čto li* “isn’t it” (see Table 2), *predstavljajes’?* “do you imagine?” or *ponjatno?* “understood?”. Some speakers use the same tag many times.¹³

These tag questions are mainly, though not exclusively, used in utterances with a low degree of questionhood. Nevertheless, they usually contain a rise-fall.

The tag *da* normally has a nuclear rise, added to intonational phrases (IPs) with a falling nuclear accent, but they are also added to IPs that already have a rising-falling accent. Most multi-syllable tags carry a rise-fall, and when not in final position, the rise on *da* is followed by a fall as well. The tag *čto li* typically has falling intonation, but added to an IP that itself has an ‘interrogative’ rising-falling accent.

Lapteva [31] claims that the tag *da*, with rising intonation standing out on the background of a falling pitch contour, is frequent in utterance-internal position in Central Standard Russian colloquial speech, but not in dialects.¹⁴ The corpus data indeed suggest that this tag particle is most common in Central Russia: although *da* was attested as question tag in all dialects, 6 of the 9 speakers in central Zvenigorod use it very often (79+4=83 out of a total of 377 tokens of question utterances, of which half are wh-questions; see Table 1).

	<i>corpus name</i> ¹⁵	<i>region</i>	<i>tokens</i> ¹⁶	<i>questions</i> ¹⁷	..., <i>da?</i> ¹⁸	..., <i>da</i> ,...? ¹⁹	<i>čto li</i> ...? ²⁰
1	Ustja	Arkhangelsk	959782	5728	217	21	101
2	Lukh and Teza	Vladimir	146350	877	14		27
3	Zvenigorod	Moscow	68324	377	79	4	3
4	Luzhnikovo	Tver’	68666	570	17	3	2
5	OPOCHETSKY	Pskov	68741	580	39	8	2
6	Khislavichi	Smolensk	296327	1694	67	4	7
7	Spiridonova Buda	Brjansk	70565	636	7	1	
9	Nekhochi	Kaluga	138943	777	27		7
10	Rogovatka	Belgorod	100048	815	30		14

Table 1. Number of questions in the LCL corpora and questions with the tags *da* (utterance-finally and utterance-internally) and *čto li*

However, Lapteva’s non-final use is rare, with only 4 out of 83 tokens of the tag “*da?*” in Zvenigorod (Table 2). Non-final “*da?*” is more frequent in the Western Russian OPOCHKA corpus, relative to the corpus’ size. The few examples of non-final *da* are found in five different corpora from North and Central Russia. It is not alien to the Northern Russian speakers of today (cf. the Ustja corpus), with 21 tokens by 10 different speakers, all born after 1945.²¹ One of them is (2):

¹³ To give an extreme example, one of the Rogovatka speakers uses the phrase “*ponjatno?*” 71 times!

¹⁴ Lapteva’s examples from the north suggest she had mainly Northern Russian dialects in mind.

¹⁵ The Malinino corpus was temporarily not available when this research was carried out.

¹⁶ The column *tokens* gives the total number of tokens in each corpus (informants only). Note that the Ustja corpus is much larger than the other corpora.

¹⁷ Utterances ending in a question mark, including all wh-questions.

¹⁸ The tag “*da?*” in utterance-final position. The numbers include a few cases where the tag “*da?*” is added to a wh-question.

¹⁹ The question tag “*da?*” in non-final position.

²⁰ The tag *čto li*...? in utterances ending in a question mark; their total number is higher.

²¹ Although 217+21 tokens of “*da?*” is a much higher number than in the other corpora, it is not high relative to the large size of the Ustja corpus.

- (2) В общем они соседи, да, Индия с Пакистаном?
 V obščem oni sosedi, da, Indija s Pakistanom?
 “Anyway, they are neighbours, aren't they, India and Pakistan?” (Ustja,
<http://www.parasolcorpus.org/Pushkino/OUT/20140703b-mgb-2-1172152-1178079.wav>)

The intonation of utterances with non-final “*da*?” that caught Lapteva’s attention is worth mentioning. In the Central-Russian tokens, the pitch level drops to low level immediately after the rise on *da*, but in several Northern Russian examples, including (2) above, there is no pitch reset to low level after the tag: the fall on *Indija s Pakistanom* is delayed, like in most other questions in the north. However, the preceding accent on *sosedi* “neighbours”, the nuclear word preceding the tag *da* itself, does carry a rising-falling pitch accent with a steep fall. This tonal configuration obviously asks for further research.

3.3 The question particles *li*, *ti* and *či*

In search for alternative constructions, I performed queries for the interrogative particles *li*, *ti* and *či* used before a question mark. The particle *li* is part of standard Russian, the other two are recorded in the border area with Belarus (mainly *ti*) and Ukraine (mainly *či*; see [12; 34], map 11). The dialects of Khislavichi, Spiridonova Buda and Rogovatka are inside this area; the OPOCHKA dialects are situated just north of it.

This distribution is corroborated in the corpora data (Table 2). Even Standard Russian *li* shows varying geographical distribution: It is almost exclusively used in the north and east.

	<i>corpus name</i>	<i>region</i>	<i>questions (total)</i>	<i>ti...?</i>	<i>či...?</i>	<i>li...?</i>
1	Ustja	Arkhangelsk	5728			59 (87)
2	Lukh and Teza	Vladimir	877		0 (1)	13
3	Zvenigorod	Moscow	377			
4	Luzhnikovo	Tver'	570			1
5	OPOCHKESKY	Pskov	580			2
6	Khislavichi	Smolensk	1694	18 (21)	5 (9)	
7	Spiridonova Buda	Brjansk	636	3	1 (2)	
8	Nekhochi	Kaluga	777			1
9	Malinino	Lipetsk	1202			
10	Rogovatka	Belgorod	815		4	1

Table 2. Number of questions with question particles in the LCL Russian dialect corpora²²

Table 2 shows that in the corpora, all three question particles are rare, though one should consider that some tokens of the inconspicuous, non-standard particles *ti* and *či* might have been overlooked. The particles are much more frequent in other functions.²³ They are also used in declaratives, often in the meaning ‘or’, even in some of the questions (cf. the numbers in brackets in Table 2), and many tokens of *li* are usages in the idiomatic expression *čto li*.²⁴ This tag is far more frequent in questions than *li* in all corpora (cf. Table 1, last column).

The questions with *ti* and *či* tend to have falling pitch accents, which otherwise are rare, but their number is too low to draw conclusions about a possible link between particle use and intonation.

²² The higher numbers in brackets include questions with *ti*, *či* or *li* in the meaning ‘or’. Empty fields indicate zero tokens. The column *li...?* gives only tokens other than *čto li*.

²³ The Khislavichi corpus has 383 tokens of *ti* and 111 tokens of *či*; in Spiridonova Buda, they are used (at least) 33 resp. 45 times.

²⁴ The Ustja corpus has 1260 tokens of *li*, of which 260 are occurrences of *čto li*.

3.4 Linguistic convergence?

Alternative tonal configurations to the rise-fall, and questions with question particles are rare in all of the corpora. The low frequency of the dialect question particles and of the dialectal rising accent might well be a consequence of linguistic convergence (dialect levelling and influence from Standard Russian on the dialects). Their distribution over speakers – most are used by the oldest speakers in the corpora – points partly in this direction.²⁵ Accommodation of the speakers to the interview situation might also play a role [44].

4 Conclusion

The present study shows that spoken dialect corpora can give valuable information on regional intonation. The polar questions represented in the Russian dialect corpora of LCL corroborate earlier observations that rising-falling pitch accents are predominant in polar questions not only in Central Standard Russian, but in Russian regional speech as well, at least in the regions of Russia that are represented in the LCL corpora. Similar to earlier studied northern dialects, but unlike the dialects further south, the fall is often situated very late in the Northern Russian dialect, close to the end of the utterance, irrespective of the presence or absence of postnuclear words.

Questions with a continuous rise, typical for southwestern dialects, are, at most, rare in the corpora, and so are the question particles *tī* and *čī* in the southwestern dialects that have them. Not only these dialectal particles, but also the Standard Russian question particle *li* shows a varying regional distribution. It is almost exclusively found in the two dialect corpora in the north and east. It is far more frequent, however, in the idiomatic expression *čto li*.

In these biographical interviews, most question utterances have a lowered degree of questionhood, and tag questions are frequent. Tag questions with the particle *da* are most frequent in the corpus with the language that is closest to CSR Russian.

These data call for more research. More detailed studies of the intonation contours and their prosodic and pragmatic conditions will be able to reveal, among others, regional differences in the phonetic implementation of the rising-falling accent that appear to be present (e.g., in timing and slope of the fall; cf. earlier findings on regional differences in rising-falling accents in continuation contexts in [19].)

Although the online queries in the LCL dialect corpora only perform searches in the speech of the local speakers, the databases are also a rich source for the study of question intonation in Standard Russian, since in interviews, most questions are uttered by dialectologists and their students.

Acknowledgements

Funding for this research was provided by a grant from Meltzerfondet.

²⁵ In Khislavichi, *čī* is exclusively used by speakers born before WW II; *tī*, however, is also attested among younger speakers. Unfortunately, the corpus has no speakers born after 1961.

References

Russian dialect corpora of the Linguistic Convergence Laboratory

- [1] Ustja = Daniel M., Nina Dobrushina N., von Waldenfels R. The language of the Ustja river basin. A corpus of North Russian dialectal speech. — Bern, Moscow, 2013–2018. Electronic resource. URL: www.parasolcorpus.org/Pushkino (Last visited on 16.03.2022)
- [2] Lukh and Teza = Kuvshinskaya Yu. Corpus of the Russian dialect spoken in the basins of Lukh and Teza rivers. — Moscow: Linguistic Convergence Laboratory, HSE University, 2020. Electronic resource. URL: <http://lingconlab.ru/lukhteza/> (Last visited on 16.03.2022)
- [3] Zvenigorod = Panova, A. Corpus of Russian spoken in Zvenigorod. — Moscow: Linguistic Convergence Laboratory, HSE University, 2021. Electronic resource. URL: <http://lingconlab.ru/zvenigorod> (Last visited on 16.03.2022)
- [4] Luzhnikovo = Malysheva, A., Ter-Avanesova A. Luzhnikovo Corpus. — Moscow: Linguistic Convergence Laboratory, HSE University; Vinogradov Russian Language Institute of the Russian Academy of Sciences, 2020. Electronic resource. URL: <http://lingconlab.ru/luzhnikovo/> (Last visited on 16.03.2022)
- [5] Opochetsky = Ronko R., Volf E., Grebyonkina M., Ershova M., Okhapkina A., Khadasevich A., Valeria V. Corpus of Opochetsky dialects. — Moscow: Linguistic Convergence Laboratory, HSE University; Vinogradov Russian Language Institute of the Russian Academy of Sciences, 2019. Electronic resource. URL: <https://lingconlab.ru/opochka> (Last visited on 16.03.2022)
- [6] Khislavichi = Ryko, A., Spiricheva, M. Corpus of the Russian dialect spoken in Khislavichi district. — Moscow: Linguistic Convergence Laboratory, HSE University, 2020. Electronic resource. URL: <http://lingconlab.ru/khislavichi/> (Last visited on 15.03.2022)
- [7] Spiridonova Buda = Garder M., Petrova N., Moroz A., Panova A., Dobrushina N. Corpus of the Spiridonova Buda dialect. — Moscow: Linguistic Convergence Laboratory, HSE University, 2018. Electronic resource. URL: <http://linghub.ru/spiridonovabuda/> (Last visited on 17.03.2022)
- [8] Malinino = Ter-Avanesova A., Balabin F., Dyachenko S., Malysheva A., Panova A., Morozova V. Corpus of the Malinino dialect. — Moscow: Linguistic Convergence Laboratory, NRU HSE; Vinogradov Russian Language Institute of the Russian Academy of Sciences, 2019. Electronic resource. URL: <https://lingconlab.ru/malinino/> (Last visited on 17.03.2022)
- [9] Nekhochi = Ter-Avanesova A., Dyachenko S., Korpechkova E., Malysheva A., Pekunova I., Tolstaya M. Corpus of the Nekhochi dialect. — Moscow: Linguistic Convergence Laboratory HSE University, Vinogradov Russian Language Institute of the Russian Academy of Sciences, Institute of Slavic Studies of the Russian Academy of Science, 2020. Electronic resource. URL: <http://lingconlab.ru/nekhochi/> (Last visited on 15.03.2022)
- [10] Rogovatka = Ter-Avanesova A., Dyachenko S., Kolesnikova E., Malysheva A., Ignatenko D., Panova A., Dobrushina N. Corpus of Rogovatka dialect. — Moscow: Linguistic Convergence Laboratory, NRU HSE, 2018. Electronic resource. URL: <http://lingconlab.ru/rogovatka/> (Last visited on 17.03.2022)

Other references

- [11] Avanesov R., Bromlej S. (eds.): Dialektologičeskij atlas russkogo jazyka (DARJa): Centr Evropeiskoj časti Rossii, Vyp. 3, čast' 1. Sintaksis. Leksika. Spravočnye materialy. — Moscow: Vinogradov Russian Language Institute of the Russian Academy of Sciences, 1996.
- [12] Avanesov R., Bromlej S. (eds.): Dialektologičeskij atlas russkogo jazyka (DARJa): Centr Evropeiskoj časti Rossii, Karty. Vypusk 3, čast' 2. Sintaksis. Leksika. [Dialectological atlas of the Russian language: Centre of European Russia, Vol. 3, part 2. Syntax. Vocabulary] — Moscow: Vinogradov Russian Language Institute of the Russian Academy of Sciences, 2004.
- [13] Bjørnflaten, J. I. et al. The Oslo Corpus of Pskov dialects. — Tekstlaboratoriet, Oslo. Electronic resource. URL: <http://tekstlab.uio.no/pskov/> (Last visited on 14.01.2022)
- [14] Boersma P., Weenink D. Praat: Doing phonetics by computer. — Computer programme, 1992–2022. URL: www.praat.org
- [15] Bogdanova-Beglarjan N. Voprositel'nye konstrukcii v ustnoj spontannoju reči: adresnost', ritoričnost', idiomatičnost' [Interrogative constructions in spontaneous speech: addresses, rhetorics, idiomatics] // Kommunikativnye issledovanija, 4 (10), 2016. — pp. 61–76.
- [16] Bryzgunova E. Intonacija [Intonation] // Russkaja grammatika, Vol. I. — Moscow: Nauka, 1980. — pp. 96–122.
- [17] Comrie B., Corbett G. The Slavonic Languages. — London – New York: Routledge, 1993.

- [18] Daniel M., von Waldenfels R., Ter-Avanesova A., Kazakova P., Schurov I., Gerasimenko E., Ignatenko D., Makhlina E., Tsfasman M., Verhees S., Vinyar A., Zhigul'skaja V., Ovsjannikova M., Say S., Dobrushina N. Dialect loss in the Russian North: Modeling change across variables // *Language Variation and Change* 31 (3), 2019. — pp. 353–376.
- [19] Grammatčikova E., Knjazev S., Luk'janova L., Požarickaja S. Ritmičeskaja struktura slova i mesto realizacii tonal'nogo akcenta v regional'nyx variantax sovremennogo ruskogo literaturnogo jazyka // *Aktual'nye voprosy teoretičeskoj i prikladnoj fonetiki*. — Moscow: Buki-Vedi, 2013. — pp. 69–90.
- [20] Janko, T. Prosodija voprosov s časticej *li* [The Russian 'li'-questions prosody] // *Komp'juternaja lingvistika i intellektual'nye tehnologii*, 18: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog» 2019. — Moscow: RGGU. — pp. 754–764.
- [21] Janko T. O prosodičeskoj variativnosti // *Trudy Instituta ruskogo jazyka im. V. V. Vinogradova XIII: Kul'tura ruskogo reči* [Proceedings of the Vinogradov Russian Language Institute, No. 13: Russian Linguistic Standards]. — Moscow: Vinogradov Russian Language Institute of the Russian Academy of Sciences, 2017. — pp. 205–214.
- [22] Kachkovskaja T., Kocharov D., Skrelin P., Volskaja N. CoRuSS – a New Prosodically Annotated Corpus of Russian Spontaneous Speech // *Proceedings of LREC 10, 2016*. — pp. 1949–1954.
- [23] Kasatkin, L. 1989. *Russkaja dialektologija* [Russian dialectology]. — Moscow: Nauka.
- [24] Kasatkina R. Zametki o južnorusskoj intonacii [Remarks on Southern Russian intonation] // *Materialy i issledovanija po slavjanskoj dialektologii I (VII)*. — Moscow, 2002. — pp. 134–150.
- [25] Kobozeva I. Opyt razrabotki priznakovoj bazy dlja xarakteristiki leksiko-sintaksičeskich, semantičeskich i pragmatičeskich svojstv voprositel'nyx replik // *Komp'juternaja lingvistika i intellektual'nye tehnologii*, 4: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog» 2005.
- [26] Kodzasov S. Issledovanie modal'nyx intonacij [Studies on modal intonation] // *Prosodičeskij stroj ruskogo reči*. — Moscow: Vinogradov Russian Language Institute of the Russian Academy of Sciences, 1996. — pp. 145–180.
- [27] Kodzasov S. Intonacija voprositel'nyx predloženij: forma i funkcii [The intonation of interrogative sentences: form and function] // *Issledovanija v oblasti ruskogo prosodii*. — Moscow: Jazyki slavjanskix kul'tur, 2009 (1985). — pp. 175–189.
- [28] Kodzasov S., Bonč-Osmolovskaja A., Zaxarov L., Kobozeva I., Krivnova O. Baza dannyx «Intonacija ruskogo dialoga»: Voprositel'nye repliki [The database “Russian intonation: Interrogative utterances”] // *Komp'juternaja lingvistika i intellektual'nye tehnologii*, 4: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog» 2005.
- [29] Krause M. Epistemische Modalität: zur Interaktion lexikalischer und prosodischer Marker: dargestellt am Beispiel des Russischen und des Bosnisch-Kroatisch-Serbischen. — Harrassowitz, 2007.
- [30] Ladd D. R. *Intonational phonology*. — Cambridge: Cambridge UP, 1996.
- [31] Lapteva O. *Russkij razgovornyj sintaksis* [Russian colloquial syntax]. — Moscow: Nauka, 1976.
- [32] Linguistic Convergence Laboratory Resources. — Moscow: HSE University, 2017–2022. Website. URL: <http://lingconlab.ru> (Last visited on 05.05.2022)
- [33] Makarova V. Acoustic cues of surprise in Russian questions // *Acoustical Science and Technology*, 21 (5), 2000. — pp. 243–250.
- [34] Marčenko I., Dolgov O., Azanova A., Zambržitskaja M., Zalivina E., S. A. Zemljanskaja S., Močul'skij D., Tsejtina E., Čistjakova D., Ron'ko R. Baza dannyx dialektologičeskogo atlasa ruskogo jazyka [Database of the Dialectal atlas of Russian]. Available online at URL: <https://da.ruslang.ru/> (Last visited on 03.05.2022)
- [35] Nikolaeva T. *Frazovaja intonacija slavjanskix jazykov* [Utterance intonation in the slavic languages]. — Moscow: Nauka, 1977.
- [36] Odé C. ToRI, Transcription of Russian Intonation: a free interactive research tool and learning module. — University of Amsterdam, ACLC/Chair Phonetic Sciences. Electronic resource. URL: www.fon.hum.uva.nl/tori/ (Last visited on 18.03.2022)
- [37] Odé C. Neutralization or truncation? The perception of two Russian pitch accents on utterance-final syllables // *Speech Communication* 47 (1–2), 2005 — pp. 71–79.
- [38] Odé C. Communicative functions and prosodic labelling of three Russian pitch accents // *Studies in Slavic and General Linguistics* 32. — Amsterdam/New York: Rodopi, 2008. — pp. 377–401.
- [39] Odé C. Transcription of Russian intonation, ToRI, an interactive research tool and learning module on the internet // *Studies in Slavic and General Linguistics* 34. — Amsterdam/New York: Rodopi, 2008. — pp. 431–449.
- [40] Paufošima R. *Fonetika slova i frazy v severnorusskix govorax* [Utterance and word phonetics in Northern Russian dialects]. — Moscow: Nauka, 1983.
- [41] Paufošima R. Ob ispol'zovanii registrovyx različij v ruskogo frazovoj intonacii (na materiale ruskogo literaturnogo jazyka i severnorusskix govorov) [About the use of pitch register in Russian utterance intonation (based on Standard Russian and on Northern Russian dialects)] // *Slavjanskoe i balkanskoe jazykoznanie. Prosodija*. — Moscow: Nauka, 1989. — pp. 53–64.

- [42] Post B., Nolan F. Data Collection for Prosodic Analysis of Continuous Speech and Dialectal Variation // *The Oxford Handbook of Laboratory Phonology*. — Oxford: Oxford UP, 2012. — pp. 538–547.
- [43] Post M. Post-Nuclear Prominence Patterns in Northern Russian Question Intonation // *Proceedings of the Speech Prosody 2008 Conference* Editora RG/CNPq, 2008. — pp. 233–236.
- [44] Post M. Which dialect are we recording? A case study of a Russian babushka's linguistic accommodation // *Definitely perfect: Festschrift for Janneke Kalsbeek*. — Pegasus Oost-Europese Studies 29. — Amsterdam: Pegasus. — pp. 541–556.
- [45] Post M., Knjazev S. Ob intonacii obščego voprosa v dvux russkix govorax s tendenciej k poslovnomu tonal'nomu oformleniju [On the intonation of polar questions in two Northern Russian dialects with word-based tonal configurations] // *Aktual'nye problemy russkoj dialektologii. Materialy Meždunarodnoj konferencii 29–31 oktjabrja 2021 g.* Rossijskaja Akademija Nauk, 2021. — pp. 175–177.
- [46] Rathcke T. *Komparative Phonetik und Phonologie der Intonationssysteme des Deutschen und Russischen*. — München: Herbert Utz, 2008.
- [47] Rathcke T. On the neutralizing status of truncation in intonation: A perception study of boundary tones in German and Russian // *Journal of Phonetics*, 41 (3–4), 2013. — pp. 172–185.
- [48] Sappok, Ch.; Krasovitskij, A.; Paschen, L.; Brabender, K.; Koch, A.; Kühn, N. RuReg: Russische Regionen. Akustische Datenbank, 2016–2022. Electronic resource. URL: <http://www.rureg.de/> (Last visited on 14.01.2022)
- [49] Ščigel' E. Osobennosti intonacionnoj organizacii frazy v nekotoryx severnorusskix govorax [Peculiarities of the intonational structure of some Northern Russian dialects]. // *Dialektografija russkogo jazyka [Dialectography of Russian]* — Moscow: Nauka, 1985. — pp. 102–111.
- [50] Svetozarova N. Intonation in Russian. — *Intonation systems: A survey of twenty languages*. Cambridge: Cambridge U. P. — pp. 264–277.
- [51] Verbickaja L., Ignatkina L., Livačuk N., Sergeeva T., Cvetkova M., Ščukin V. Regional'nye osobennosti realizacii russkoj reči (na fonetičeskom urovne) [Regional peculiarities in the realisation of Russian speech (on the level of phonetics)] // *Vestnik Leningradskogo universiteta 8: Istorija. Jazyk. Literatura*, 1984 № 2. — pp. 71–80.
- [52] Volskaja N. Konflikt pokolenij v zerkale russkoj intonacii [Generational conflict in the mirror of Russian intonation] // *Aktual'nye voprosy teoretičeskoj i prikladnoj fonetiki: Sbornik stat'ej k jubileju O.F. Krivnojvoj*. — Moscow: Buki-Vedi, 2013. — pp. 53–62.
- [53] von Waldenfels R., Daniel M., Dobrushina N. Why standard orthography? Building the Ustya River Basin Corpus, an online corpus of a Russian dialect // *Komp'juternaja lingvistika i intellektual'nye tehnologii*, 13: Po materialam ežeodnoj Meždunarodnoj konferencii «Dialog» 2014. — Moscow: RGGU. — pp. 720–729.

Machine Reading Comprehension Model in RuNNE Competition

Igor Rozhkov

Lomonosov Moscow State University,
Moscow, Russia

fulstocky@gmail.com

Natalia Loukachevitch

Lomonosov Moscow State University,
Moscow, Russia

louk_nat@mail.ru

Abstract

The paper studies machine reading comprehension model (MRC) (Li et al., 2020) in its application to extracting nested named entities (nested NER) in the RuNNE-2022 evaluation (Artemova et al., 2022). The model transforms named entity recognition tasks to a question-answering task. In this paper we compare several approaches to formulating "questions" for the MRC model such as entity type names (keywords), entity type definitions, most frequent examples for the train set, combinations of definitions and examples. We found that using two most frequent examples from the training set is comparable in quality of nested NER with gathering qualitative definitions from different dictionaries, which is much more complicated. In the RuNNE evaluation, the MRC model obtained the best results among models without any manual work (rules or additional manual annotation of texts).

Keywords: Nested named entities, RuNNE evaluation, Machine reading comprehension

DOI: 10.28995/2075-7182-2022-21-488-496

Модель машинного понимания текстов (MRC) в тестировании RuNNE

Рожков И. С.

МГУ имени М.В. Ломоносова,
Москва, Россия

fulstocky@gmail.com

Лукашевич Н. В.

МГУ имени М.В. Ломоносова,
Москва, Россия

louk_nat@mail.ru

Аннотация

В статье исследуется модель машинного чтения (MRC) (Li et al., 2020) в ее применении для извлечения вложенных именованных сущностей в тестировании RuNNE-2022 (Artemova et al., 2022). Модель преобразует задачи распознавания именованных сущностей в задачи ответы на вопросы. В данной работе мы изучаем несколько подходов к формулированию «вопросов» для модели MRC. В тестировании RuNNE модель MRC показала лучшие результаты среди моделей, применяемых без какой-либо ручной работы (правил или дополнительной ручной аннотации текстов).

Ключевые слова: Вложенные именованные сущности, RuNNE, Модель машинного чтения

1 Introduction

Named entity recognition (NER) is one of the known task in natural language processing. Traditionally, NER task setting and datasets are devoted to extraction of so-called flat named entities, which presumes that a named entity cannot contain another named entity. For example, only one external ORGANIZATION entity should be extracted in *Lomonosov Moscow State University*, which leads to the loss of two internal named entity. During last years, due to the development of neural network models, the task of extracting nested named entities became much more frequent. Nested named entities allow for enhancing the coverage of found named entities, which is useful for such tasks as relation extraction, entity linking, knowledge graph population, etc. Specialized datasets are annotated with 2-6 levels of nestedness (Ringland et al., 2019; Plank et al., 2020; Loukachevitch et al., 2021). New NER methods specially devoted

to extracting nested named entities have been developed and significantly improved the performance in nested NER tasks (Shibuya and Hovy, 2020; Jue et al., 2020; Yu et al., 2020).

For Russian, two datasets annotated with nested named entity exist. The first dataset, FactRuEval (Starostin et al., 2016), is quite small for training machine learning models. Recently, new dataset NEREL (Loukachevitch et al., 2021) with nestedness up to 6 levels has been created. The NEREL dataset became a basis for organization of RuNNE-2022 evaluation (Artemova et al., 2022), devoted to recognition of nested named entities and also few-shot setting of nested NER.

In this paper we describe an approach applied to the RuNNE tasks, which is based on machine reading comprehension model (MRC) (Rajpurkar et al., 2016; Li et al., 2020). The model transforms NER tasks to question-answering tasks and achieve state-of-the art results on various NER datasets. We compare several approaches to formulating "questions" for the MRC model such as entity type names (keywords), entity type definitions, most frequent examples for the train set, combinations of definitions and examples. We found that using two most frequent examples from the training set is comparable in quality of nested NER with gathering of qualitative definitions from different dictionaries, which is much more complicated. In the RuNNE evaluation, the MRC model obtained the best results among models without any manual work (rules or additional manual annotation of texts).

2 Related Work

Early works regarding nested NER involved mainly hybrid methods that combined rules with supervised learning algorithms (Shen et al., 2003; Zhang et al., 2004). Another approach to the nested NER task relies on hand-crafted features (Alex et al., 2007; Muis and Lu, 2018). These methods mostly failed to take advantage of the dependencies among nested entities.

Later, LSTM-based models were developed to process nested named entities. LSTM-CRF model (Ju et al., 2018) was already able to capture context representation of input sequences and globally decode predicted labels for nested entities even of the same entity type. Dynamically stacked multiple layers recognize outer entities by taking full advantage of information encoded in their corresponding inner entities. Straková et al. (Straková et al., 2019) identify nested named entities by a seq2seq model exploring combinations of different context-based embeddings (ELMo, BERT, Flair). Sohrab and Miwa (Sohrab and Miwa, 2018) proposed to concatenate the LSTMs outputs for the start and end positions of spans and then calculate a score for each span. In Biaffine model, Yu et al. (Yu et al., 2020) demonstrated that the model provides a global view on the input and performs better results – the model scores pairs of start and end tokens to form a named entity. Pyramid model (Jue et al., 2020) consists of a stack of inter-connected layers. Each layer l predicts whether a l -gram is a complete entity mention. The Second-best Sequence Learning coupled with Decoding (Second Best) model (Shibuya and Hovy, 2020) uses the Conditional Random Field output layer. The model treats the tag sequence for nested entities as the second best path within the span of their parent entity. In addition, the decoding method for inference extracts entities iteratively from outermost ones to inner ones in an outside-to-inside way.

Machine Reading Comprehension (MRC) (Rajpurkar et al., 2016) treats the nested NER as a question-answering task (Li et al., 2020), when for each named entity type, a specialized question is created. The model should find answers to the questions in a sentence, which is equivalent to extracting corresponding named entities. In (Loukachevitch et al., 2021), several models (Biaffine, MRC, Pyramid) were studied for extracting nested named entities in Russian. The best results were obtained by the MRC model.

3 Machine Reading Comprehension Model

The MRC model treats the NER task as extracting answer spans to specialised questions, each entity type is associated with a specific question. The dataset sentences are converted into triples (Question Q , Answer A , Context C). Question Q is either generated or selected supplementary sequence (described below); the Answer A is the annotated named entity, the subsequence of the given sentence; the Context C is the given sentence. The MRC model is constructed over the BERT (Devlin et al., 2018) model, which obtains the following string as an input:

$$\{[CLS], q_1, q_2, \dots, q_m, [SEP], t_1, t_2, \dots, t_n\}$$

where q_i are words of the question sequence, t_i are words of the given sentence, $[CLS]$ and $[SEP]$ are special tokens of the BERT model. The MRC model should extract a continuous span A in the context C :

$$A = \{t_i, \dots, t_{i+k}, 1 \leq i \leq i+k \leq n\}$$

such that A is now a retrieved named entity.

The model backbone is as follows. BERT, given aforementioned input, outputs a context representation matrix $E \in \mathbb{R}^{n \times d}$, where d is the size of last layer dimension of BERT. The query part of output is dropped.

Next, given matrix E , model first predicts the probabilities of each word to be start index, to be end index and then probability of each start-end indices pair to be matched onto one named entity.

In more detail: model first predicts two values, P_{start} and P_{end} as follows:

$$P_{start} = \text{softmax}_{eachrow}(E \cdot T_{start}) \in \mathbb{R}^{n \times 2}$$

$$P_{end} = \text{softmax}_{eachrow}(E \cdot T_{end}) \in \mathbb{R}^{n \times 2}$$

where $T_{start}, T_{end} \in \mathbb{R}^{n \times 2}$ are the weights learned. Then for \hat{I}_{start} and \hat{I}_{end} sets

$$\hat{I}_{start} = \{i \mid \text{argmax}(P_{start}^{(i)}) = 1, i = \overline{1, n}\}$$

$$\hat{I}_{end} = \{j \mid \text{argmax}(P_{end}^{(j)}) = 1, j = \overline{1, n}\}$$

where superscripts (i) and (j) denote i -th and j -th row of a matrix respectively binary classification model is trained to predict value of matching probability:

$$P_{i_{start}, j_{end}} = \text{sigmoid}(M \cdot \text{concat}(E_{i_{start}}, E_{j_{end}}))$$

where $M \in \mathbb{R}^{1 \times 2d}$ is weights learned. Now this value predicts whether each occurred span i_{start}, j_{end} in the context C is a desired answer A , i.e. named entity of given type.

There are different approaches to creating questions. (Li et al., 2020) proposed several of them:

- **Position index:** question is generated based on the position index of given tag, i.e. "first", "second", etc. or "one", "two", etc.
- **Keyword usage:** question is given or generated keyword describing tag, e.g. "profession", "person".
- **Rule-based template filling:** generates a sequence from given template, e.g. "Find named entities of type "person" in the given sentence."
- **Wikipedia definition retrieval:** question is generated with Wikipedia definition of a given tag, e.g. "An organization is an entity comprising multiple people, such as an institution or an association."
- **Synonyms:** words that have the same or close meaning to the original tag, e.g. for tag "profession" that would be "occupation", "job", etc.
- **Concatenation of keyword and synonyms:** question is constructed from both keywords and synonyms, e.g. "profession, occupation, job".
- **Annotation guideline notes:** the guidelines of labeled entities provided by the dataset builder, e.g. for *location* it could be "Find locations in the text including nongeographical locations, mountain ranges and bodies of water".

The last approach achieves best results in the original work.

4 RuNNE task and data

RuNNE competition (Artemova et al., 2022) sets the few-shot version of the nested named entity recognition task. While most of the entities have considerable number of examples in the training set, several others occur much less frequently: the amount of such entities is limited in the training set. Dev and test

NE type	Number of mentions	
	train	test
PROFESSION	4566	848
PERSON	4517	961
ORGANIZATION	4049	675
EVENT	2850	683
COUNTRY	2521	456
DATE	2268	523
CITY	1101	239
NUMBER	1026	230
ORDINAL	565	107
AGE	554	138
NATIONALITY	394	66
LAW	389	61
FACILITY	371	63
STATE_OR_PROVINCE	343	112
AWARD	322	119
IDEOLOGY	300	43
LOCATION	270	62
PRODUCT	237	53
CRIME	180	35
MONEY	171	43
TIME	154	47
DISTRICT	98	25
RELIGION	94	24
PERCENT	82	7
LANGUAGE	43	8
DISEASE	32	57
PENALTY	32	17
WORK_OF_ART	30	88
FAMILY	17	14
Total amount	27576	5804

Table 1: Number of entities in the RuNNE training and test sets.

sets both contain the usual (non-limited) amount. Therefore, the main goal of the competition is to create models capable of retrieving both common and uncommon named entity types.

The dataset of RuNNE evaluation was created from the NEREL dataset (Loukachevitch et al., 2021). This data was collected from WikiNews texts in Russian language, manually labeled by the annotators of the NEREL dataset using the *brat* annotation tool (Stenetorp et al., 2012). After that the initial dataset was mixed and split into train, dev and test sets. The dataset contains 29 different named entity types, with maximum nestedness of 6 levels.

For the few-shot task formulation, three classes were chosen and decreased in the amount for the training set, namely *disease*, *work_of_art* and *penalty*. Table 1 shows the amount of each labeled entity type both in the training and test sets.

As a result, we can see that the classes are not balanced, and there are no more than 32 mentions of the aforementioned entity types in the training set. Moreover, other types have similar amount of the mentions, e.g. *language* has 43 mentions, while *family* has even less - 17.

For evaluation on the RuNNE dataset, the macro-average precision, recall and F1 both for only new (few-shot NER task) and all (general NER task) entity types are used.

5 Approach and Results

In this work we study what approaches to question generation help the MRC model in few-shot and general NER tasks ¹.

Though annotation guidelines allows achieving the best results in the original work, they do not always exist for some dataset. Sometimes it is quite difficult to retrieve or generate such. In our case, the RuNNE dataset was not provided with annotation guidelines, and thus we cannot employ this approach. Therefore, aside from previously described approaches, this work proposes few new techniques for question generation:

- **Definition selection.** The questions are definitions of entity types, carefully selected from multiple dictionaries.
- **N most frequent examples.** The N most frequent examples of entities are obtained from the given training set and questions are generated. Number N is pre-defined from the start.
- **N most frequent entity components.** Each entity example in the training set is split into single words and then lemmatized. After that the N most frequent words are retrieved, which then compose the question.
- **Concatenation of definitions and most frequent examples.**

Examples of dictionary definitions are as follows (translated from Russian): "**Age** is the period of time when someone was alive or something existed."; "A **city** is a place where many people live, with many houses, shops, businesses, etc."

Question examples of N most frequent examples (here we presume $N = 5$) are as follows (translated from Russian): "**Date** is an entity such as in Monday, in Tuesday, today, in year 2011, in year 2004." "**Law** is an entity such as Constitution, CC (Criminal Code), CC of RF, Yarovaya package, constitution."

Question examples of N most frequent entity components ($N = 5$) are as follows (translated from Russian): "**Disease** is an entity such as cancer, hurt, heart, heart as adjective, pain". "**City** is an entity such as moscow, moskovsky, london, new-york, kyiv."

In this work, we study the results of utilizing following aforementioned approaches to question generation:

- **Keyword usage**
- **Definition selection**
- **N most frequent examples, for $N = 2, 5, 10$.**
- **N most frequent entity components, for $N = 2, 5, 10$.**
- **Concatenation of definitions and most frequent examples**

As baseline we use following models:

- **RuBERT** (Kuratov and Arkhipov, 2019): Baseline model of the RuNNE competition.
- **2nd-best-path-RuBERT** (Shibuya and Hovy, 2020): treats the tag sequence as the second best path within in the span of their parent entity based on RuBERT.

We utilize RuBERT (Kuratov and Arkhipov, 2019) as a basis for MRC model. We use the batch size of 32 and learn the MRC model for 16 epochs on the 8 GPUs over the RuNNE data. We use AdamW optimizer (Loshchilov and Hutter, 2017), with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$. RuBERT configuration was set to default values after (Kuratov and Arkhipov, 2019). We use OneCycleLR learning rate scheduler (Smith and Topin, 2019) with maximum learning rate of $2 * 10^{-5}$, final div factor = 10^4 , linear anneal strategy. Weight decay was set to 0.01. Other hyperparameters were set to default values.

Table 2 shows experimental results on the RuNNE dataset. We can see that using two most frequent entity components from the training set is even slightly better than using well-constructed entity definitions. For few-shot setting, the results based on definitions are slightly better, but the extracting frequent entity components is much simpler than gathering well-written definitions from various dictionaries. The increase in the number of components leads to a decrease in quality of name extraction. Furthermore, we can see that using original entity examples and not split ones shows lower results. Though this approach is even simpler than previous one, it acts poorer. Also combinations of definitions and examples do not

¹https://github.com/fulstock/mrc_nested_ner_ru

Model and approach	General Task			Few-shot Task			
	Precision	Recall	F1	Precision	Recall	F1	
RuBERT-Tagger	-	-	67.44	-	-	44.66	
2nd-best-path-RuBERT	74.83	61.78	67.68	77.61	09.77	17.36	
MRC	Keyword	78.27	71.92	73.79	88.09	45.22	59.02
	Definitions	78.76	72.44	74.31	80.62	50.77	61.21
	2 most frequent examples	78.59	72.19	74.17	84.32	45.03	57.98
	5 most frequent examples	79.23	71.58	73.89	84.76	47.29	58.98
	10 most frequent examples	78.13	70.64	73.09	81.60	45.56	56.96
	2 most fr. entity components	78.65	73.05	74.63	86.15	49.07	60.80
	5 most fr. entity components	78.54	72.77	74.62	83.39	48.35	60.30
	10 most fr. entity components	78.04	71.82	73.76	83.62	47.82	59.52
	Def. + 2 most frequent ex.	78.37	71.74	73.96	80.21	49.89	60.83
	Def. + 5 most frequent ex.	77.83	72.62	74.26	78.47	48.71	58.69
	Def. + 10 most frequent ex.	77.60	71.36	73.22	82.50	45.68	57.24
RuNNE	pullenti	-	-	81.12	-	-	71.03
	MSU-RCC (ours)	-	-	74.93	-	-	60.39
	SibNN	-	-	74.25	-	-	40.37
	user:abrosimov_kirill	-	-	74.08	-	-	64.41

Table 2: Results (macro-averaged, %), compared with other models of the RuNNE competition.

task and the few-shot setting. The MRC model transforms named entity recognition tasks to question-answering tasks. We compared several approaches to formulating "questions" for the MRC model such as entity type names (keywords), entity type definitions, most frequent entity components and most frequent examples for the training set, combinations of definitions and examples. We found that using two most frequent entity components from the training set is even slightly better than using well constructed entity definitions. For few-shot setting, the results based on definitions are slightly better, but the extracting frequent entity components is much simpler than gathering well-written definitions from dictionaries.

In the RuNNE evaluation, the MRC model utilizing definitions as questions obtained the best results among machine learning models used without additional manual annotation of training texts.

Acknowledgments

The work is supported by the Russian Science Foundation, grant # 20-11-20166. The research is carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University.

References

- Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. // *Biological, translational, and clinical language processing*, P 65–72.
- Ekaterina Artemova, Maksim Zmeev, Natalia Loukachevitch, Igor Rozhkov, Tatiana Batura, Pavel Braslavski, Vladimir Ivanov, and Elena Tutubalina. 2022. RuNNE-2022 Shared Task: Recognizing Nested Named Entities. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog"*.
- 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.
- Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 1446–1459.

- Wang Jue, Lidan Shou, Ke Chen, and Gang Chen. 2020. Pyramid: A layered model for nested named entity recognition. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 5918–5928.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 5849–5859.
- Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.
- Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Iliia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. Nerel: A russian dataset with nested named entities, relations and events. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, P 876–885.
- Aldrian Obaja Muis and Wei Lu. 2018. Labeling gaps between words: Recognizing overlapping mentions with mention separators. *arXiv preprint arXiv:1810.09073*.
- Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. Dan+: Danish nested named entities and lexical normalization. // *Proceedings of the 28th International Conference on Computational Linguistics*, P 6649–6662.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.
- Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. Nne: A dataset for nested named entity recognition in english newswire. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 5176–5181.
- Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Effective adaptation of hidden markov model-based named entity recognizer for biomedical domain. // *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*, P 49–56.
- Takashi Shibuya and Eduard Hovy. 2020. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.
- Leslie N. Smith and Nicholay Topin. 2019. Super-convergence: very fast training of neural networks using large learning rates. // *Defense + Commercial Sensing*.
- Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. // *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, P 2843–2849.
- 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. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog”*, P 702–720.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. // *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics.
- Jana Straková, Milan Straka, and Jan Hajič. 2019. Neural architectures for nested ner through linearization. *arXiv preprint arXiv:1908.06926*.
- Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 6470–6476.
- Jie Zhang, Dan Shen, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2004. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of biomedical informatics*, 37(6):411–422.

NE type	10 most frequent entity components	Mentions in the test, %		
		2 ex.	5 ex.	10 ex.
AGE	год, 42-летний, 40-летний, 31-летний, 55-летний, 60-летний, 25-летний, 62-летний, 50, годовщина	20.81	22.84	27.41
AWARD	премия, нобелевский, мир, чемпион, медаль, чемпионка, золотой, за, год, олимпийский	11.82	23.92	30.55
CITY	москва, московский, лондон, нью-йорк, киев, Санкт-Петербург, римский, доха, бостон, столица	11.57	16.12	20.66
COUNTRY	россия, США, российский, РФ, украина, американский, израиль, Великобритания, федерация, япония	20.34	29.56	39.83
CRIME	убийство, коррупция, домогательство, преступление, сексуальный, нарушение, незаконный, связь, насилие, шпионаж	08.43	09.64	15.66
DATE	год, в, 2016, 2013, 2017, декабрь, октябрь, июнь, 2012, день	28.91	32.37	39.82
DISEASE	рак, ушибить, сердце, сердечный, боль, брюшной, полость, отравление, грудной, клетка	07.45	14.89	17.02
DISTRICT	район, округ, военный, северокавказский, московский, федеральный, химкинский, косовский, СКФО, сибирский	18.18	18.18	18.18
EVENT	выборы, отставка, назначить, погибнуть, чемпионат, родиться, пост, задержать, человек, арестовать	01.58	04.73	08.41
FACILITY	памятник, гора, храмовый, собор, улица, дом, здание, аэропорт, площадь, дворец	00.00	01.95	09.09
FAMILY	семья, королевский, романов, дом, Бекхэм, Abanyiginya, Хантсмен, монарший, Обама	20.00	20.00	20.00
IDEOLOGY	демократ, республиканец, демократический, оппозиционный, террорист, консерватор, коммунистический, социалистический, оппозиция, левый	06.12	08.16	30.61
LANGUAGE	английский, русский, арабский, французский, испанский, немецкий, итальянский, чувашский, марийский, татарский	25.00	37.50	50.00
LAW	закон, о, конституция, кодекс, УК, статья, ., РФ, «, федеральный	06.02	17.67	34.96
LOCATION	европа, европейский, западный, море, запад, северокавказский, берег, америка, иордан, южный	12.86	14.29	17.14
MONEY	доллар, млн, рубль, миллиард, миллион, млрд, США, тысяча, евро, ..	22.39	31.34	49.25
NATIONALITY	россиянин, гражданин, американец, российский, американский, серб, русский, канадец, США, афроамериканец	04.48	23.88	28.36
NUMBER	два, тысяча, четыре, один, около, три, двое, шесть, 1, трое	08.98	16.02	24.22
ORDINAL	первый, второй, третий, пятый, ii, четвертый, xvi, v, 1, 2	47.27	55.45	63.64
ORGANIZATION	правительство, россия, совет, полиция, парламент, партия, РФ, МВД, Госдума, комитет	03.13	06.19	09.25
PERCENT	%, процент, 1, 30, 90, 50, 20, 75, 49, 24	40.00	46.67	46.67
PERSON	Владимир, Путин, Сергей, Александр, Дмитрий, Обама, Медведев, Виктор, Кастро, Андрей	02.28	06.56	07.90
PENALTY	казнь, штраф, тюрьма, заключение, год, смертный, 20, въезд, пожизненный, денежный	17.65	29.41	29.41
PRODUCT	интернет, твиттер, facebook, сеть, википедия, youtube, союз, twitter, tumblr, як-18г	09.52	20.63	26.98
PROFESSION	президент, глава, министр, премьер-министр, губернатор, россия, директор, депутат, председатель, «	07.28	11.32	15.95
RELIGION	мусульманин, исламский, православный, мусульманский, ислам, католический, католик, христианин, баптистский, итский	25.00	58.33	58.33
STATE_OR_PROVINCE	область, край, каталония, тexas, чечня, Калининградский, Архангельский, Крым, Калифорния, Массачусетс	07.04	08.45	16.20
TIME	час, минута, год, время, вечером, ночь, в, утром, около, местный	08.21	18.66	30.60
WORK_OF_ART	рим, друг, список, Шиндлера, старый, спасатель, малиб, проповедь, падение, le	00.00	00.00	00.00

Table 3: Most frequent components of each entity type in the training set (in decreasing order), and their corresponding amount in the test set.

Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian

Tatiana Shamardina^{1*}, Vladislav Mikhailov^{2*}, Daniil Chernianskii^{3,4},

Alena Fenogenova², Marat Saidov⁶, Anastasiya Valeeva⁷,

Tatiana Shavrina^{2,3}, Ivan Smurov^{1,7}, Elena Tutubalina^{5,6,8}, Ekaterina Artemova⁶

¹ABBY, ²SberDevices, ³AIRI, ⁴Skolkovo Institute of Science and Technology, ⁵Sber AI,

⁶HSE University, ⁷Moscow Institute of Physics and Technology, ⁸Kazan Federal University

Correspondence: tatiana.shamardina@abby.com

Abstract

We present the shared task on artificial text detection in Russian, which is organized as a part of the Dialogue Evaluation initiative, held in 2022. The shared task dataset includes texts from 14 text generators, i.e., one human writer and 13 text generative models fine-tuned for one or more of the following generation tasks: machine translation, paraphrase generation, text summarization, text simplification. We also consider back-translation and zero-shot generation approaches. The human-written texts are collected from publicly available resources across multiple domains.

The shared task consists of two sub-tasks: (i) to determine if a given text is automatically generated or written by a human; (ii) to identify the author of a given text. The first task is framed as a binary classification problem. The second task is a multi-class classification problem. We provide count-based and BERT-based baselines, along with the human evaluation on the first sub-task. A total of 30 and 8 systems have been submitted to the binary and multi-class sub-tasks, correspondingly. Most teams outperform the baselines by a wide margin. We publicly release our codebase, human evaluation results, and other materials in our GitHub repository.

Keywords: artificial text detection, natural language generation, shared task, neural authorship attribution, transformers

DOI: 10.28995/2075-7182-2022-21-497-511

RuATD-2022: Соревнование по автоматическому распознаванию сгенерированных текстов

Татьяна Шамардина^{1*}, Владислав Михайлов^{2*}, Даниил Чернявский^{3,4},

Алена Феногенова², Марат Саидов⁶, Анастасия Валеева⁷,

Татьяна Шаврина^{2,3}, Иван Смуров^{1,7}, Елена Тутубалина^{5,6,8}, Екатерина Артемова⁶

¹ABBY, ²SberDevices, ³AIRI, ⁴Сколтех, ⁵Sber AI,

⁶НИУ ВШЭ, ⁷МФТИ, ⁸КФУ

Для связи: tatiana.shamardina@abby.com

Аннотация

Данная статья представляет собой отчет организаторов соревнования RuATD-2022, посвященного автоматическому распознаванию сгенерированных текстов на материале русского языка. Соревнование RuATD-2022 проходило в рамках кампании Dialogue Evaluation в 2022 году. Набор данных, использованный в соревновании, частично составлен автоматически с использованием моделей генерации текстов. Мы использовали модели, обученные решать различные задачи генерации текстов: машинного перевода, генерация парафразов, автоматического реферирования и упрощения предложений. Мы также рассматриваем популярные постановки задач, такие как обратный перевод и zero-shot генерация. Вторая часть набора данных – тексты, написанные людьми – собрана из открытых источников, относящихся к ряду предметных областей.

Участникам соревнования предлагается решить две задачи: (i) определить, был ли данный текст написан человеком или сгенерирован моделью (бинарная классификация), или (ii) определить автора текста (мультиклассовая классификация). В рамках соревнования мы предоставляем базовые решения в стандартной постановке задачи классификации на основе счетных признаков (TF-IDF) и модели архитектуры BERT. Кроме того, мы проводим оценку решения первой задачи разметчиками на краудсорсинговой платформе (human baseline). В общей сложности, соревнование привлекло внимание 38 решений: 30 для первой постановки задачи и 8 – для второй. Большая

*Equal contribution.

часть участников преодолела уровень базовых решений и уровень разметчиков. Используемая кодовая база, результаты оценки на краудсорсинговой платформе и другие материалы соревнования доступны в публичном GitHub репозитории соревнования.

Ключевые слова: распознавание сгенерированных текстов, генерация текстов, соревнование, автоматическое определение автора текста, нейронные сети

1 Introduction

Modern text generative models (TGMs) have demonstrated impressive results in generating texts close to the human level in terms of fluency, coherence, and grammar (Keskar et al., 2019; Zellers et al., 2019; Brown et al., 2020; Rae et al., 2021). However, the misuse potential of TGMs increases with their capabilities to generate more human-like texts. Malicious users can deploy TGMs for spreading propaganda and fake news (Zellers et al., 2019; Uchendu et al., 2020; McGuffie and Newhouse, 2020), augmenting fake product reviews (Adelani et al., 2020), and facilitating fraud, scams, and other targeted manipulation (Weidinger et al., 2021). The increasing difficulty for laypeople and users to discriminate machine-generated texts from human-written ones facilitates the spread of such misuse (Karpinska et al., 2021; Uchendu et al., 2021). This motivates the *artificial text detection* task (Jawahar et al., 2020), a fast-growing niche field aimed at mitigating the misuse of TGMs.

The Russian Artificial Text Detection (RuATD) shared task explores the problem of artificial text detection in Russian. Unlike existing datasets for English, our approach includes a range of task-specific TGMs, that is, models fine-tuned for common text generation tasks at the sentence- and document-level. On the one hand, such a setting challenges the participants and crowd-sourced annotators. On the other hand, it also enables many research and development purposes, such as training and benchmarking artificial text detectors, warning users about potentially fake content on social media and news platforms, filtering corpora augmented with TGMs, exploring detectors’ robustness w.r.t. TGMs’ architecture, size, downstream task, or domain. The shared task dataset consists of publicly available texts across multiple domains and texts generated by various monolingual and cross-lingual TGMs. The setup includes two sub-tasks: (i) to determine if a given text is automatically generated or written by a human (binary classification), and (ii) to identify the author of a given text (multi-class classification).

The main contributions of this paper are the following:

1. We propose a diverse automatic text detection corpus in Russian, the first of its kind (§2.2);
2. We model two competition sub-tasks (§4.1) after the traditional concepts of “Turing test” and authorship attribution for neural text generation models (Uchendu et al., 2021). We establish two count-based and neural-based baseline solutions (§4.4) and the human evaluation on the binary classification problem (§4.6);
3. We conduct an extensive analysis of the received submissions for both sub-tasks (§5) and discuss potential research directions (§6);
4. We set up the shared task environment, which remains open for the community submissions to facilitate future research in the area (§4.3).

2 Dataset

2.1 Text Generation

The corpus includes texts from 14 text generators, i.e., one human writer and 13 monolingual/multilingual TGMs varying in their number of parameters, architecture choices, and pre-training objectives. Each model is fine-tuned for one or more of the following text generation tasks: MT, paraphrase generation, text simplification, and text summarization. We also consider back-translation and zero-shot generation approaches. Figure 1 outlines the dataset creation pipeline. Text generation hyperparameters for each model are presented in 1. Note that we mostly use the default hyperparameters under the considered libraries.

Human We collect human-written texts from publicly available resources among six domains (see Section 2.2 for more details). Gold standard references from task-specific datasets are also used as human texts, since they are generally written and/or validated by crowd-source annotators (Artetxe and

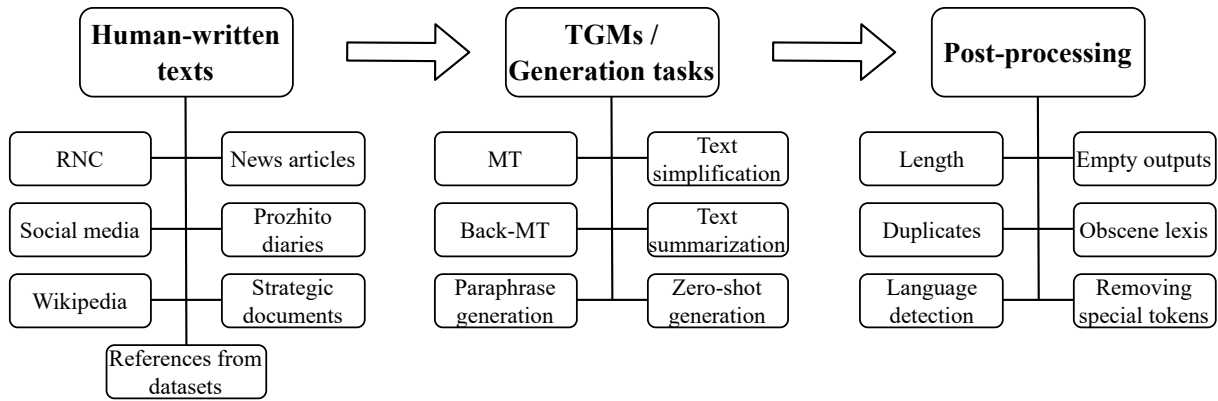


Figure 1: Data collection, text generation, and post-processing procedure.

Task	Model	Hyperparameters
Machine Translation	OPUS-MT	BEAM SEARCH; num beams=5
	M-BART50	
	M2M-100	
Back Translation	OPUS-MT	BEAM SEARCH; num beams=5
	M-BART50 M2M-100	
Paraphrase Generation	mT5-Small, mT5-Large	TOP-P SAMPLING; p=0.9
	ruT5-Base-Multitask ruGPT2-Large, ruGPT3-Large	
Simplification	mT5-Large, ruT5-Large	TOP-P SAMPLING; p=0.9
	ruGPT3-Small, ruGPT3-Medium, ruGPT3-Large	
Summarization	M-BART	BEAM SEARCH; num beams=5
	ruT5-Base	
Zero-shot Generation	ruGPT3-Small, ruGPT3-Medium	TOP-P SAMPLING; p=0.95; max length=90 percentile of length distribution by domain
	ruGPT3-Large	

Table 1: A brief description of the text generation hyperparameters and decoding strategies by text generation task.

Schwenk, 2019; Schwenk et al., 2021; Scialom et al., 2020; Hasan et al., 2021). The human texts serve as the input to the TGMs.

MT & Back-translation We use three MT models via the EasyNMT framework¹: OPUS-MT (Tiedemann and Thottingal, 2020), M-BART50 (Tang et al., 2020), and M2M-100 (Fan et al., 2021). We use subsets of the Tatoeba (Artetxe and Schwenk, 2019) and WikiMatrix (Schwenk et al., 2021) datasets to obtain translations among three language pairs: English-Russian, French-Russian, and Spanish-Russian. In the back-translation setting, the input sentence is translated into one of the target languages, and then back into Russian.

Paraphrase Generation Paraphrases are generated with models available under the `russian-paraphrasers` library (Fenogenova, 2021a): `ruGPT2-Large`², `ruT5-Base-Multitask`³, and `mT5` (Xue et al., 2021) of Small and Large versions.

Text Simplification We fine-tune `ruGPT3-Small`⁴, `ruGPT3-Medium`⁵, `ruGPT3-Large`⁶, `mT5-Large`, and `ruT5-Large`⁷ for text simplification on a filtered version of the RuSimpleSentEval-2022 dataset (Sak-

¹github.com/UKPLab/EasyNMT

²hf.co/sberbank-ai/rugpt2large

³hf.co/cointegrated/rut5-base-multitask

⁴hf.co/sberbank-ai/rugpt3small

⁵hf.co/sberbank-ai/rugpt3medium

⁶hf.co/sberbank-ai/rugpt3large

⁷hf.co/sberbank-ai/rugt5-large

hovskiy et al., 2021; Fenogenova, 2021b). Fine-tuning of each model is run for 4 epochs with the batch size of 4, learning rate of 10^{-5} , and weight decay of 10^{-2} .

Text Summarization We use two abstractive summarization models fine-tuned on the Gazeta dataset (Gusev, 2020): ruT5-base⁸ and M-BART⁹.

Zero-shot Generation We generate texts in a zero-shot manner by prompting the model and specifying the maximum number of generated tokens. The models include ruGPT3-Small, ruGPT3-Medium, ruGPT3-Large.

2.2 Data

Pre-training corpora of TGMs can cover multiple versatile domains (Liu et al., 2020), which prompt their abilities to generate texts with specific lexical, syntactic, discourse and stylistic properties. Despite this, the ATD task is generally explored w.r.t. only one particular domain, e.g., product reviews (Adelani et al., 2020), social media posts (Fagni et al., 2021), or news (Uchendu et al., 2021). Such setting limits the scope of evaluation of artificial text detectors. A few studies show that performance of modern detectors can vary drastically across domains (Bakhtin et al., 2019; Kushnareva et al., 2021), which stimulates the development of more generalizable and robust methods (Jawahar et al., 2020).

This paper aims at providing a diverse shared task data, taking into account the current limitations in the niche ATD field, and the diversity of TGMs widely used in the industry and NLP research for Russian. To this end, we consider domains which represent normative Russian, as well as general domain texts, social media posts, texts of different historical periods, bureaucratic texts with complex discourse structure and embedded named entities, and other domains included in the task-specific datasets, such as subtitles and web-texts. Recall that aside from linguistic and stylometric properties, texts differ in their length (e.g., sentence-level vs. document-level), and specifics attributable to the downstream tasks. We now list domains of texts that are fed into the previously described TGMs.

Russian National Corpus We use the diachronic sub-corpora of the Russian National Corpus¹⁰ (RNC), which covers three historical periods of the society and the Modern Russian language (“pre-Soviet”, “Soviet”, and “post-Soviet”).

Social Media We parse texts from multiple social media platforms that are marked with certain hashtags, such as dates, months, seasons, holidays, the names of large cities in Russia, etc. These texts are typically short, written in informal style and may contain emojis and obscene lexis.

Wikipedia We select the top-100 most viewed Russian Wikipedia pages spanning the period of 2016-2021 according to the PageViews¹¹ statistics.

News Articles The news segment covers different news sources in the Taiga corpus (Shavrina and Shapovalova, 2017) and the corus library¹², including but not limited to Lenta, KP, Interfax, Izvestia, Gazeta. We additionally parse more recent news articles to prevent potential data leakage and cheating.

Prozhito Diaries Prozhito is a corpus of digitized personal diaries, written during the 20th century (Melnichenko and Tyshkevich, 2017).

Strategic Documents are produced by the Ministry of Economic Development of the Russian Federation. The documents are written in bureaucratic style, rich in embedded entities, and have complex syntactic and discourse structure. This dataset has been previously used in the RuREBus shared task (Ivanin et al., 2020).

2.3 Post-processing

Each generated text undergoes a post-processing procedure based on a combination of language processing tools and heuristics. First, we discard duplicates, copied inputs, empty outputs, and remove special tokens from the generated texts (e.g., `<s>`, `</s>`, `<pad>`, etc.). Next, we empirically define length

⁸hf.co/IlyaGusev/rut5-base-sum-gazeta

⁹hf.co/IlyaGusev/mbart-ru-sum-gazeta

¹⁰ruscorpora.ru

¹¹<https://pageviews.wmcloud.org/>

¹²github.com/natasha/corus

Task	Text Generator	Domain	Num. samples	Num. tokens	IPM
Machine Translation	Human OPUS-MT M-BART50 M2M-100	Tatoeba WikiMatrix	35860	11.5	0.89
Back Translation	Human OPUS-MT M-BART50 M2M-100	Strategic documents News Prozhito RNC Wikipedia Tatoeba WikiMatrix	35588	12.9	0.88
Paraphrase Generation	Human mT5-Small mT5-Large ruT5-Base-Multitask ruGPT2-Large ruGPT3-Large	Strategic documents News Prozhito RNC Wikipedia Social media	44298	13.0	0.85
Simplification	Human mT5-Large ruT5-Large ruGPT3-Small ruGPT3-Medium ruGPT3-Large	Strategic documents News Prozhito RNC Wikipedia Social media	44700	18.3	0.86
Summarization	Human M-BART ruT5-Base	Strategic documents News Prozhito RNC Wikipedia	17164	33.5	0.86
Zero-shot Generation	Human ruGPT3-Small ruGPT3-Medium ruGPT3-Large	Strategic documents News Prozhito RNC Wikipedia	37499	141.5	0.85

Table 2: Text generators, domains and the final number of samples per task. The number of human-written texts is same as machine-generated texts.

intervals for each generation task based on a manual analysis of length distributions in `razdel`¹³ tokens. The texts are filtered by the following token ranges: 5-to-25 (**MT**, **Back-translation**, **Paraphrase Generation**), 10-to-30 (**Text Simplification**), 15-to-60 (**Text Summarization**), and 85-to-400 (**Zero-shot Generation**). We additionally discard the social media texts containing obscene lexis according to the corpus of Russian obscene words¹⁴, and keep the MT/Back-translation texts classified as Russian with the confidence of more than 0.9 (`langdetect`¹⁵).

3 Dataset Statistics

This section describes various count-based statistics of our dataset for human-written and machine-generated texts.

General Statistics Table 2 shows general dataset statistics w.r.t. text generation task, text generator, and domain. On average, there are 37.9 tokens in each text, with variations depending on the task. We estimate the frequency of each text according to the Russian National Corpus (RNC)¹⁶. It is computed as the number of frequently used tokens (i.e., the number of instances per million, that is, IPM in RNC is higher than 1) divided by the number of tokens in a sentence. The average IPM is 0.86 for the human-written texts and 0.87 for the machine-generated ones.

Diversity Metrics We estimate the diversity of the texts in terms of their k -gram statistics and lexical richness. We calculate two diversity metrics upon k -gram statistics: $\text{Dist-}k$ (Li et al., 2016) and Ent-

¹³github.com/natasha/razdel

¹⁴github.com/odaykhovskaya/obscene-words

¹⁵github.com/fedelopez77/langdetect

¹⁶ruscorpora.ru/new/en

	k	H	M
Dist-k	1	0.35	0.40
	2	0.75	0.76
	4	0.74	0.77
Ent-k	1	7.86	7.30
	2	9.99	9.03
	4	10.16	9.19

Table 3: Dist- k and Ent- k diversity measures by the target level. **H**=Human-written texts; **M**=Machine-generated texts.

	Back-MT		MT		Zero-shot Gen.		Paraphrase Gen.		Simplification		Summarization		Overall	
	H	M	H	M	H	M	H	M	H	M	H	M	H	M
Words	10.04	10.66	9.65	8.84	106.05	112.48	10.21	11.64	14.48	14.70	24.46	30.34	28.82	30.72
Terms	9.70	10.07	9.33	8.47	73.52	95.40	9.87	11.33	13.72	13.55	22.32	26.16	22.64	26.95
TTR	0.95	0.95	0.98	0.97	0.70	0.86	0.96	0.97	0.95	0.93	0.93	0.87	0.91	0.93
CTTR	2.09	2.13	2.08	1.97	4.91	6.20	2.13	2.31	2.52	2.47	3.14	3.33	2.76	3.01

Table 4: Lexical richness metrics per text generation task.

k (Zhang et al., 2018). Dist- k is the total number of k -grams divided by the number of tokens in the text set. Ent- k is an entropy metric that weights each k -gram so infrequent k -grams are penalized and contribute less to diversity. We compute the diversity scores for texts grouped by label and report them for $k \in \{1, 2, 4\}$ in Table 3.

To measure the lexical diversity¹⁷ of the texts in our dataset, we calculate four types of metrics: word count, terms count, type-token ratio (TTR), and corrected type-token ratio (CTTR). Type-token ratio is computed as t/w and corrected type-token ratio is computed as $t/\sqrt{2} * w$, where t is the number of unique terms/vocabulary, and w is the total number of words.

We can see that the ratio of the diversity measures between the natural and artificial texts depends on the task, which is explained by the very task formulation. At the same time, artificial texts may include non-existent words, degenerated textual segments, or rare words, which can be attributed to more significant lexical richness metrics overall.

4 Setup

4.1 Tasks

The RuATD Shared task features two sub-tasks:

- I. Determine if a given text is automatically generated or written by a human. This sub-task is framed as a binary classification problem with two labels: **H** (human) and **M** (machine).
- II. Identify the author of a given text. This sub-task is modeled after the traditional problem of authorship attribution (Coyotl-Morales et al., 2006), particularly in the context of neural models (Uchendu et al., 2020). It is a multi-class classification problem with 14 target classes – a human writer and 13 TGMs.

Evaluation Each sub-task uses the accuracy score, a standard metric for classification with balanced classes, as the official evaluation metric.

4.2 Dataset Splits

We split the dataset into four sets in the 60/10/15/15 proportion ratio: train (130k), development (21k), public test (32k), and private test (32k). Each set is balanced by the number of target classes, text generator, text generation task, and domain¹⁸. These sets are used for both sub-tasks, with only the target

¹⁷Lexical richness

¹⁸The number of human-written texts is equal to the number of machine-generated texts for each domain and text generation task.

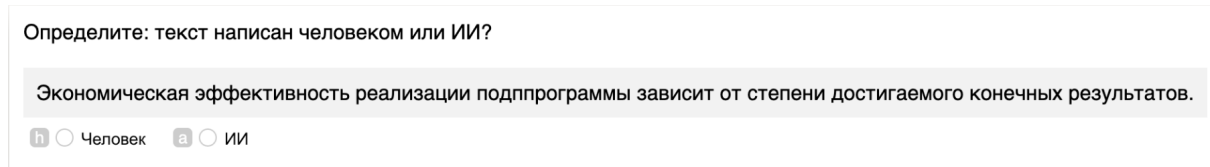


Figure 2: An example of the Toloka interface for the human evaluation setup.

classes changed, i.e., the **M** label is broken into 13 TGMs’ names in the multi-class sub-task.

4.3 Kaggle Setup

We use the Kaggle competition platform to run the shared task. The sub-tasks are set as separate competitions and leaderboards:

- I. The binary sub-task is hosted under this link;
- II. The multi-class sub-task is hosted under this link.

The participants are allowed to take part solely or in teams in both sub-tasks. The shared task comprises two stages: **public** and **private** testing. The first stage provides access to the public test set and 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 three submissions selected by the participants. Otherwise, the Kaggle platform automatically selects the three best submissions based on the participants’ public test scores. Participants are allowed to use any additional materials and pre-trained models, except for direct markup of the test set and search on the Internet.

4.4 Baseline

We provide the participants with two open-source baseline solutions: count-based (TF-IDF baseline) and BERT-based (BERT baseline) (Devlin et al., 2019). TF-IDF baseline is based on TF-IDF features coupled with the SVD dimensionality reduction and a Logistic Regression classifier. The TF-IDF has 50k features, further reduced to 5000 by SVD. The BERT baseline follows the default fine-tuning and evaluation procedure for the classification task under the HuggingFace transformers framework (Wolf et al., 2020).

4.5 Peer Review

Each participant is asked to publicly release their solutions and peer review other participants’ submissions. This step allows for a fair evaluation, eliminating the risks of potential cheating, such as solving the sub-tasks via a web search or other heuristics. After analyzing the assigned submission, the peer-reviewer should answer two questions in the Google form and provide comments, if any:

- Does the submission use a web search?
- Does the submission violate any other rules¹⁹ of the shared task?

4.6 Human Baseline

We conduct a human evaluation on the binary classification problem using stratified sub-samples from the public and private test sets. Each subset of 2.5k samples is balanced by the number of target classes, text generator, text generation task, and domain. We report the human baseline results during both public and private testing stages. The evaluation is run via the Toloka platform. The annotation setup follows the conventional crowd-sourcing guidelines for the ATD task and accounts for methodological limitations discussed in (Ippolito et al., 2020; Clark et al., 2021; Karpinska et al., 2021). We provide a full annotation instruction in Figure 1 (see Appendix A), and an example of the Toloka interface in Figure 2.

¹⁹The shared task rules are provided in the GitHub repository.

Binary sub-task		Multi-class sub-task	
Team	Accuracy	Team	Accuracy
MSU ✓	0.82995	Posokhov Pavel ✓	0.65035
Igor	0.82725	Yixuan Weng ✓	0.64731
Orzhan ✓	0.82629	Orzhan ✓	0.64573
mariananieva ✓	0.82427	MSU ✓	0.62856
Ivan Zakharov	0.82294	BERT baseline	0.59813
Yixuan Weng ✓	0.81767	Nikita Selin	0.58967
ilya koziev	0.81699	Victor Krasilnikov	0.55012
miso soup ✓	0.81178	Petr Grigoriev ✓	0.45814
Eduard Belov	0.80862	TF-IDF baseline	0.44280
Posokhov Pavel ✓	0.80630	Anastasiya Shabaeva	0.05411
Kirill Apanasovich	0.80308		
Tumanov Alexander	0.79778		
BERT baseline	0.79666		
Elizaveta Nosova	0.79595		
mipatov	0.78591		
akstar	0.78442		
Nikita Selin	0.78228		
David Avagyan	0.77869		
Mikhail Yumanov	0.77181		
Gregory Kuznetsov	0.75237		
Anastasiya Shabaeva	0.75178		
Shershunya	0.74534		
Ekaterina Kostrykina	0.74326		
Victor Krasilnikov	0.74091		
Alena	0.73589		
Alexander Tesemnikov	0.73204		
Lera Lelik	0.72727		
Dmitriy Vahrushev	0.71559		
Human baseline	0.66666		
Molostvov Pavel	0.68543		
Mental Sky	0.65326		
Petr Grigoriev ✓	0.64232		
TF-IDF baseline	0.64223		

Table 5: The official shared task results sorted in the descending order. Left: the binary leaderboard; Right: the multi-class leaderboard. Baseline submissions are colored in grey. ✓ stands for peer-reviewed submissions.

We grant access to the human evaluation project to only top-70% annotators according to the in-house Toloka rating system. Each annotator must first finish the training task by completing at least 80% of samples correct to get onto the main annotation task. We use the dynamic overlap of 3-to-5 annotators per sample. We discard votes from those annotators whose quality rate on the control tasks is less than 50%. We also filter out votes with the response time of less than 15 seconds per annotation task page (5 samples). The resulting vote is aggregated as the majority vote label.

5 Results and Analysis

We report the official shared task results of the private testing stage in Table 5. Only peer-reviewed submissions (marked with ✓ in the Table) are considered official.

As one can notice from Table, top-4 systems in the binary classification task have been peer-reviewed. In the multi-class setting, all four top-4 systems have been also peer-reviewed.

The results demonstrate that state-of-the-art classification models can be relatively successful in distinguishing human-written texts from machine-generated ones for the Russian language and determining the exact model used for generation for the latter class. However, one can quickly notice a rather stark contrast between the best scores obtained on the RuATD test set in binary setup (0.830 accuracy for *MSU*, the top-system in binary classification task) and scores obtained for a similar setup in English (0.970 accuracy; see (Uchendu et al., 2020) for reference). We attribute this contrast not to the difference in the languages but mainly to the nature of texts: in the English setup, an average text length is 432 words (compared to 31 in RuATD). This claim can be validated by splitting evaluation scores of the best binary RuATD models by length: on the texts longer than 23 words (about a quarter of all RuATD texts), top models can score over 0.95 accuracy.

Unsurprisingly all models that can outperform our BERT baseline used fine-tuned language models

(LMs) from the BERT family. Specific models that can achieve the best scores on the test set are mDeBERTa (He et al., 2021), and Russian-language implementations of RoBERTa (Liu et al., 2019). Top models experiment with learning-rate scheduling as well as other training techniques (e.g., adversarial training with fast-gradient method (Dong et al., 2018), or child-tuning training (Xu et al., 2021)).

Using additional features (e.g., lexical richness, perplexity, number of characters, number of sentences, TF-IDF of POS tags, punctuation, tonality, reading ease) provided only limited benefit. While there are competitive solutions with such features (e.g., *mariananieva*, 4th-placed solution in the binary setup), none of the three best models in either task used any additional features.

Ensembling models proved to be beneficial, although competitive results could be achieved using single models. For example, *Posokhov Pavel*, the best model in the multi-class setup task, does not use the ensembling of any kind, nor does *orzhan*, the third-placed model in both tasks.

Human Baseline The overall accuracy of the human evaluation is 0.66, which scores below the BERT baseline. The low results are consistent with recent studies (Karpinska et al., 2021; Uchendu et al., 2021), which underpin the difficulty of the task for crowd-sourcing annotators. These works advise hiring experts trained to evaluate written texts or conduct multiple crowd-sourcing evaluation setups with extensive training phases. We leave the human evaluation experiments for future work.

6 Discussion

On indistinguishable examples The reasons for the errors of various systems on the RuATD corpus are of separate research interest. A short meaningful sentence of frequency n-grams may often occur in a web-corpus and be easily reproduced by a simple statistical LM. Thus, the very definition of a specific automatic text can be a challenging task for an attentive annotator and even for an engineer directly involved in developing TGMs. This can be illustrated, for example, by the case of Ilya Sutskever from the GPT-3 project, who tweeted spring of 2022, that *large neural networks may be “slightly conscious.”*²⁰. The methodological problem of obtaining some significant phrases or texts randomly using LMs, however, is raised much earlier than the onset of “indistinguishability by the engineers themselves”: critical works on the Turing test (Turing and Haugeland, 1950) offer various variations of tests that level this problem. For example, (Bringsjord, 1996) explicitly note that a state machine that generates random sentences could be randomly considered meaningful by a judge in a good mood. In general, various methodological variations offer 1) interactive work with models/people, checking the maintenance of the context (Kugel, 1990) and even the consistency of the author’s “cognitive profile” (Watt, 1996). These areas can be considered topics for future work for the following shared tasks.

Ethical considerations Setting the task of detecting non-human texts is timely due to the rapid development of LMs. The very issue of detecting non-human texts affects the fundamental right of the user to understand when they interact with a subjectless technological solution and when - with a person. Problems of this kind are actively discussed in reviews of recent years. In particular, (Bommasani et al., 2021) define the scope of problems as:

1. foundation model misuse, including both purposeful generated text misuse and the unconditional reliance on automatic text classification results that can be false negative;
2. development of legal grounds to mitigate generative model misuse and detection model misuse;
3. widespread deployment of automatic text detection systems: the presented models can lead to an "arms race" between malicious content generators and detectors.

Although the improvement of language modeling is undoubtedly a fundamental task of machine learning, we are of the position that a thorough study of models that classify automatic texts is necessary. As practice shows, the percentage of their errors in the Russian language is non-zero.

7 Related Work

Many research efforts are related to natural language generation (NLG) models. These works can be characterized into two broad categories - (i) training LMs on large-scale data and (ii) learning to dis-

²⁰<https://towardsdatascience.com>

tinguish between machine-authored and human-written content. (Jawahar et al., 2020) provides a good survey on the automatic detection of machine-generated text for English.

Prior work has focused on training classifiers on samples from a model (Brown et al., 2020) and directly using a model distribution (Gehrmann et al., 2019). (Gehrmann et al., 2019) propose a visual and statistical tool named GLTR for the detection of generation artifacts across different sampling schemes. (Ippolito et al., 2020) compare human raters and automatic classifiers depending on the decoding strategy. They observe that classifiers can detect statistical artifacts of generated sequences while humans quickly notice semantic errors. Classifier accuracy ranges between 70% and 90% depending on the decoding strategy for short texts (64 tokens). (Dugan et al., 2020) propose a RoFT (Real or Fake Text) tool to detect the boundary between a human-written text passage and machine-generated sentences showing NLG models are capable of fooling humans by one or two sentences. A recent study of (Gallé et al., 2021) focuses on the unsupervised detection of machine-generated documents leveraging repeated higher-order n-grams. They show that specific well-formed phrases over-appear in machine-generated texts as compared to human ones. (McCoy et al., 2021) propose a suite of analyses called RAVEN for assessing the novelty of generated text, focusing on sequential structure (n-grams) and syntactic structure. Experiments show that random sampling result in generated text with a more significant number of novel n-grams.

Recent studies (Carlini et al., 2022; Lee et al., 2022) have raised a concern about model memorization due to data privacy leakage. (Carlini et al., 2022) confirm that memorization scales with model size and current LMs do accurately model the distribution of their training data. (Lee et al., 2022) investigate memorization and plagiarism when generating artificial texts. They observe that fine-tuned LMs demonstrate different patterns of plagiarism based on characteristics of auxiliary data. (Schuster et al., 2020) propose two benchmarks demonstrating the stylistic similarity between malicious and legitimate uses of LMs.

(Liyanage et al., 2022) propose a benchmark for detecting automatically generated research content that consists of a synthetic dataset and a partial text substitution dataset. The latter is created by replacing several sentences of abstracts with sentences generated by an NLG model. (Stiff and Johansson, 2021) adopt a wide variety of datasets of news articles, product reviews, forum posts, and tweets and investigated several classifiers to predict whether a text has been automatically generated. Their experiments show that classifiers perform reasonably accurately in the news domain, while the same task is more challenging for shorter social media posts.

8 Conclusion

We presented RuATD shared task, the first shared task on artificial text detection for the Russian language. As a result of the competition, 38 solutions have been obtained, solving the problem in two tasks modeled after the traditional concepts of the Turing test and authorship attribution for NLG methods.

The best solution of the shared task has achieved

- 83.0% accuracy in a binary task setup;
- 65.0% accuracy in a multi-class task setup.

The shared task dataset, codebase, human evaluation results, participant solutions, and other materials are now available online under Apache 2.0 license²¹.

The competition problem can be further treated as a Turing test in a non-interactive setting. First of all, its direct methodological extensions are possible in such applied areas as:

- dialogue systems, and
- applications for editors and writers.

Another direction for future work is to conduct a critical study on the human evaluation guidelines on artificial text detection, which is still an open methodological question in the field (Karpinska et al., 2021). We welcome the communities of NLP developers, linguists, and engineers to contribute to further research in the area and next criteria formulations.

²¹<https://github.com/dialogue-evaluation/RuATD>

Acknowledgments

The experiments were partially carried out on computational resources of HPC facilities at HSE University (Kostenetskiy et al., 2021). Ekaterina Artemova and Marat Saidov were supported by the framework of the HSE University Basic Research Program.

References

- David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. // *International Conference on Advanced Information Networking and Applications*, P 1341–1354. Springer.
- Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc’Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models.
- S Bringsjord. 1996. The inverted turing test is provably redundant. *Psychology*, 7(29).
- 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.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that’s ‘human’ is not gold: Evaluating human evaluation of generated text. // *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 7282–7296, Online, August. Association for Computational Linguistics.
- Rosa María Coyotl-Morales, Luis Villaseñor-Pineda, Manuel Montes-y Gómez, and Paolo Rosso. 2006. Authorship attribution using word sequences. // *Iberoamerican Congress on Pattern Recognition*, P 844–853. Springer.
- 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.
- Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. // *Proceedings of the IEEE conference on computer vision and pattern recognition*, P 9185–9193.

- Liam Dugan, Daphne Ippolito, Arun Kirubakaran, and Chris Callison-Burch. 2020. Roft: A tool for evaluating human detection of machine-generated text. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 189–196.
- Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Alena Fenogenova. 2021a. Russian paraphraser: Paraphrase with transformers. // *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, P 11–19, Kiyv, Ukraine, April. Association for Computational Linguistics.
- Alena Fenogenova. 2021b. Text simplification with autoregressive models.
- Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. 2021. Unsupervised and distributional detection of machine-generated text. *arXiv preprint arXiv:2111.02878*.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, P 111–116.
- Ilya Gusev. 2020. Dataset for automatic summarization of russian news. // *Artificial Intelligence and Natural Language*, P 122–134, Cham. Springer International Publishing.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 4693–4703, Online, August. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 1808–1822.
- 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.
- Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. // *Proceedings of the 28th International Conference on Computational Linguistics*, P 2296–2309, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.
- Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 1265–1285, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- 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*.
- Pavle Kostenetskiy, Roman Chulkevich, and Viacheslav Kozyrev. 2021. HPC resources of the Higher School of Economics. // *Journal of Physics: Conference Series*, volume 1740, P 012050. IOP Publishing.
- Peter Kugel. 1990. Is it time to replace turing’s test?’. // *1990 Workshop Artificial Intelligence: Emerging Science or Dying Art Form. Sponsored by SUNY Binghamton’s Program in Philosophy and Computer and Systems Sciences and AAAI*.

- Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 635–649, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2022. Do language models plagiarize? *arXiv preprint arXiv:2203.07618*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. // *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 110–119, San Diego, California, June. Association for Computational Linguistics.
- 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*.
- Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.
- Vijini Liyanage, Davide Buscaldi, and Adeline Nazarenko. 2022. A benchmark corpus for the detection of automatically generated text in academic publications. *arXiv preprint arXiv:2202.02013*.
- R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *arXiv preprint arXiv:2111.09509*.
- Kris McGuffie and Alex Newhouse. 2020. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. *arXiv preprint arXiv:2009.06807*.
- Michail Melnichenko and Natalia Tyshkevich. 2017. Prozhito from manuscript to corpus. *ISTORIYA*, 8(7 (61)).
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. Rusimplementeval-2021 shared task: evaluating sentence simplification for russian. // *Proceedings of the International Conference “Dialogue*, P 607–617.
- Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 1351–1361, Online, April. Association for Computational Linguistics.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8051–8067, Online, November. Association for Computational Linguistics.
- Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: «taiga» syntax tree corpus and parser. *Proceedings of the “Corpora*, P 78–84.
- Harald Stiff and Fredrik Johansson. 2021. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, P 1–21.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. // *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, P 479–480, Lisboa, Portugal, November. European Association for Machine Translation.

- Alan M Turing and J Haugeland. 1950. Computing machinery and intelligence. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, P 29–56.
- Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8384–8395, Online, November. Association for Computational Linguistics.
- Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. // *Findings of the Association for Computational Linguistics: EMNLP 2021*, P 2001–2016, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Stuart Watt. 1996. Naive psychology and the inverted turing test. *Psychology*, 7(14):463–518.
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.
- 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.
- Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. // *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 483–498, Online, June. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.
- Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. *Advances in Neural Information Processing Systems*, 31:1810–1820.

A Annotation Protocols

Задача. Человек или ИИ?

Вам необходимо выбрать:

- 1) текст написан человеком
- 2) текст написан искусственным интеллектом (ИИ).

Задание подробнее

- Прочтите предложенный текст.
- Определите: текст написан человеком или ИИ?
- Если вы считаете, что текст написал человек, — выберите ответ **“ЧЕЛОВЕК”**.
- Если вам кажется, что текст сгенерировал искусственный интеллект, — выберите ответ **“ИИ”**.

Примеры

Текст: *А это моя семья моя семья моя семья.*

Определите: текст написан ИИ или человеком?

Нажмите **“ИИ”**. Этот текст создан ИИ: текст будто “зациклен” на повторе *“моя семья”*.

Текст: *Софийский собор в Киеве, построенный в XI веке, является одним из самых значительных архитектурных сооружений данного периода. Первоначально Софийский собор представлял собой пятинефный крестово-купольный храм с 13 главами. С трёх сторон он был окружён двухъярусной галереей, а снаружи — ещё более широкой одноярусной.*

Определите: текст написан ИИ или человеком?

Нажмите **“ЧЕЛОВЕК”**. Это написал человек. Связность текста и исторические факты не нарушены.

Обратите внимание!

Вам будут встречаться тексты разной длины и разных видов: это могут быть новости, официальные документы, тексты из социальных сетей или художественные. По следующим особенностям можно определить, что текст написан ИИ:

- Несоответствие фактов и/или потеря связности текста
- Нарушение логической связи или знаний о мире
- Бессмысленные повторы или внезапное по смыслу завершение текста

Следующие признаки могут ввести в заблуждение, так как иногда встречаются и у людей и у машин:

- Грамматические/орфографические ошибки. Иногда даже люди совершают ошибки и опечатки, это нормально.
- Стиль. Современный ИИ может имитировать стиль текста и использовать разнообразную лексику. Если текст написан правильно, это необязательно человек!

Спасибо!

Figure 1: An example of the annotation instruction for the human evaluation.

Topic modeling of the Russian short stories of 1900–1930s: the most frequent topics and their dynamics

Tatiana Sherstinova, Anna Moskvina, Margarita Kirina,
Asya Karysheva, and Evgenia Kolpashchikova

National Research University Higher School of Economics, Saint Petersburg
121 Kanala Griboedova Embankment, 190068, Saint Petersburg, Russia
{tsherstinova, admoskvina, mkirina}@hse.ru
{askarysheva, eokolpaschikova}@edu.hse.ru

Abstract

The article describes the results of an experiment on topic modeling of Russian short stories for three successive historical periods of the early 20th century: 1) the beginning of the 20th century until 1913, 2) the war-revolutionary period (1914–1922), and 3) the early Soviet period (1923–1930). Using the Latent Dirichlet Allocation (LDA) algorithm, 9 models were built — 3 samples of different sizes (100, 500, and 1000 stories) for each of the periods. It turned out that in every model there are very frequent “themes” (topics) that characterize with a high probability a fairly significant share of texts in each sample. Moreover, one can also observe a meaningful dynamics of these frequent topics over different time periods, which allows us to consider them as thematic and stylistic markers of the analyzed text collections along with the more traditional quantitative measures of text analysis. The variety of frequent topics turned out to be higher in the second and third periods, which can be explained by the greater lexical and stylistic diversity of the prose of the “era of change”.

Keywords: quantitative text analysis; topic modeling; fiction; Russian short story; topic variety; dynamics of language and style

DOI: 10.28995/2075-7182-2022-21-512-526

Тематическое моделирование русского рассказа 1900–1930: наиболее частотные темы и их динамика

Шерстинова Т. Ю., Москвина А. Д., Кирина М. А.,
Карышева А. С., Колпащикова Е. О.

Национальный исследовательский университет «Высшая школа экономики»,
Санкт-Петербург
Россия, 190068, Санкт-Петербург, наб. канала Грибоедова, 121
{tsherstinova, admoskvina, mkirina}@hse.ru
{askarysheva, eokolpaschikova}@edu.hse.ru

Аннотация

В статье описаны результаты эксперимента по построению тематических моделей малой русской прозы (русского рассказа) трех последовательных исторических периодов начала XX века: 1) начала XX века до 1913 г. включительно, 2) военно-революционного периода (1914–1922) и 3) раннесоветского периода (1923–1930). С помощью алгоритма латентного размещения Дирихле (LDA), построено 9 моделей (по 3 выборки разного размера для каждого из периодов – по 100, 500 и 1000 рассказов). Оказалось, что в каждой из моделей присутствуют весьма частотные «темы» (топики), характеризующие довольно существенную долю текстов каждой выборки с высокой вероятностью, а также наблюдается содержательная динамика этих частотных тем по разным временным периодам, что позволяет считать их тематико-стилистическим маркерами анализируемых коллекций текстов наряду с более традиционными количественными мерами анализа текстов. Разнообразие частотных топиков оказалось выше во втором и третьем периоде (для

выборки в 500 и 1000 рассказов), что можно объяснить большим лексико-стилистическим разнообразием прозы «эпохи перемен».

Ключевые слова: квантитативный анализ лексики; тематическое моделирование; художественная проза; русский рассказ; тематическое разнообразие; динамика языка и стиля

1 Введение

Построение тематических моделей для коллекции текстовых документов — активно развивающееся направление автоматической обработки текста [5; 18; 19; 25; 26]. Этот метод машинного обучения позволяет выявлять из корпуса текстов скрытые семантические структуры — темы. Под «темой» понимается «набор ключевых слов, характеризующий отдельный документ или набор документов» [19, с. 221]. Каждый текст в анализируемом корпусе представляется как набор тем, т. е. один текст описывается несколькими темами одновременно. Стоит отметить, что зачастую при тематическом моделировании реализуется подход, называемый *bag-of-words*, т. е. не учитываются грамматические и синтаксические характеристики слов. Любой текст осмысливается как «случайная выборка слов, порожденная неким множеством тем» [там же]. Важно также и то, что при тематическом моделировании в результате бикластеризации происходит объединение в семантически схожие группы не только слов, но и текстов [45, с. 216].

Изначально методы тематического моделирования разрабатывались для анализа специальных (научных, технических, новостных и т. п.) документов. В последние годы появляется все больше примеров, когда эти методы применяются для интеллектуального анализа и кластеризации художественной прозы [8; 20; 21; 22; 24]. Однако результаты, которые получаются при обработке литературных текстов, принципиальным образом отличаются от результатов обработки текстов специальных [3; 28; 31; 42]. Метафоричность языка художественных произведений, образность повествования, применение разнообразных стилистических приемов, а во многих случаях и отсутствие ярко выраженной «темы повествования» приводят к тому, что математические подсчеты совместного употребления слов, лежащие в основе методов тематического моделирования, дают не совсем те результаты, которые исследователь ожидает получить от применения этих методов. С другой стороны, в отличие от традиционного анализа художественного текста, результатом построения такой тематической модели в большинстве случаев является не информация «в чем главная идея/содержание рассматриваемых текстов», сколько информация о тематико-стилистическом разнообразии анализируемой текстовой коллекции, в которой каждая из выделенных тем-топиков состоит из семантически связанных слов, формирующих некоторую интерпретируемую лексико-семантическую группу (или тематическое ядро).

Данное исследование продолжает серию работ по изучению языка, стиля и тематического разнообразия малой русской прозы первых трех десятилетий XX века, начатого в [6; 14; 15; 33; 35; 37; 38; 44], осуществляемых главным образом на материале Корпуса русского рассказа 1900–1930 гг. [16; 17; 34], и ставит своей целью построение тематических моделей для трех последовательных исторических периодов, каждый из которых соотносится со значимой исторической эпохой: I период (1900–1913 гг.) — начало XX века до Первой мировой войны, II период (1914–1922 гг.) — эпоха острых социальных катаклизмов, войн и революций (Первая мировая войны, Февральская и Октябрьская революция, Гражданская война), III период (1923–1930 гг.) — становление молодого советского государства. Вслед за [38; 39] мы полагаем, что исторический фон эпохи, в котором создаются литературные тексты, так или иначе будет проявляться в художественном творчестве современных ему писателей и оказывать влияние не только на язык, но и на содержание и тематику литературных произведений, косвенным отражением которых можно считать темы/топики, полученные в результате тематического моделирования.

В статье описаны результаты эксперимента по тематическому моделированию малой русской прозы (русского рассказа) начала XX века. С помощью алгоритма латентного размещения Дирихле (LDA) построено 9 тематических моделей, из которых выделены и описаны наиболее частотные топики, а также продемонстрировано изменение состава частотных топиков в зависимости от объема выборки.

2 Материал и методика

2.1 Отбор литературных текстов

Анализ динамики тематических моделей художественной прозы проводится на материале русского рассказа. Выбор рассказа как жанра для проведения исследования определяется тем, что рассказ является наиболее распространенным жанром прозы, охватывающим все литературные направления — рассказы присутствуют в творчестве практически всех прозаиков (и даже многих поэтов!), что позволяет вовлечь в исследование тексты максимального количества авторов, получить тем самым наиболее статистически достоверный «литературный портрет» эпохи и оценить все его тематическое разнообразие. Малый текстовый объем рассказа способствует тому, что рассказы значительно быстрее, чем повести и романы, проходят издательский цикл, достаточно большая их доля публикуется в литературных журналах, которые также вовлекаются в единый художественно-литературный процесс эпохи [17]. Можно утверждать, что рассказ, как особый жанр, выполняет «разведочную» функцию и даже работает на опережение, чутко улавливая и реагируя на изменения в общественном сознании и культуре общества» [там же], является «диагностом социальных процессов» [16]. Наконец, «*в русской литературе рассказ традиционно был сильным жанром. Пожалуй, лишь американская литература приближается в этом отношении к нашей*» [46].

Приведенные аргументы стали причиной того, что для задачи моделирования языка русской прозы были начаты именно для жанра рассказа. С этой целью был создан и продолжает развиваться Корпус русского рассказа¹ [13; 16; 17; 34].

По сравнению с предыдущими исследованиями, посвященными тематическому моделированию русского рассказа [6; 14; 15; 33; 35; 37; 38; 44], для получения более достоверных результатов мы существенно (в 10 раз) расширяем объем исследовательского материала — до 1000 рассказов в каждом из изучаемых периодов. Более того, нам показалось целесообразным посмотреть, насколько меняются результаты тематического моделирования при последовательном расширении исследовательской выборки — для 100, 500 и 1000 текстов, тем самым рассмотреть зависимость тематической модели от объема выборки. Ограниченный объем публикации вынуждает нас остановиться на рассмотрении наиболее частотных тем (топиков) художественных текстов для каждого временного среза, которые, однако, представляются достаточно показательными для русской литературы изучаемого периода.

Чтобы оценить изменение «тематического разнообразия» в динамике, было решено подготовить 9 выборок: по 100, 500 и 1000 рассказов для трех последовательных временных срезов — довоенного (1900-1913), военно-революционного (1914-1922) и раннесоветского (1923-1930). Поскольку конечной целью проводимых исследований является моделирование национального литературного процесса [36; 41], при формировании выборки основной акцент делался на включение текстов максимального количества русских писателей, работавших в жанре рассказа [16; 17], не только «известных» и хрестоматийных, но и малоизвестных и даже фактически забытых.

Основными источниками для формирования выборки стали два открытых литературных ресурса — библиотека Lib.ru Максима Машкова [9], являющаяся одним из старейших и представительных ресурсов русского литературного мира и уже упомянутый Корпус русского рассказа 1900-1930 гг. [13; 16; 17; 34], содержащий большой объем редких текстов, специально оцифрованных для этого корпуса. Однако для поставленных задач исследования этих двух текстовых коллекций оказалось недостаточно, и нам пришлось обращаться к другим открытым интернет-ресурсам (в частности, [10; 11; 12; 30; 43]), для пополнения текстовых коллекций раннесоветского периода.

При формировании выборок ставилась две задачи: обеспечение относительно равномерного распределения текстов по году написания/первой публикации рассказа и обеспечение максимальной представительности разных авторов внутри как периода в целом, так и для каждого отдельного года. Конкретные рассказы отбирались в случайном порядке, вне зависимости от тематики и содержания. В выборки не включались как достаточно крупные

¹В настоящее время работа идет над периодом первых трех десятилетий XX в.

рассказы (больше 10000 словоупотреблений), так и очень краткие (менее 200 слов). Отобранные тексты анализировались целиком, вне зависимости от размера. Выборки меньшего объема входили подмножеством в выборки большего объема (то есть тексты выборки из 100 рассказов входят в выборку 500 и 1000).

Всего для исследования отобраны были отобраны 3000 текстов: 74% из них взяты Lib.ru, 24 % — из Корпуса русского рассказа и 2 % — из других открытых источников. В табл. 1 приведены данные по объему итоговых выборок в словоупотреблениях.

Период	Объем выборки			Количество разных писателей		
	100	500	1000	100	500	1000
1 (1900-1913)	369113	1820446	3564493	100	320	690
2 (1914-1922)	321693	1355994	2702207	100	254	366
3 (1923-1930)	302682	1516696	2541865	100	307	313

Таблица 1: Объем исследовательских выборок в словоупотреблениях.

Данные табл. 1 показывает, что суммарный объем текстов первого периода превышает выборки второго и третьего, а количество писателей, вовлеченных в литературный процесс в довоенный период, по нашим данным, также оказалось больше, чем в последующие годы. Это еще раз подтверждает выводы о том, что писательская активность в начале века была выше, чем в периоды острых социальных конфликтов и преобразований [17].

Далее по тексту выборки обозначаются в виде кода из двух чисел, первое из которых обозначает период, а второе — объем выборки в рассказах (напр., 2-500 — выборка текстов из второго периода в 500 рассказов).

2.2 Методика обработки данных

Тексты были лемматизированы с помощью библиотеки spaCy [7]. Из текстов удалялись все имена собственные и стоп-слова, служебные части речи и другие стоп-слова.

Для построения тематических моделей использовался алгоритм латентного размещения Дирихле (LDA), реализованный в библиотеке gensim [1; 2]. Оптимальное количество тем (топиков) для каждой выборки определялось автоматически на основе меры когерентности [29]: выбиралось то количество топиков, при котором достигалось наибольшее ее значение. Для подсчета метрики использовалась функция CoherenceModel из модуля models библиотеки gensim [27]. Диапазон количества топиков при вычислении когерентности был определен в интервале от 10 до 45; метрика вычислялась при построении моделей в цикле с шагом в 5 топиков.

Графики когерентности для определения оптимального количества тем приведены на рис. 1, а итоговая статистика о количестве автоматических полученных тем для каждого из периодов представлена в табл. 2. Неожиданно, для всех трех периодов максимальный «тематический» разброс показала выборка среднего размера в 500 текстов. Для первого и второго периода количество рекомендуемых тем для крупной выборки в 1000 текстов меньше, чем для малых выборок, для раннесоветской прозы, напротив, это число несколько возрастает.

Период	Объем выборки		
	100	500	1000
1 (1900-1913)	20	40	15
2 (1914-1922)	35	40	15
3 (1923-1930)	25	45	30

Таблица 2: Количество тем (топиков), выявленных для каждой выборки.

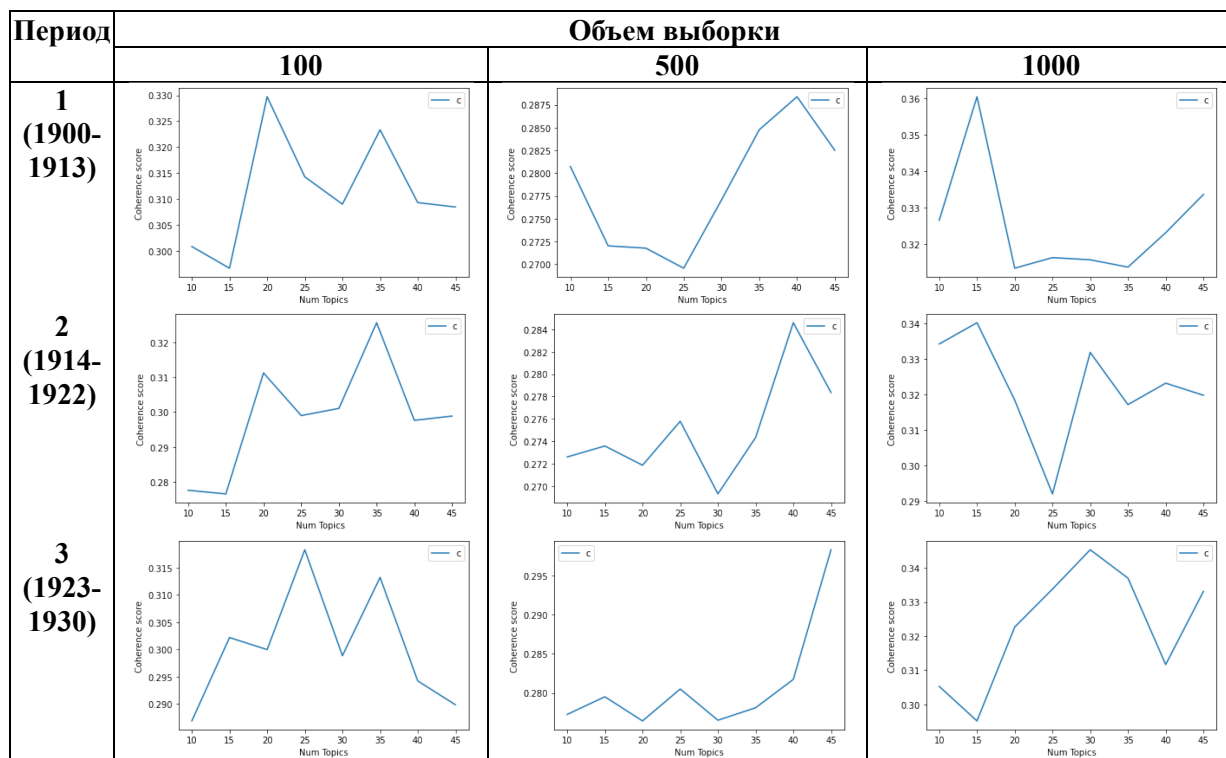


Рисунок 1: Графики когерентности для исследуемых выборок.

3 Результаты

Итогом работы алгоритмов тематического моделирования являются: 1) собственно список тем (топиков), представленный в виде набора ключевых слов, каждый из которых встречается в тексте с определенной вероятностью, 2) вероятности отнесения к каждой конкретной теме для каждого документа текстовой коллекции. Таким образом, имеет место «мягкая кластеризация» [19] исследуемой выборки. В центре нашего исследования находятся именно полученные темы, а также их частотность (распространенность) для выборки того или иного среза.

Ввиду больших объемов выборки и большого количества выявленных тем, в рамках данного исследования было решено ограничиться наиболее частотными темами, характеризующими максимальное количество анализируемых текстов.

Для этого среди всех полученных топиков, были отобраны темы, которые показали встречаемость для наибольшего количества документов. В качестве нижнего порогового значения было выбрано 25%, то есть ниже рассматриваются только те темы, которые превышают установленное моделью пороговое значение для более чем четверти документов выборки. Таких тем оказалось следующее количество (см. табл. 3).

Период	Объем выборки		
	100	500	1000
1 (1900-1913)	2	3	4
2 (1914-1922)	2	4	6
3 (1923-1930)	2	4	6

Таблица 3: Количество наиболее частотных тем (топиков) для каждой выборки.

Однако детальное рассмотрение наполнения каждой из этих тем показало, что многие рассказы входят туда с небольшими вероятностями. Поэтому было решено взять еще два среза данных — подсчитать количество документов, входящих в каждую из тем: 1) с вероятностью не

менее 75%, 2) с вероятностью не менее 95%. В табл. I–III Приложения² представлены эти самые частотные темы для довоенного, военно-революционного и раннесоветского периодов соответственно. Темы представлены их номером (столбец № 2), с точки зрения статистики являющимся номинальной переменной [4], содержание топика с вероятностями ключевых слов, округленных до 10000-тысячных, представлено в правом столбце, и также приводится относительное количество рассказов, содержащих соответствующую тему с высокой вероятностью.

Рассмотрим полученные результаты.

Характеризуя слова, образующие топики, мы решили использовать понятие «мотив», поскольку для описания литературных текстов это кажется более привычным, чем «ключевые слова» или термины.

При выборке в 100 текстов во всех трех периодах по предложенной методике выделяются всего по 2 темы, все они достаточно общего плана. Для первого периода это топики 17(1-100) и 18(1-100). Темы представляют собой оппозицию статики и динамики (*стоять* – *пойти*, *прийти*), молодости-старости (*молодой* – *старик*), личного и общего (*душа* – *народ*), одиночества – семьи (*женщина* – *отец*, *мать*), духовного и материального (*душа* – *деньги*). Есть в темах и существенные пересечения — это «голос», «спросить» и «земля». Мотив «голоса» дважды появляется в частотных темах выборки 500, для выборки в 1000 он уже уходит из верхней зоны тем, «земля» появляется в двух темах больших выборок, а слово «спросить» является абсолютным лидером довоенного периода, характеризуя 7 из 9 частотных тем.

Оценить соответствие выделенных топикиков «тематике» текста можно на примере отдельных рассказов. Так, продажа/покупка становится центральным мотивом рассказа П. И. Астрова «Из жизни человека» (1907), с вероятностью 0,91 относящегося к теме 18(1-100)³: ('отец', 0.0095), ('пойти', 0.0069), ('старик', 0.0065), ('мать', 0.0057), ('деньга', 0.0046), ('прийти', 0.0041), ('земля', 0.0039), ('голос', 0.0039), ('народ', 0.0038), ('спросить', 0.0037). Главный герой — «старик» Алейка звонким «голосом» вопрошает на улице: «Нет ли чего продавать?», на что человек «с глухим сдавленным *голосом*» высовывается из окна своей комнаты и просит Алейку зайти к нему. Умиравший барин предлагает Алейке свои последние вещи; Алейка долго не соглашается на сделку, повторяя, что барин назначил слишком большую сумму. Однако все-таки Алейка покупает вещи барина, отдав за них меньше денег, чем изначально просил «продавец»... В рассказе много глаголов движения «*пойти*»/«*прийти*», во время торгов старика с барином неоднократно встречается «*спросить*» и «*деньги*». «*Народ*» и «*земля*» появляется в самом конце рассказа, в котором говорится о похоронах умершего барина. Два слова из топика («*отец*» и «*мать*») не встретились в рассказе ни разу⁴.

При расширении выборки до 500 текстов, «динамичная» тема 18(1-100) модифицируется в тему 3(1-500), появляется мотив «мужика», значительная доля ее образующих слов — это глаголы, причем не только динамичные, но и статичные (*смотреть*, *стоять*). А тема первого периода 17(1-100), которую можно условно назвать «молодая женщина в комнате», при выборке в 500 текстов распадается на две — тему «любви» 2(1-500) и тему «семьи» 22(1-500).

С высокой вероятностью (0,999) относится к теме 2(1-500) рассказ В. Березовского «Утро» (1901): Иванов идет через лес к своей возлюбленной. Он наслаждается природой и своей любовью. В то же время его возлюбленная не спит всю ночь, ждет его у окна. Наконец, с приходом утра, Иванов приходит к дому. Возлюбленная вспоминает, что забыла отпереть дверь, и еще несколько минут не опирает, предвкушая встречу со счастьем. Наконец встреча происходит. Как видно, содержание рассказа действительно хорошо коррелирует с автоматически определенной для него темой.

На выборке в 1000 текстов мы уже имеем 4 частотные темы. Тему 6(1-1000) можно считать «преемницей» тем 18(1-100) и 3(1-500), но возникает новая окраска — «ночь», «лес». Стоит

² Содержание частотных тем вынесено в приложение в конце статьи, чтобы не превышать требуемого объема.

³ См. Табл. I Приложения.

⁴ Чем больше вероятность вхождения рассказа в топик, тем больше вероятность встретить в тексте рассказа ключевые слова топика, причем неоднократно.

обратить внимание на высокую частоту этого топика: 10% или 100 рассказов из 1000 содержат в себе эту тему с вероятностью более 95%. Примером хорошего соответствия топика 6(1-1000) с вероятностью 0,999 можно считать рассказ Б. А. Верхоустинского «Лесное озеро» (1912) — действие происходит в лесу, на озере, потом дома, присутствует и дед-старик. Две следующие по частоте темы 5(1-1000) и 14(1-1000), которые можно условно назвать «любовь и семья» и «любовь и душа», восходят к темам 2(1-500) и 22(1-500), и ранее — к теме 17(1-100). Наконец, появляется тема 3(1-1000), которую можно условно назвать «Деньги – дом – служба». Это единственная из всех частотных тем с ориентацией на финансовое состояние и понятие «служба». При этом сохраняются мотив семьи («жена», «дом»).

Частотные топика второго, военно-революционного, периода во многом наследуют тематику рассказов начала века. Но предсказуемо появляются и новые мотивы. Так, на малой выборке в 100 рассказов максимальную частоту показали две темы 16(2-100) и 22(2-100), обе из которых можно считать «преемницами» темы «молодая женщина в комнате» первого периода. Между темами есть существенное пересечение в виде глаголов «пойти», «спросить», «сидеть». Время сдвигается в сторону «вечера» и «ночи», акцент с нейтрального «дом» в теме 16(2-100) смещается на «дверь» и «окно», которые можно считать метафорами *расставания/встречи* и *ожидания*. В теме 22(2-100) возникает мотив «письма», весьма частотный для русской прозы того периода [35], но особенно значимый во время войны и вынужденной разлуки, а также мотив *мысли*, наполняющий человеческое сознание в тревожные времена. Примером рассказа, относящегося и к 16, и к 22 топикам, могут служить «Ситцевые колокольчики» Ю. Л. Слезкина (1922). Один из героев рассказывает другому о том, как в молодости гостил в семье, где было 5 дочерей-красавиц, в одну из которых он был влюблен, но мог лишь позволить любоваться ею со стороны, поскольку она рано вышла замуж по любви. Переехав жить в другое место, герой ждет писем от своей старой знакомой и получает их, что коррелирует с мотивами ожидания и письма. Другой рассказ, относящийся к этим двум топикам — «Тень счастья» Н. Д. Телешева (1921), в котором главного героя разыгрывают коллеги, посылая любовные письма от прекрасной незнакомки. Герой никак не может встретиться с ней и страдает, особенно когда коллегам надоедает эта игра и они «убивают» его возлюбленную, прислав ему письмо с новостями о смерти девушки.

Новым мотивом тем этого периода является понятие «жизнь», которое никак не проявилось в частотных темах малой выборки из 100 рассказов, но присутствует с высокими вероятностями во всех четырех топиках выборки из 500 рассказов и в половине частотных тем выборки из 1000 рассказов. *Ценность жизни* наиболее остро проявляется во времена социальных катастроф, поэтому появление этого мотива вполне закономерно. Надо отметить, что мотив «жизни» сохранится и в прозе третьего, раннесоветского, периода. Самая частотная тема выборки 500 20(2-500) наиболее неоднородна: помимо «жизни на земле» и «голоса сердца» она окрашена оппозициями: «ночь – солнце», «белый – черный». Тема 24(2-500) формируется вокруг «дома», содержит много глаголов (*пойти, сидеть, спросить, смотреть*), субъект исчезает, но появляется мотив «ночи». Тема 25(2-500) — это традиционная тема женской любви, но с акцентом на мотивы «жизни» и «мысли». Наконец, тема 29(2-500) — первая в этом блоке, в котором угадывается революционный дух эпохи: действие переносится в «город», на «улицу», действующим лицом становится «толпа», но сквозь призму семейных (*жена, отец*) и духовных (*душа*) ценностей. К этому топика был отнесен, например, рассказ Н.Н. Никандрова «Катаклизма» (1917), показывающий митинги глазами простых городских жителей, которые даже не знают слова «митинг», а их протест напрямую связан со страхом за семью и близких.

Из 6 наиболее частотных тем выборки 1000 две — 0(2-1000) и 10(2-1000) — относятся к традиционной тематике «женщина и любовь», при этом 0(2-1000) восходит к типичной довоенной теме «женщина в комнате» (14(1-1000), 5(1-1000)), но с новым мотивом «нужности». Тема 10(2-1000) является наиболее светлой во всем этом блоке, ее темообразующие слова легко преобразуются в вполне связный текст, напр., «*Жизнь. Душа. Любовь. Женщина любит сердцем. Счастье для прекрасной девушки — письмо*». Тема 8(2-1000) весьма близка к теме 6(1-1000) первого периода (*земля, лес, отец, старый*), но окрашена оппозициями *белый-черный*, а также *земля-небо*, а тема 4(2-1000) имеет много пересечений с темой 3(1-500) — *пойти, отец, дом, смотреть, сидеть, мужик* (причем порядок слов, отражающий вероятности, тоже во многом совпадает). Новыми словами темы становятся

«мать» и отмеченные нами мотивы, общие для всего второго периода — «жизнь» и «окно». Уникальными в этом периоде является «созерцательный» топик 1(2-1000): «смотреть в окно в ночи» на «воду, лес и небо», но с оттенком оппозиции «белый-черный» и единственный «военный» топик 3(2-1000), который можно описать как «между жизнью и смертью», «мысли солдата/капитана в последнюю минуту». К «военному» топику был отнесен, к примеру, рассказ Н. Уклеина «Поезд мертвых» (1915), в котором на празднично подготовленный перрон приходит поезд, полный вражеских трупов. В последней теме кажется несколько неожиданной появление такого персонажа как «князь», впрочем, его вероятность явно меньше других темообразующих слов этого периода.

Ожидаемо, частотные топики третьего, раннесоветского, периода содержательно ближе к военно-революционному периоду, чем к рассказам начала века. Следует отметить, что у большинства частотных тем этого периода темообразующим становится слово «лицо» (появляются и другие слова, связанные описанием внешности героев — «голова», «глаз»), становится меньше глаголов движения, максимум для которых наблюдался во втором периоде, сохраняется актуальность понятия «жизнь», повышается значимость (частота) «земли» и «воды». На малой выборке в 100 рассказов выделяются две частотные темы, первая 9(3-100) из которых напоминает темы 8(2-1000), 6(1-1000), 1(2-1000) — *земля, вода, ночь, старик, белый-черный*, а вторая 12(3-100) содержательно представляет собой синтез уже знакомых с начала века мотивов (*голос, спросить, комната*) и слов, отражающих реалии нового времени — *тетка, работа, товарищ*. Так, образ *женщины*, частотный в 1-2 периодах, замещается образом *тетки-товарища*. Можно предположить, что две частотные темы первой выборки символизируют противопоставление старого и нового мира. Так, среди героев рассказа В. Инбер «Квартира № 32» (1924), принадлежащего к топику 12, есть Эсфирь Абрамовна, почти карикатурная обитательница коммунальной квартиры; слово «товарищ», впрочем, в тексте относится только к мужчинам.

Для выборки в 500 рассказов третьего периода определены 4 частотные темы, три из которых достаточно традиционны и встречались с небольшими изменениями ранее. Выделяется на их фоне тема 40(3-500), которая ассоциируется с состоянием *болезни*. Одним из рассказов, принадлежащих к этому топику, стало «Общежитие» В. Зазубрина (1923), в котором описываются соседи по общежитию; среди них есть доктор, пишущий научный труд, и упоминаются проблемы со здоровьем других соседей. Неожиданно, на этом срезе не проявилось ни одной частотной темы, маркированной советскими реалиями.

Наконец, на большой выборке в 1000 рассказов, мы имеем 6 наиболее частотных топиков, тематика которых распределяется следующим образом: 20(3-1000) имеет основным лейтмотивом «мужик/товарищ на земле», 26(3-1000) можно было бы назвать «собрание граждан/товарищей/рабочих» (кстати, здесь присутствует и более свойственные первому периоду «голос» и «спросить», но в этом контексте они воспринимаются совсем иначе). Примером рассказа о собрании товарищей являются «Именины» М.Я. Козырева (1925). В этом тексте переосмыслиется ритуал празднования (и другие части жизни) с учетом наступления новой эпохи. Герои празднуют вместе со своим начальником и обсуждают актуальные перемены в обществе. Интересна по составу тема 10(3-1000): связующими элементами являются объекты природы (*вода, лес, земля, снег*), при этом отсутствует явный субъект, хотя описывается его/ее *глаза, лицо, голова*. Тема 29(3-1000) представляет собой возвращение к традиционной «семейной жизни в любви» (см. 5(1-1000) в первом периоде). Последние две темы 23(3-1000) и 2(3-1000) можно условно отнести к развитию науки и техники и освоению новых территорий.

Обобщая полученные данные, можно сделать следующие выводы:

- 1) Наблюдается содержательная динамика частотных тем по разным временным периодам. Поэтому хотя тематические модели далеко не всегда являются отражением собственно «тематик» литературного текста, тем не менее их можно рассматривать как тематико-стилистический маркер анализируемой коллекции текстов.
- 2) Частотные темы, построенные на небольших выборках, трансформируются или перераспределяются на выборках большего объема, некоторые темы можно считать универсальными (сквозными) для всех трех рассмотренных периодов. Топики на

- больших выборках в среднем выглядят более конкретными с содержательной точки зрения.
- 3) Разнообразие частотных топиков выше во втором и третьем периоде — для выборок в 500 и 1000 рассказов. Очевидно, это определяется большим лексико-стилистическим разнообразием прозы «эпохи перемен». И это несмотря на то, что для первого периода выборка содержит максимальное количество разных авторов.
 - 4) В случае работы с художественным текстом оценка адекватности модели посредством оценки интерпретируемости порожденных тем является непростой задачей. Традиционные подходы, предполагающие понимание экспертом значения «удачного» топика, в некоторых случаях оказываются затруднительными, поскольку тема может представлять собой как группу семантически близких слов, так и свертку сюжета, а также подвергаться влиянию образности художественного языка и имплицитных смыслов произведения. Поэтому вопрос правильного подхода к оценке интерпретируемости результатов работы модели на материале художественной прозы должен стать темой отдельного исследования.
 - 5) Даже в том случае, когда конкретно выделенный топик является хорошо интерпретируемым и рассказ относится к нему с большой вероятностью, ключевые слова одного единственного топика как правило не исчерпывают основные тематические категории рассматриваемого текста.

4 Заключение

Проведенное исследование позволило выявить наиболее частотные темы рассказов для трех последовательных исторических периодов, полученные автоматически в результате построения тематической модели. Под частотными здесь понимаются темы, отнесенные к большему количеству проанализированных документов текстовой коллекции с высокой вероятностью. Разумеется, тематическое моделирование в большинстве случаев приводит к результатам, отличным от того, что понимается под темой литературного произведения при экспертной оценке. Тем не менее, построенные модели кажутся во многом осмысленными, позволяют посмотреть на литературные тексты с неожиданной стороны и выявить отличительные особенности больших массивов текстов, прочитать и переосмыслить которые не в состоянии ни один эксперт. Поэтому эксперименты с тематическим моделированием литературных произведений безусловно имеет смысл продолжать.

В нашем случае, как и в большинстве других компьютерных исследований литературного материала, встает вопрос об интерпретируемости полученных тем. Разумеется, полученные темы не являются (и не могут быть) отражением «темы рассказа» в литературоведческом понимании. В целом, на выборках большего объема рассмотренные частотные топика выглядят более «содержательными» и «интерпретируемыми», чем на меньших выборках. Тем не менее, полученные данные показывают, что содержание топиков между тремя исследуемыми периодами отличается, что позволяет считать их тематико-стилистическим маркером анализируемой коллекции текстов наряду с более традиционными количественными мерами анализа текстов.

Задача интерпретации автоматически полученных тем является непростой задачей, поскольку традиционные подходы, используемые при работе со специальными текстами, при анализе художественной литературы оказываются недостаточными вследствие содержательных и стилистических ее особенностей. Предполагается, что разработка правильного подхода к оценке интерпретируемости результатов работы модели на материале художественной прозы должна стать темой отдельного исследования.

При этом следует иметь в виду, что на сегодняшний день существует большое количество алгоритмов тематического моделирования, и результаты их применения к одной и той же текстовой коллекции могут несколько отличаться. На построенные модели оказывают влияние особенности предобработки текстов, а также и сама методология проведения исследований. Поэтому полученные частотные топика, описанные в данной статье, не могут рассматриваться как единственно возможная и, тем более, оптимальная схема. Выявление оптимальной схемы работы алгоритмов тематического моделирования на художественных текстах — задача

будущих исследований, весьма важная для решения многих гуманитарных задач, связанных с *distant reading* [23].

Проведенный эксперимент стоит считать пилотным, а его результаты — предварительными. Тем не менее сравнение выявленных тем по периодам отчасти согласуется с информацией о тематике русских рассказов, полученных при экспертной литературоведческой оценке для тех же временных срезов [37]. Согласно этим исследованиям, в частности, в довоенный период по сравнению с последующими максимально преобладают семейные ценности, романтическая любовь, интерес к финансовому благополучию. Эти темы хорошо видны на частотных топиках построенной модели. Выявленные экспертом особенности малой прозы раннесоветского периода, в частности — высокая доля рассказов о жизни на селе, а также интерес к науке и техническому прогрессу, — также нашли свое отражение в полученных частотных топиках.

Однако у построенных моделей есть и недостатки. В первую очередь, таковыми кажутся повторения темообразующих слов в нескольких частотных темах (*спросить, дом, комната* и др. для 1-го периода, *пойти, жизнь, спросить* и др. для 2-го периода, *лицо, жизнь, земля* и др. для 3-го периода). С одной стороны, эти слова можно считать «ключевыми» для прозы соответствующих эпох, но с другой стороны, такие повторы «размывают» содержательные отличия между топиками, что снижает общее качество модели. Другим недостатком моделей можно считать «потерю» военно-революционной тематики, ожидаемо частотной для рассказов 2-го периода [38], а также темы насилия, довольно частотной для русской прозы изучаемого периода [6]. Объяснения этому явлению могут быть разные. Во-первых, возможно, указанные топики просто не попали в ограниченный список частотных, рассмотренных в данной статье. Во-вторых, это может быть связано с тем, что при построении моделей была отфильтрована частотная лексика, встречающаяся более чем в 80% текстов для каждой выборки. Тем самым могло быть утрачено «военное» своеобразие второго периода. Отсюда можно сделать вывод, что при анализе художественных текстов, в отличие от специальных текстов, нужно очень осторожно подходить к «отбрасываемой» лексике и имеет смысл повторить расчеты для более мягких условий построения модели. В-третьих, возможно, что статистика, полученная в предыдущих исследованиях [37; 44] на материале 310 текстов не является достаточно представительной, а как раз новые данные, выполненные на выборке в 3000 рассказов, точнее показывают реальную дистрибуцию тем.

Кроме того, стоит отметить, что без первичного фильтра в 25% рассказов от общего объема выборки количество частотных тем могло быть существенно выше и расширило бы перечень общих частотных тем. Будем считать это задачей уже следующего этапа исследования. Поскольку рассмотрены результаты тематической модели, построенной без применения частеречных фильтров, стоит попробовать другие алгоритмы и методы — например, рассмотреть только существительные. Кажется целесообразным сравнить полученные результаты с частотными словарями художественной прозы и со списками ключевых слов [32; 40], а также провести эксперименты, предназначенные для оценки интерпретируемости полученных тематических моделей.

Благодарности

Публикация подготовлена в результате проведения исследования по проекту № 21-04-053 «Методы искусственного интеллекта для филологических исследований» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)».

References

- [1] Blei D. M. Probabilistic topic models, *Communications of the ACM*. — 2012. — Vol. 55(4) — pp. 77–84.
- [2] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation, *The Journal of machine Learning research*. — 2003. — Vol. 3. — pp. 993–1022.
- [3] Da N. Z. The computational case against computational literary studies, *Critical Inquiry*. — 2019. — Vol. 45(3). — pp. 601–639.
- [4] Glass V, Stanley J. *Statistical Methods in Education and Psychology*. — Englewood Cliffs, NJ: Prentice-Hall, 1972.

- [5] Greene D., Cross J.P. Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis, Proceedings of the ACM Web Science Conference (WebSci'15), Oxford, UK. — 2015.
- [6] Gryaznova E., Kirina M. Defining Kinds of Violence: A Comparison of Topic Modelling with Latent Dirichlet Allocation and Principal Component Analysis for Russian Short Stories of 1900–1930, 2021 International Conference “Internet and Modern Society”, IMS 2021, CEUR Workshop Proceedings, 2021, Vol. 3090, pp. 281–290.
- [7] Honnibal M., Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing — 2017.
- [8] Jockers M. L., Mimmo D. Significant themes in 19th-century literature, *Poetics*. — 2013. — Vol. 41(6). — pp. 750–769.
- [9] Lib.ru: "Classics" (Maxim Moshkov's Library), Web: <http://az.lib.ru>.
- [10] Library CoolLib. 2012 – 2022, Web: <https://coollib.net/>
- [11] Library LitMir, Web: <https://www.litmir.me/>
- [12] Library RoyalLib.Com, 2010-2022, Web: <https://royallib.com/>
- [13] Corpus of Russian Short Stories of 1913-1930s: <https://russian-short-stories.ru/>
- [14] Martynenko G. Stylized syntactic triads in Russian short story of the first third of the 20th century [Stilizovannyye sintaksicheskiye triady v russkom rasskaze pervoy treti XX veka], Proceedings of the Int. Conf. ‘Corpus Linguistics – 2019’, St. Petersburg State University, St. Petersburg — 2019. — pp. 395–404.
- [15] Martynenko G., Sherstinova T. Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture, Digital Transformation and Global Society. DTGS 2018. *Communications in Computer and Information Science*, Springer, Switzerland. — Vol. 859 — 2018. — pp. 299–309.
- [16] Martynenko G.Ya., Sherstinova T.Yu., Melnik A.G., Popova T.I. Methodological problems of creating a Computer Anthology of the Russian story as a language resource for the study of the language and style of Russian artistic prose in the era revolutionary changes (first third of the 20th century), Proc. of the XXI Int. United Conference ‘The Internet and Modern Society’, IMS-2018, Computational linguistics and computational ontologies. ITMO University, St. Petersburg — 2018. — Iss. 2 — pp. 99–104.
- [17] Martynenko G.Ya., Sherstinova T.Yu., Popova T.I., Melnik A.G., Zamirajlova E.V. On the Principles of Creation of the Russian Short Stories Corpus of the First Third of the XX Century [O printsipakh sozdaniya korpusa russkogo rasskaza pervoy treti XX veka] // Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics ‘TEL 2018’. Kazan Federal University, Kazan. — 2018. — pp. 180–197.
- [18] McFarland D. A. et al. Differentiating language usage through topic models // *Poetics*. — Vol. 41(6). — 2013. — pp. 607–625.
- [19] Mitrofanova O. Topic modeling of special texts based on LDA algorithm [Modelirovaniye tematiki special'nyh tekstov na osnove algoritma LDA] // XLII International philological conference [XLII Mezhdunarodnaya filologicheskaya konferenciya]. — 2014.
- [20] Mitrofanova O.A. Analysis of Fiction Text Structure by means of Topic Modelling: Case Study of “Master and Margarita” Novel by M. A. Bulgakov [Issledovanie strukturnoj organizacii hudozhestvennogo proizvedeniya s pomoshh'ju tematicheskogo modelirovaniya (opyt raboty s tekstom romana «Master i Margarita» M.A. Bulgakova)], *Korpusnaya lingvistika-2019*. — 2019. — pp. 387–394.
- [21] Mitrofanova O.A. Topic modelling of the Corpus of ‘Russian folk tales by A. N. Afanasiev’, *Structural and applied linguistics [Strukturnaya i prikladnaya lingvistika]*. — Vol. 11. — 2015. — pp. 146–154.
- [22] Mitrofanova O.A., Sedova A.G. Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose) // *Information Technology and Computational Linguistics (ITCL 2017)*, ICPS Proceedings, IMS2017: Proceedings of the International Conference IMS-2017 — 2017. — pp. 175–180.
- [23] Moretti F., *Distant Reading*, London: Verso, 2013.
- [24] Navarro-Colorado B. On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry // *Frontiers in Digital Humanities*. — 2018. — Vol. 5.
- [25] Nikolenko S., Koltcov S., Koltsova O. Topic modelling for qualitative studies, *Journal of Information Science*. — Vol. 43(1). — 2017. — pp. 88–102.
- [26] Panicheva P., Litvinova O., Litvinova T. Author Clustering with and Without Topical Features, *Speech and Computer*, Proceedings of the 21st Int. Conf., SPECOM 2019, LNAI 11658, Springer, Cham — 2019. — pp. 348–358.
- [27] Rehurek R., Sojka P. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. — Vol. 3(2). — 2011.
- [28] Rhody L. M. Topic Modelling and Figurative Language, *Journal of Digital Humanities*. — 2012.
- [29] Röder M., Both A., and Hinneburg A.: Exploring the Space of Topic Coherence Measures, In Proceedings of the eighth International Conference on Web Search and Data Mining. — 2015.

- [30] Ruthenia.ru <https://www.ruthenia.ru/>
- [31] Schöch C. Topic modeling genre: an exploration of french classical and enlightenment drama, arXiv preprint arXiv:2103.13019. — 2021.
- [32] Sherstinova T., Grebennikov A., Skrebtsova T., Guseva A., Gukasian M., Egoshina I., Turygina, M. Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900-1930), 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy. — 2020. — pp. 366–373.
- [33] Sherstinova T., Kirina M. Normalization Issues in Digital Literary Studies: Spelling, Literary Themes and Biographical Description of Writers, 6th International Conference on Digital Transformation and Global Society, DTGS 2021, *Communications in Computer and Information Science (CCIS)*. — Vol. 1503, — pp. 332–346.
- [34] Sherstinova T., Martynenko G. Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century, R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019), St. Petersburg, Russia, CEUR Workshop Proceedings — Vol. 2552. — 2020. — pp. 105–120.
- [35] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction, *Advances in Computational Intelligence, MICAI 2020, Lecture Notes in Computer Science*. — Vol. 12469. — 2020. — pp. 134–151.
- [36] Sherstinova T., Moskvina A., Kirina M. Towards Automatic Modelling of Thematic Domains of a National Literature: Technical Issues in the Case of Russian, 29th Conference of Open Innovations Association FRUCT, FRUCT 2021. — pp. 313-323.
- [37] Sherstinova T., Skrebtsova T., Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900-1930, Proceedings of the International Conference "Internet and Modern Society" IMS-2020. CEUR Workshop Proceedings. — pp. 117-128.
- [38] Skrebtsova T. G. Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories, Proceedings of the International Conference "Internet and Modern Society" IMS-2020. CEUR Workshop Proceedings. — Vol. 2813. — 2020. — pp. 265-276.
- [39] Skrebtsova T. Narrative structure of the Russian short story in the early XX century [Struktura narrativa v russkom rasskaze nachala XX veka], Proc. of the Int. Conf. Corpus Linguistics-2019, St. Petersburg. — 2019. — pp. 426–431.
- [40] Skrebtsova T., Grebennikov A., Sherstinova T. The Dynamics of Vocabulary in Russian Prose (Based on Frequency Dictionaries of the Corpus of Russian Short Stories 1900-1930), 21st Annual International Conference on Computational Linguistics and Intellectual Technologies, Dialogue 2021 [Komp'juternaja Lingvistika i Intellektual'nye Tehnologii]. — 2021. — pp. 646-659.
- [41] Tynyanov Yu. Archaists and Innovators [Arkhaisty i novatory]. Priboi Publ., Leningrad, 1929.
- [42] Uglanova I., Gius E. The Order of Things. A Study on Topic Modelling of Literary Texts, Proc. of the CHR 2020: Workshop on Computational Humanities Research, CEUR Workshop Proceedings. — 2020. — Access mode: <http://ceur-ws.org/Vol-2723/long7.pdf>.
- [43] Wikisource: <https://ru.wikisource.org/>
- [44] Zamiraylova E., Mitrofanova O. Dynamic topic modelling of Russian fiction prose of the first third of the XXth century by means of non-negative matrix factorization, R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019), *CEUR Workshop Proceedings*. — Vol. 2552. — 2020. — pp. 321–339.
- [45] Korshunov A., Gomzin A. Tematicheskoye modelirovaniye tekstov na yestestvennom yazyke [Thematic modeling of natural language texts]. In: Trudy Instituta sistemnogo programmirovaniya RAN [Proceedings of the Institute for System Programming RAS], 2012, no. 23, https://ispranproceedings.elpub.ru/jour/article/view/982?locale=ru_RU.
- [46] Nagibin Yu. M. Antologiya russkogo sovetского rasskaza. Predisloviye [Anthology of Russian Soviet short stories. Foreword], Bibliotekha «Knizhnoye obozreniye». Moscow: Knizhnoye obozreniye, 1987.

Приложение

Выборка	Тема	Кол-во текстов		Содержание топики
		P>75%	P>95%	
1-100	17	29%	17%	('голос', 0.0040), ('стоять', 0.0034), ('спросить', 0.0034), ('минута', 0.0030), ('душа', 0.0030), ('комната', 0.0028), ('женщина', 0.0028), ('молодой', 0.0027), ('дом', 0.0027), ('земля', 0.0027)
	18	9%	5%	('отец', 0.0095), ('пойти', 0.0069), ('старик', 0.0065), ('мать', 0.0057), ('деньга', 0.0046), ('прийти', 0.0041), ('земля', 0.0039), ('голос', 0.0039), ('народ', 0.0038), ('спросить', 0.0037)
1-500	2	13,8%	5,6%	('смотреть', 0.0037), ('любить', 0.0035), ('окно', 0.0031), ('душа', 0.0028), ('сидеть', 0.0027), ('комната', 0.0027), ('голос', 0.0026), ('белый', 0.0025), ('ночь', 0.0025), ('тёмный', 0.0025)
	3	13,4%	6%	('пойти', 0.0041), ('отец', 0.0037), ('дом', 0.0033), ('смотреть', 0.0032), ('спросить', 0.0028), ('земля', 0.0027), ('стоять', 0.0027), ('мужик', 0.0026), ('сидеть', 0.0025), ('прийти', 0.0024),
	22	7,4%	3%	('жена', 0.0033), ('спросить', 0.0028), ('комната', 0.0026), ('голос', 0.0025), ('дом', 0.0024), ('женщина', 0.0024), ('сидеть', 0.0023), ('час', 0.0023), ('нужный', 0.0022), ('минута', 0.0022)
1-1000	6	23%	10%	('земля', 0.0035), ('пойти', 0.0031), ('отец', 0.0028), ('стоять', 0.0026), ('старик', 0.0026), ('старый', 0.0026), ('дом', 0.0024), ('ночь', 0.0024), ('сидеть', 0.0023), ('лес', 0.0023)
	5	8,4%	2,5%	('любить', 0.0030), ('жена', 0.0028), ('ребёнок', 0.0028), ('сидеть', 0.0026), ('дом', 0.0026), ('муж', 0.0025), ('спросить', 0.0024), ('комната', 0.0024), ('отец', 0.0024), ('минута', 0.0023)
	14	8%	2,3%	('любить', 0.0034), ('душа', 0.0034), ('комната', 0.0029), ('любовь', 0.0028), ('ночь', 0.0028), ('мысль', 0.0028), ('спросить', 0.0028), ('женщина', 0.0028), ('странный', 0.0027), ('сердце', 0.0026)
	3	0,9%	0,3%	('деньга', 0.0055), ('комната', 0.0046), ('жена', 0.0044), ('рубль', 0.0043), ('спросить', 0.0039), ('квартира', 0.0038), ('час', 0.0035), ('пойти', 0.0031), ('дом', 0.0031), ('служба', 0.0029)

Таблица I: Наиболее распространенные топики периода 1900–1913 гг.
(для объемов выборки в 100, 500 и 1000 рассказов)

Выборка	Тема	Кол-во текстов		Содержание топика
		P>75%	P>95%	
2-100	16	10%	9%	('женщина', 0.0067), ('пойти', 0.006323), ('сидеть', 0.0056), ('ночь', 0.0045), ('дверь', 0.0041), ('стоять', 0.0038), ('белый', 0.0038), ('спросить', 0.0036), ('голос', 0.0034), ('окно', 0.0034)
	22	10%	6%	('любить', 0.0061), ('письмо', 0.0051), ('сидеть', 0.0041), ('душа', 0.0040), ('мысль', 0.0040), ('дом', 0.0038), ('пойти', 0.0037), ('спросить', 0.0037), ('молодой', 0.0036), ('вечер', 0.0035)
2-500	20	9%	3,60%	('жизнь', 0.0049), ('земля', 0.0033), ('ночь', 0.0032), ('солнце', 0.0030), ('пойти', 0.0029), ('белый', 0.0029), ('голос', 0.0027), ('смотреть', 0.0026), ('сердце', 0.0026), ('чёрный', 0.0024)
	24	6,60%	2%	('пойти', 0.0030), ('дверь', 0.0029), ('смотреть', 0.0027), ('сидеть', 0.0027), ('спросить', 0.0027), ('комната', 0.0027), ('жизнь', 0.0027), ('ночь', 0.0026), ('голова', 0.0025), ('дом', 0.0025)
	25	4,60%	2,20%	('жизнь', 0.0046), ('душа', 0.0036), ('женщина', 0.0032), ('мысль', 0.0031), ('любить', 0.0029), ('сидеть', 0.0029), ('спросить', 0.0027), ('минута', 0.0027), ('пойти', 0.0025), ('прийти', 0.0024)
	29	3,60%	2,60%	('жизнь', 0.0044), ('душа', 0.0036), ('город', 0.0036), ('жена', 0.0034), ('смотреть', 0.0033), ('дом', 0.0033), ('толпа', 0.0031), ('спросить', 0.0028), ('улица', 0.0026), ('отец', 0.0026)
2-1000	0	17%	6,5%	('жизнь', 0.0041), ('спросить', 0.0028), ('женщина', 0.0027), ('комната', 0.0027), ('любить', 0.0026), ('дом', 0.0025), ('смотреть', 0.0024), ('сидеть', 0.0024), ('нужный', 0.0024), ('пойти', 0.0021)
	8	7,7%	2,9%	('земля', 0.0053), ('ночь', 0.0036), ('белый', 0.0031), ('отец', 0.0029), ('душа', 0.0027), ('пойти', 0.0027), ('лес', 0.0027), ('старый', 0.0027), ('чёрный', 0.0026), ('нёбо', 0.0026)
	1	2%	0,9%	('вода', 0.0041), ('чёрный', 0.0039), ('ночь', 0.0037), ('белый', 0.0036), ('лес', 0.0032), ('окно', 0.0028), ('пойти', 0.0027), ('смотреть', 0.0027), ('нёбо', 0.0026), ('стоять', 0.0025)
	4	2,4%	0,6%	('пойти', 0.0058), ('отец', 0.0046), ('дом', 0.0041), ('смотреть', 0.0041), ('сидеть', 0.0040), ('мать', 0.0033), ('спросить', 0.0031), ('мужик', 0.0031), ('окно', 0.0030), ('жизнь', 0.0029)
	3	1,2%	0,5%	('капитан', 0.0042), ('жизнь', 0.0034), ('смерть', 0.0031), ('минута', 0.0027), ('последний', 0.0025), ('мысль', 0.0022), ('солдат', 0.0022), ('стоять', 0.0020), ('князь', 0.0020), ('сторона', 0.0020)
	10	1,1%	0,3%	('жизнь', 0.0128), ('душа', 0.0093), ('любовь', 0.0082), ('женщина', 0.0077), ('любить', 0.0056), ('сердце', 0.0044), ('счастье', 0.0041), ('прекрасный', 0.0038), ('письмо', 0.0032), ('девушка', 0.0033)

Таблица II: Наиболее распространенные топики периода 1914–1922
(для объемов выборки в 100, 500 и 1000 рассказов)

Выборка	Тема	Кол-во текстов		Содержание топики
		P>75%	P>95%	
3-100	9	25%	11%	('лицо', 0.0052), ('земля', 0.0049), ('белый', 0.0046), ('старик', 0.0038), ('вода', 0.0037), ('ночь', 0.0036), ('жизнь', 0.0034), ('мать', 0.0034), ('сидеть', 0.0034), ('чёрный', 0.0034)
	12	13%	8%	('отец', 0.0078), ('голос', 0.0062), ('тётка', 0.0057), ('спросить', 0.0056), ('работа', 0.0054), ('товарищ', 0.0051), ('старик', 0.0050), ('лицо', 0.0049), ('комната', 0.0045), ('жизнь', 0.0044)
3-500	36	11%	4,2%	('пойти', 0.0041), ('лицо', 0.0038), ('голова', 0.0032), ('земля', 0.0030), ('вода', 0.0029), ('лес', 0.0029), ('ночь', 0.0027), ('старик', 0.0027), ('мужик', 0.0027), ('белый', 0.0026)
	0	7,2%	2,4%	('лицо', 0.0043), ('дом', 0.0039), ('земля', 0.0037), ('жизнь', 0.0033), ('пойти', 0.0032), ('старик', 0.0030), ('ночь', 0.0029), ('отец', 0.0029), ('сидеть', 0.0028), ('смотреть', 0.0028)
	40	3,6%	1,6%	('жизнь', 0.0037), ('комната', 0.0035), ('белый', 0.0028), ('книга', 0.0027), ('лицо', 0.0027), ('нужный', 0.0025), ('женщина', 0.0025), ('доктор', 0.0024), ('дверь', 0.0023), ('последний', 0.0023)
	34	3%	2%	('жизнь', 0.0034), ('земля', 0.0033), ('вода', 0.0032), ('лицо', 0.0030), ('город', 0.0030), ('чёрный', 0.0027), ('море', 0.0026), ('ветер', 0.0025), ('белый', 0.0025), ('ночь', 0.0025)
3-1000	20	11,5%	2,5%	('глаз', 0.0068), ('пойти', 0.0042), ('лицо', 0.0039), ('сидеть', 0.0031), ('мужик', 0.0029), ('спросить', 0.0028), ('голос', 0.0028), ('голова', 0.0026), ('товарищ', 0.0026), ('земля', 0.0026)
	26	3,1%	1%	('товарищ', 0.0061), ('рабочий', 0.0038), ('город', 0.0029), ('председатель', 0.0029), ('глаз', 0.0028), ('лицо', 0.0027), ('ответить', 0.0025), ('голос', 0.0024), ('спросить', 0.0022), ('гражданин', 0.0022)
	10	2,9%	0,8%	('глаз', 0.0054), ('вода', 0.0052), ('лес', 0.0047), ('ночь', 0.0041), ('пойти', 0.0032), ('земля', 0.0032), ('лицо', 0.0030), ('белый', 0.0030), ('голова', 0.0029), ('снег', 0.0028)
	29	1,7%	0,1%	('любить', 0.0059), ('жена', 0.0059), ('жизнь', 0.0057), ('лицо', 0.0048), ('женщина', 0.0047), ('глаз', 0.0046), ('любовь', 0.0043), ('смотреть', 0.0041), ('нужный', 0.0040), ('муж', 0.0040)
	23	1,4%	0,2%	('жизнь', 0.0047), ('глаз', 0.0040), ('земля', 0.0040), ('мир', 0.0036), ('профессор', 0.0030), ('тело', 0.0029), ('дом', 0.0028), ('комната', 0.0027), ('мысль', 0.0026), ('стена', 0.0024)
	2	1,2%	0,2%	('жизнь', 0.0037), ('книга', 0.0033), ('город', 0.0027), ('последний', 0.0027), ('русский', 0.0024), ('остров', 0.0023), ('нужный', 0.0020), ('вода', 0.0020), ('лицо', 0.0018), ('ответить', 0.0018)

Таблица III: Наиболее распространенные топики периода 1923–1930
(для объемов выборки в 100, 500 и 1000 рассказов)

The Delimitative in Russian: Linear Ordering or a Path towards the Culmination?

Tatevosov S. G.

Lomonosov Moscow State University,
Interdisciplinary Scientific and
Educational 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

In this paper, we explore the distribution of the delimitative in Russian in combination with eventuality descriptions that denote culminating processes. With H.R. Mehlig, we assume that homogenization of a process is a *sine qua non* condition on the delimitative. We argue for a modal treatment of homogeneity whereby a process is (locally) homogeneous at an interval if the cardinality of the set of metaphysically accessible non-culminating worlds does not change significantly at that interval.

Keywords: the delimitative, culmination, modality

DOI: 10.28995/2075-7182-2022-21-527-537

Русский делимитатив: линейный порядок или движение к кульминации?

Татевосов С. Г.

МГУ им. 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 делимитативов от глаголов, описывающих кульминирующие процессы. Опираясь на идею Х. Р. Мелига о гомогенизации таких процессов как необходимым условием образования делимитатива, мы предлагаем модальную интерпретацию понятия гомогенности. Гомогенными оказываются такие процессы, при осуществлении которых их позиция на шкале, отражающей объем множества метафизически доступных некульминирующих миров, существенно не меняется.

Ключевые слова: делимитатив, кульминация, модальность

1 Делимитатив и его дистрибуция

Делимитативы и пердуративы, иллюстрируемые в (1a-b), занимают особое место в русской аспектуальной системе. Если бы они внезапно исчезли, система подчинилась бы обобщению (2):

- (1) a. Франс ещё немного **поразрушал** свою комнату, а потом позвонил [ficbook.net]
b. Я бы этой шестерке и руки бы не подал. "Настоящий офицер", который полжизни **проткрывал** перед "хозяином". [newslab.ru]

- (2) Обобщение о перфективности
Глаголы совершенного вида предельны.

Согласно (2), любой перфективный глагол, независимо от его морфологического состава, создает предельные глагольные группы и предложения. Обычный способ проверить предельность — сочетаемость с обстоятельствами типа *за два часа*.

- (3) *Володя решил задачу / прочитал книгу / дозаписал обращение к народу за два часа.*

(3) показывает, что глаголы *решить*, *прочитать*, *дозаписать*, как и бесчисленное множество других глаголов совершенного вида, предельны. При выполнении (2) предельность — **необходимое** условие перфективности:

- (4) Совершенный вид → предельный предикат

И обратно: если бы не делимитативы и пердуративы, аспектуальные свойства русской глагольной системы полностью предсказывались бы предельностью/непредельностью глагольного предиката. Предельность была бы также **достаточным** условием перфективности:

- (5) Предельный предикат → совершенный вид

Делимитативы и пердуративы, однако, будучи перфективными, непредельны. В этом качестве они сочетаются с обстоятельствам типа *два часа*.

- (6) а. *Володя **посидел** в кресле десять минут.*
б. *Двигатель **проработал** холостую полчаса.*

Делимитативы и пердуративы тем самым обеспечивают возможность соединить семантическую перфективность не только с предельными предикатами, но и с непредельными. Как пишет по этому поводу С. Дики (Dickey 2005), “*po-* delimitatives perform a crucial systemic function in the Russian aspectual system — the extension of the aspect opposition to atelic activity predicates... Without *po-* delimitatives, the Russian aspect opposition would be restricted to telic predicates (accomplishments and achievements) and thus be a much more lexical category.”

Далее мы оставим в стороне пердуратив и сосредоточимся на делимитативе. Нас будет интересовать делимитатив в той интерпретации, когда он описывает **единичное событие** и не предполагает итерации. Последнее свойство важно контролировать, поскольку при итеративном и не-итеративном прочтениях делимитатив имеет разную дистрибуцию.

Например, (7) допустим только при итеративной интерпретации (и к тому же нуждается в узкой сфере действия именной группы *таблетка аспирина*):

- (7) ^{2/OK} *Володя **попринимал** таблетку аспирина (каждый день в течение месяца, и ему стало лучше).*

Прототипический делимитатив — это комбинация префикса *по-* и глагольной основы, которая описывает некульминирующий процесс или состояние стадийного уровня¹. Формального определения (не)кульминирующих процессов мы не даем ввиду его громоздкости; в Tatevosov 2022 оно опирается на теорию имперфективности, предложенную в Varasdi 2014. С эвристической точки зрения отнести основу к этому классу можно на основании простого критерия: отсутствия среди деривационно связанных с ней лексем такой, которая описывает кульминацию того же самого процесса. В терминах Tatevosov 2016 — это основы, находящиеся вне предельных

¹ В этой статье мы не затрагиваем проблему того, почему делимитативы не образуются от еще одного подкласса непредельных предикатов — предикатов состояния индивидуального уровня (**Володя познал английский язык / полюбил Машу* ‘побыл в состоянии влюбленности’) / *повесил 87 килограммов*).

акциональных групп. Несколько примеров показаны в Таблице 1; основы, описывающие некульминирующие процессы, выделены затенением.

Основа, описывающая процесс	<i>чита-</i>	<i>открыва-</i>	<i>жари-</i>	<i>сиде-</i>	<i>работа-</i>	<i>бега-</i>
Основа, описывающая его кульминацию	<i>написа-</i>	<i>откры-</i>	<i>пожари-, поджари-</i>	—	—	—

Таблица 1. Кульминирующие и некульминирующие процессы

Делимитативы от основ, описывающих некульминирующие процессы и состояния стадийного уровня, образуются практически без ограничений: *посидеть, поработать, побегать* и т. д.²

Основная проблема, на решение которой нацелена эта статья, связана с тем, что делимитативы образуются также от основ, описывающих кульминирующие процессы, однако не от всех. В работах Х. Р. Мелига (Mehlig 2006, 2012, *i.a.*), а также в наших более ранних работах (Kisseleva, Tatevosov 2011, Tatevosov 2020) обсуждаются примеры типа *повязать свитер, поучить урок, полечить больного, покопать яму, пооткрывать дверь*. Корпусной пример, иллюстрирующий последний случай, показан в (8):

- (8) *Мужик в красной толстовке... вышел, даже сигарету не выбросил, продолжал курить, пооткрывал дверь — не получилось, ну и ладно, пошел дальше.* [74.ru]

В противоположность этим случаям предикаты типа *выдавать книгу* (в библиотеке), *запивать таблетку, расстреливать пленного, надевать рубашку* или *сообщать новость* делимитатив, по-видимому, не образуют.

- (9) а. ??*Володя **повыдавал** книгу читателю.*³
 б. ??*Володя **позапивал** таблетку от головной боли.*
 с. ??*Володя **порасстреливал** пленного белогвардейца.*
 с. ??*Володя **посообщал** новость Феликсу.*

Ограничение на образование делимитатива у глаголов в (9) и аналогичных примерах является **лексическим** и не зависит от лингвистического и экстралингвистического контекста. В (Kisseleva, Tatevosov 2011, Tatevosov 2020) выделяются два систематических случая, когда приемлемость делимитатива варьирует в зависимости от **экстралингвистического сценария**. (10)-(11) иллюстрируют первый такой случай:

- (10) Сценарий 1. Замок в двери сломан, а Володя пытается попасть внутрь. Сначала он использует обычный ключ, потом отмычку, потом пытается воздействовать на дверь ломом, но ничего не помогает. На самом деле дверь открывается ударом кувалды, но Володя не знает об этом и прекращает свои попытки без надежды на успех.
Володя пооткрывал дверь (полчаса и бросил это занятие).
- (11) Сценарий 2. Дверь открывается введением кода, состоящего из последовательности цифр 7-9-6-8-1. Набрав первые четыре цифры, Володя останавливается.
 * *Володя пооткрывал дверь (полминуты и передумал)*

² Мы исходим из того, что образование делимитатива от имперфективного глагола сохраняет акциональные свойства: если исходный глагол описывает состояние стадийного уровня, это же верно и для делимитатива; имперфективы, которые представляют собой дескрипцию динамического процесса, создают процессный делимитатив. Мы признательны рецензенту «Диалога», рекомендовавшему внести в текст это уточнение.

³ Представленные здесь и далее суждения о приемлемости отражают интуицию авторов. Разделяя до известной степени беспокойство одного из анонимных рецензентов «Диалога» об интроспекции как источнике семантических суждений, мы хотели бы отметить, что корпус не может быть источником отрицательного материала, а проведение полномасштабного экспериментального исследования по этому вопросу выходит далеко за рамки скромных задач нашего исследования.

Как видно из (10)-(11), делимитативом *пооткрывать* описывается деятельность, нацеленная на открытие двери, которая останавливается на попытке использовать лом, но не деятельность, прерываемая после нажатия кнопок 7-9-6-8.

Ограничение второго типа иллюстрируется в (12):

- (12) *Володя немного почитал* ^{OK}роман / ^{OK}статью / ^{OK?}письмо / [?]записку / ^{???}абзац / ^{??}предложение / *слово / *символ.

Приемлемость глагольной группы вида «*почитать* DP» убывает с уменьшением длины объекта, выступающего как пациентивный актанта. Предельный случай — *почитать символ*, когда делимитатив делается полностью аномален. Этот паттерн можно без труда воспроизвести с другими глаголами с инкрементальной темой, или с накопителем эффекта — *писать, пить, петь, пахать* и т.д.

Эти наблюдения подводят нас к двум обобщениям в (13):

- (13) Обобщения о дистрибуции делимитатива
1. Делимитатив образуется не только от глаголов, описывающих некульминирующие процессы, но и от глаголов, предполагающих достижение кульминации
 2. Во втором случае, в отличие от первого, действуют по меньшей мере три типа ограничений:
 - (i) лексические, как в (9), когда глагол вовсе не способен к образованию делимитатива;
 - (ii) сценарные, как в (10)-(11), когда делимитатив совместим лишь с определенными экстралингвистическими сценариями
 - (iii) актантно-обусловленные, как в (12), когда делимитатив предъявляет требования к семантическим характеристикам внутреннего аргумента.

В следующем разделе мы обсудим возможные объяснения обобщений в (13).

2 Гомогенность

Не имея возможности дать обзор всех имеющихся работ по делимитативу, мы ограничимся обсуждением одной ключевой идеи, сформулированной Х. Р. Мелигом (Mehlig 2006, 2012, i.a.). С точки зрения Х. Р. Мелига, основной семантический ингредиент делимитативности — **гомогенность** описываемой ситуации. Делимитатив вводит в рассмотрение «ограниченные в видовом плане гомогенные ситуации, начальное и конечное состояния которых являются идентичными и которые поэтому без дальнейших уточнений в контексте не могут продвинуть время повествования вперед» (Mehlig 2006: 248) и «допускается только тогда, когда глагольная предикация обозначает гомогенную ситуацию, т. е. ситуацию, для которой верен принцип произвольной делимости на идентичные фазы» (Mehlig 2006: 258).

Мы полностью согласны с описанием семантики делимитатива, которое дает Х. Р. Мелиг. Однако, чтобы в полной мере воспользоваться его преимуществами, в частности, объяснить с его помощью обобщения в (13), описание желательно сделать более эксплицитным. В первую очередь это относится к понятию гомогенности, к определению которого надо подойти очень тщательно, чтобы под него подпадали ровно те случаи, которые описываются в (13).

Здесь мы сталкиваемся с несколькими затруднениями.

Если гомогенность определить через «идентичные начальное и конечное состояния», возникает вопрос, применимо ли это к случаям типа *почитать роман, попахать поле* или *позаполнять анкету*. Инкрементальные глаголы обладают свойством отображения в подобъекты: чем дальше разворачивается событие, тем большая часть объекта в нем задействуется (в конце чтения прочитана большая часть романа, чем в начале, вспахана большая часть поля и т. д.). Следовательно, начальное и конечное состояния у таких делимитативов нельзя назвать в строгом смысле идентичными.

Если для гомогенности требуется «произвольная делимость на идентичные фазы», мы сталкиваемся с проблемой в случаях типа (10). Из этого примера ясно, что деятельность, описываемая как *пооткрывал дверь*, может состоять из весьма различных частей. Если гомогенность —

необходимое условие делимитатива, а компоненты делимитативной деятельности могут быть так разнородны, как в (10), гомогенность вряд ли следует определять через «произвольное деление на идентичные фазы».

Как ответ на эти затруднения во Tatevosov 2020 предлагается описывать гомогенность Х. Р. Мелига как **отсутствие уникального временного упорядочивания**.

- (14) Гипотеза об уникальном временном упорядочивании
 Предикаты, не допускающие образование делимитатива, отличаются тем, что подсобытия, составляющие процесс, упорядочены во времени единственно возможным образом.

Покажем, как работает эта гипотеза применительно к примерам выше. Рассмотрим (11), *пооткрывать дверь* со Сценарием 2.

В этом случае *e*, деятельность агенса, направленная к кульминации, складывается из пяти подсобытий, которые должны быть расположены во времени единственным образом:

- (15) a. e_7 : ввод цифры 7
 e_9 : ввод цифры 9
 e_6 : ввод цифры 6
 e_8 : ввод цифры 8
 e_1 : ввод цифры 1
 b. $e = e_7 \oplus e_9 \oplus e_6 \oplus e_8 \oplus e_1$
 c. $e_7 \llcorner e_9 \llcorner e_6 \llcorner e_8 \llcorner e_1$
 где “ \oplus ” — мереологическая сумма, а “ \llcorner ” — отношение временного предшествования

Если подсобытия происходят в любом другом порядке, пропускаются или повторяются, их сумма более не является деятельностью, открывающей дверь. Иными словами, в мире, в котором *e* из (15b) — это процесс, кульминирующий открытием двери, любой процесс из экстенционала *открывать* идентичен *e* с точки зрения состава и временного порядка подсобытий.

Аналогичное наблюдение можно сделать по поводу кульминирующих процессов из (9), где последовательность подсобытий, ведущих к кульминации, лексически упорядочена (см. также Mehlig 2006).

Рассмотрим *пооткрывать дверь* со сценарием 1 («сломанный замок») в (10). Релевантные подсобытия процесса открывания выглядят как (16).

- (16) $e_{\text{ключ}}$ = использование ключа
 $e_{\text{отмычка}}$ = использование отмычки
 $e_{\text{лом}}$ = использование лома
 $e_{\text{кувалда}}$ = использование кувалды

При таком сценарии уникального временного упорядочивания подсобытий не предполагается. Единственное, что существенно, — это чтобы конечным подсобытием было $e_{\text{кувалда}}$; все прочие могут происходить в любом порядке, повторяться сколько угодно раз или вовсе отсутствовать. Открытие двери от них каузально не зависит. В экстенционал открывать дверь в такой ситуации входят и $e = e_{\text{кувалда}}$ и, например, $e = e_{\text{лом}} \oplus e_{\text{отмычка}} \oplus e_{\text{лом}} \oplus e_{\text{лом}} \oplus e_{\text{ключ}} \oplus e_{\text{кувалда}}$. Как и предсказывает гипотеза (14), сценарий в (10), не предполагающий, в отличие от (11), уникального порядка событий, совместим с делимитативом.

Рассмотрим инкрементальные предикаты типа *читать роман* или *пахать поле*, когда кульминация определяется полной задействованностью пациенса в ситуации. Деятельность, описываемая как *пахать поле*, кульминирует в тот момент, когда ею охвачено все поле целиком; в этот момент возникает состояние ‘поле вспахано’. Однако последовательность вспахивания может быть любой, и временной порядок образующих деятельность подсобытий (‘вспахать часть поля *a*’, ‘вспахать часть поля *b*’ и т.д.) неважен. Процесс чтения романа до какой-то степени упорядочен структурой романа, однако и здесь возможны отступления от линейной

последовательности глав и разделов, возвращение к уже прочитанному и т.д. Существенно то, что чем короче объект чтения, тем хуже, во-первых, вычлняются самостоятельные контекстно-релевантные подсобытия, соответствующие его частям, а во-вторых, тем труднее встроить эти подсобытия в последовательность неединственным способом. Соответственно, все менее и менее приемлемым становится делимитатив.

В (12) *почитать записку* — это, по-видимому, граничный случай. Трудно вообразить контекст, в котором, во-первых, чтение записки распадается на отдельные автономные подсобытия, а во-вторых, последовательность прочтения частей записки задается неединственным образом. В еще большей степени это относится к абзацу, предложению, слову и тем более символу. Последний, очевидно, атомарен, его прочтение содержит единственное подсобытие, и уникальная временная упорядоченность возникает тривиальным образом.

Таково в общих чертах объяснение ограничений на дистрибуцию делимитатива, опирающееся на идею уникальной временной упорядоченности подсобытий: там, где есть такая упорядоченность, нет места для делимитатива. Гомогенность в таком случае — это отсутствие уникальной упорядоченности.

3 Кульминация и ее приближение

Описанная выше темпоральная теория гомогенности рассматривает временную неупорядоченность подсобытий как характеризующее свойство событийных дескрипций, допускающих делимитатив. Возможно, однако, и альтернативное эмпирическое обобщение, которое увязывает гомогенность не с внутренней структурой описываемого процесса, а с его кульминацией. Это обобщение в нестрогой формулировке представлено в (17):

(17) Гомогенность и кульминация (предварительный вариант)

Гомогенные процессы таковы, что при их осуществлении кульминация не приближается.

Поясним обобщение применительно к примерам из раздела 1.

Для делимитативов, образованных от дескрипций некульминирующих процессов и состояний типа *побегать*, *поспать* или *посидеть*, обобщение тривиально: кульминация не приближается, поскольку глаголы *бегать*, *спать*, *сидеть* и аналогичные вовсе ее не предполагают.

Гомогенные процессы типа *пооткрывать дверь* (со сценарием сломанного замка в (10)) также удовлетворяют (17). Любое из подсобытий в (16), кроме заключительного $e_{\text{кувалда}}$, не продвигает процесс к кульминации (открытию двери); после каждого из них кульминация так же далека от осуществления, как и до. Тем самым можно сказать, что процесс, обозначаемый предикатом *открывать дверь*, при таком сценарии глобально является кульминирующим, однако его локальная временная часть, которую описывает делимитатив, выглядит как обычная неопределяемая дескрипция.

Этим сценарий сломанного замка отличается от сценария числового кода: чем больше подсобытий из (16) осуществляется, тем ближе момент, когда дверь открывается; такой сценарий предполагает неизбежное приближение к кульминации с каждым следующим подсобытием.

Как соотносится идея (не)приближения кульминации с идеей о временной (не)упорядоченности подсобытий? По всей видимости, из второго свойства ассиметрично следует первое. Если кульминирующий процесс складывается из подсобытий с фиксированным временным порядком, как в случае со сценарием числового кода, очевидно, что каждое следующее подсобытие делает кульминацию ближе.

Однако неверно, что приближение кульминации требует уникальной временной упорядоченности. Неизбежное приближение кульминации возможно и для процессов, не предполагающих уникального временного порядка своих подсобытий. Последняя возможность показана в (18): Сценарий 3 выстроен так, что подсобытия возможны в любом порядке, но кульминация тем не менее с неизбежностью приближается.

(18) Сценарий 3. Дверь открывается введением кода, который представляет собой последовательность любых пяти цифр в любом порядке. Набрав первые четыре цифры, Володя останавливается.

*Володя *пооткрывал* дверь (*и передумал*)

Примеры типа (18) позволяют сравнить эмпирические предсказания, опирающиеся на идею уникальной упорядоченности в (14) и на идею о приближении кульминации в (17). По всей видимости, Сценарий 3 в (18) (любые пять цифр в любом порядке) так же неприемлем для делимитатива, как и Сценарий 2 в (11) (конкретные пять цифр в определенном порядке). При Сценарии 3 в контекстно ограниченный экстенционал предиката *открывать дверь* входят любые события e вида (19):

$$(19) \quad \begin{aligned} e &= e_7 \oplus e_9 \oplus e_6 \oplus e_8 \oplus e_1 \\ e &= e_1 \oplus e_2 \oplus e_3 \oplus e_4 \oplus e_5 \\ e &= e_6 \oplus e_6 \oplus e_6 \oplus e_6 \oplus e_6 \\ &\dots \end{aligned}$$

Уникальной временной упорядоченности подсобытий в (19), очевидно, не предполагается. Однако если дверь открывается пятью нажатиями любых кнопок, каждое нажатие делает кульминацию ближе. Если это так, (18) дает серьезный аргумент в пользу (17) (гомогенность как приближение к кульминации) против (14) (гомогенность как отсутствие уникальной временной упорядоченности).

В этом месте возникает вопрос: что обобщение (17) предсказывает по поводу делимитативов от инкрементальных глаголов типа *почитать книгу*, *попахать поле* или *пописать письмо* и как оно объясняет контрасты в (12)?

Без дополнительных уточнений (17), очевидно, делает неверные предсказания. Как отмечалось выше в связи с описанием гомогенности у Х. Р. Мелига, рассматриваемые предикаты инкрементальны, а кульминация задается полной вовлеченностью инкрементального актанта в событие. Для *почитать книгу*, *попахать поле* и *пописать письмо* в конце события в нем участвует большая часть объекта, чем в начале, а значит, кульминация становится ближе.

(17), соответственно, нуждается в уточнении. Новая формулировка предлагается в (20):

$$(20) \quad \begin{aligned} &\text{Гомогенность и кульминация (уточненный вариант)} \\ &\text{Гомогенные процессы таковы, что при их осуществлении значимого (с точки зрения те-} \\ &\text{кущего контекста) приближения к кульминации не происходит.} \end{aligned}$$

За уточнением в (20) стоит следующая интуиция. Если верно, что субъект почитал роман, то при осуществлении события объем прочитанного, конечно, возрос. Однако с точки зрения кульминации — прочтения всего романа — это возрастание не имеет критического значения; значимого приближения кульминации не происходит. Аналогично *попахал поле* и *пописал письмо* дают приращение вспаханного и написанного, но в то же время сигнализируют, что фактическим приближением к кульминации в текущем контексте можно пренебречь; ее вероятность в начале и конце события неразличима.

(20) делает гомогенность контекстно-зависимой. Контекстная зависимость может, например, проявляться в том, что в одних контекстах, но не в других кульминация определяется не только мереологической вовлеченностью объекта, но и целеполагающим поведением агенса. Если достижение кульминации не входит в его намерения, чтение даже значительной части романа не приближает кульминацию и состояние ‘быть прочитанным’. Агенси, не намеревающийся прочитать роман до конца, может прервать чтение в произвольной точке, и о каждой можно сказать, что вероятность кульминации в ней примерно одинакова — очень невелика. В еще большей степени это относится к *почитать газету*: чтение газеты целиком обычно не является целью читающего, соответственно, прочтение двух полос не делает более вероятным прочтение оставшихся пяти, необходимых, чтобы процесс кульминировал.

Еще одним примером контекстной зависимости мы обязаны рецензенту «Диалога», отметившему, что даже для примеров типа *немного почитал абзац* в (12), которые безусловно аномальны в большинстве воображимых контекстов, «всегда найдется носитель языка, который построит специальный контекст: ...Х читает зашифрованный или эзотерический текст или текст на языке, который ему почти непонятен». И действительно: в таком контексте делимитатив описывает процесс, который, даже занимая относительно продолжительный интервал, не ведет к существенному приближению кульминации.

С учетом сказанного ограничения в (12), когда приемлемость делимитатива зависит от «размера» инкрементального объекта, получают объяснение с опорой на (20) без дополнительных допущений. Чем короче читаемый объект и чем более линейно упорядочен в обычной ситуации процесс его чтения, тем труднее подобрать контекст, когда при осуществлении описываемой делимитативом ситуации кульминация не приближается. В этом отношении предсказания (20) не слишком отличаются от предсказаний темпоральной теории в (14) из раздела 2. Содержательно, однако, две теории предполагают принципиально разное понимание гомогенности. В одном случае гомогенность — это возможность составить процесс из подсобытий неединственным образом, в другом — отсутствие значимых для наступления кульминации изменений в мироздании.

4 Модальная теория

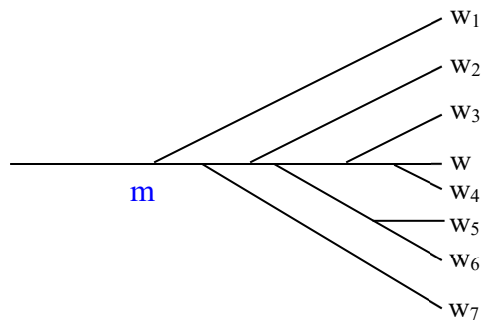
Эмпирическая успешность обобщения в (20) зависит от того, можно ли сделать эксплицитным понятие расстояния до кульминации (близости кульминации), на которое опирается обсуждение в разделе 3. Что значит, что кульминация (не) приблизилась? Цель этого раздела — предложить план ответа на этот вопрос на минимальном уровне формализации. Полный теоретико-модельный анализ содержит значительное количество технических подробностей, которые нуждаются в более подробном обосновании и обсуждении, чем позволяет объем этой статьи; он намечен в Tatevosov 2022.

Первое понятие, которое нам понадобится, — понятие метафизически доступных миров (см., например, Condoravdi 2002).

- (21) **Метафизически доступные миры**
 Для мира w и момента времени t **метафизически доступны** все миры, которые имеют с w одинаковое прошлое, т.е. идентичны ему вплоть до момента t .

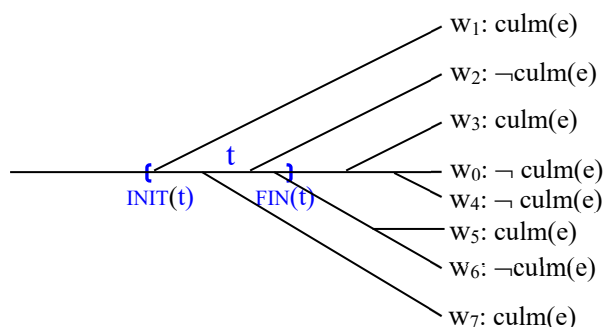
Метафизически доступные миры, иными словами, — это возможные варианты развития текущего мира, начиная с некоторого момента времени. Множество метафизически доступных миров можно изобразить в виде ветвящейся структуры в (22):

- (22) **Миры, метафизически доступные для w в t**



Пусть e — это процесс, который имеет место в мире w на интервале t . Если e — кульминирующий процесс, то среди миров, метафизически доступных для w в начале t , $INIT(t)$ есть такие, где e кульминирует, $culm(e)$ (см. Parsons 1990), и такие, где этого не происходит ($\neg culm(e)$). Назовем первые кульминирующими мирами, вторые — некульминирующими:

- (23) **Кульминирующие и некульминирующие миры**



В таком случае мерой близости кульминации может служить отношение количества некульминирующих и кульминирующих миров. Если на некотором интервале метафизически доступно значительное количество некульминирующих миров, это означает, что кульминация не близка, и у текущего мира есть много возможностей развиваться так, что ее не происходит. Движение к кульминации — это последовательное отсеечение некульминирующих «веток» в структуре типа (22), исключение таких вариантов развития событий, когда процесс или останавливается до кульминации или вовсе направляется по пути, к кульминации не ведущему. Иначе говоря, приближение кульминации — это существенное уменьшение количества метафизически доступных некульминирующих миров по мере развертывания процесса.

В начале интервала t в (23), обозначенном как $INIT(t)$, метафизически доступны восемь миров, включая текущий, из них три — некульминирующие. К концу t , обозначенному как $FIN(t)$, остаются доступны три мира (w_0, w_3, w_4), из них два некульминирующие. Два других некульминирующих мира w_2 и w_6 к концу интервала t становятся недоступными.

Если вернуться к сценарию 2 с кодовым ключом в (11), можно отметить, что каждое подсобытие в (15) элиминирует целые ветки в структуре метафизически доступных миров. Например, подсобытие e_7 , с которого начинается процесс открытия двери, исключает группы миров, в которых нажата любая другая кнопка или не нажата никакая. Все эти миры — некульминирующие. То же самое происходит при реализации каждого следующего подсобытия. В результате на протяжении процесса метафизическую доступность теряет очень значительное количество миров.

В сценарии 3 в (18) каждое подсобытие, когда нажимается произвольная кнопка, элиминирует те миры, где нажатия не происходит. С каждым нажатием некоторое количество некульминирующих миров перестает быть доступным, и кульминация, как и в предыдущем случае, приближается.

Напротив, сценарий 1 в (10) предполагает, что каждое подсобытие в (16), кроме $e_{\text{квалда}}$, не приводит к существенному уменьшению числа метафизически доступных некульминирующих миров. Например, $e_{\text{ключ}}$ элиминирует все некульминирующие миры, в которых первым подсобытием было что-то еще. Однако общее количество некульминирующих миров от этого меняется крайне незначительно — за счет того, что к отсутствию кульминации приводят любые подсобытия, кроме $e_{\text{квалда}}$, и любые их комбинации.

Если такое описание близости кульминации имеет право на существование, оно открывает возможность придать (20) следующий смысл:

- (24) Гомогенность и кульминация (уточненный более эксплицитный вариант)
Гомогенные процессы таковы, что при их осуществлении количество метафизически доступных некульминирующих миров уменьшается на пренебрежимо малую величину с точки зрения контекстного порога пренебрежимости.

Имея (24), семантику делимитатива можно определить так, как показано в (25)-(26):

- (25) Для любого событийного предиката P и события e верно следующее
а. **Пресуппозиция.** $\| \text{DELIM } P(e) \|^{w, t, C}$ определен ровно в том случае, если событие e из экстенционала предиката P обладает свойством мелиговой гомогенности, М-НОМ:
 $\| \text{М-НОМ}(P)(e) \|^{w, t, C} = 1$.
б. **Ассерция.** Если определен, $\| \text{DELIM } P(e) \|^{w, t, C} = 1$ ровно в том случае, если $P(e)=1$ в $w \wedge \tau(e) \subseteq t$,
где $\tau(e)$ — функция временного следа, отображающая событие в интервал, который оно занимает во времени
- (26) Мелигова гомогенность
 $\| \text{М-НОМ}(P)(e) \|^{w, t, C} = 1$ ровно в том случае, если $P(e)=1$ в w
 $|\{w': w' \approx_{INIT(\tau(e))} w \wedge \neg \text{culm}(e) \text{ в } w'\}| - |\{w': w' \approx_{FIN(\tau(e))} w \wedge \neg \text{culm}(e) \text{ в } w'\}| \leq \mu_C$.
где “ $w' \approx_m w$ ” читается как «мир w' метафизически доступен для w в момент m », $INIT(t)$ и $FIN(t)$ — левая и правая границы интервала t , μ_C — контекстно-зависимый порог пренебрежимости.

Согласно (25), делимитатив содержит два компонента — ассертивный и пресуппозициональный. В соответствии с ассертивным компонентом, делимитатив — это перфективный оператор в духе В. Кляйна (Klein 1994 и последующая литература): фокусное время t содержит в себе время ситуации $\tau(e)$. Пресуппозициональный компонент требует, чтобы дескрипция процесса обладала свойством мелиговой гомогенности, M-НОМ. Мелигова гомогенность определяется в (26) соотношением в (24): мощность множества метафизически доступных некульминирующих миров в начале события отличается от мощности такого же множества в конце события на пренебрежимо малую в текущем контексте величину.

5 Вместо заключения

Основной результат этой статьи состоит в том, что ограничение на дистрибуцию делимитатива можно описать через понятие гомогенности, введенное Х. Р. Мелигом, придав этому понятию следующий смысл: **гомогенными являются ситуации, при осуществлении которых не происходит приближения кульминации**. Приближение кульминации, в свою очередь, можно описать как значимое уменьшение количества метафизически доступных некульминирующих миров.

В более широком эмпирическом контексте представленный выше результат, как кажется, позволяет уточнить и дополнить наши представления о свойствах непредельных предикатов, которые существенны не только с точки зрения теоретической семантики, но и, например, при создании лексико-семантических ресурсов, в частности, событийных онтологий (Brown et al. 2017), для решения различных задач обработки естественного языка.

Acknowledgments

Авторы признательны анонимным рецензентам «Диалога» за критику и комментарии. Вклад первого автора отражает исследование, проводившееся при поддержке Российского научного фонда, грант № 22-18-00285.

References

- [1] Condoravdi Cleo. 2002. Temporal interpretation of modals. In Beaver David, Kaufmann Stefan, Clark Brady, Martinez Luis C. (eds.), *The construction of meaning*. — Stanford: CSLI Publications. — P. 59–88.
- [2] Brown Susan, Bonial Claire, Obrst Leo, Palmer Martha. 2017. The Rich Event Ontology. In *Proceedings of the Events and Stories in the News Workshop*. — Vancouver, Canada: Association for Computational Linguistics. — P. 87–97.
- [3] Dickey Stephen. 2000. *Parameters of Slavic aspect: A cognitive approach*. — Stanford: CSLI.
- [4] Dickey Stephen. 2006. Aspectual pairs, goal orientation and PO- delimitatives in Russian. — *Glossos* 7. — P. 1–32.
- [5] Dickey Stephen, Hutcheson Julie. 2003. Delimitative verbs in Russian, Czech and Slavic. In Maguire Robert, Timberlake Alan (eds.), *American contributions to the 13th International Congress of Slavists*. — Vol. 1. — Bloomington: Slavica. — P. 23–36.
- [6] Kisseleva Xenia, Tatevosov Sergei. 2011. Ordered activities and semantics of the delimitative. Paper presented at the Sixth Annual Meeting of the Slavic Linguistics Society, L'Université de Provence, September 1–3.
- [7] Klein Wolfgang. 1994. *Time in language*. — New York: Routledge.
- [8] Krifka Manfred. 1992. Thematic relations as links between nominal reference and temporal constitution. In Sag Ivan, Szabolsci Anna (eds.), *Lexical matters*. — Stanford, CA: CSLI. — P. 29–53
- [9] Mehlig Hans Robert. 2006. Verbal aspect and secondary homogenization of an eventuality: On the distribution of the Delimitative Aktionsart in Russian [Glagol'nyj vid i vtorichnaja gomogenizacija oboznachaemoj situacii: K upotrebleniju delimitativnogo sposoba dejstvija v ruskom jazyke]. In Volkmar Lehmann (ed.), *Glagol'nyj vid i leksikografija: Semantika i struktura slavjanskogo vida IV*. — Munich: Otto Sagner. — P. 235–276.
- [10] Mehlig Hans Robert. 2012. Hybrid predications in Russian. *Journal of Slavic linguistics*. — Vol. 20, no. 2. — P. 171–227.

- [11] Piñon Christopher. 1994. Aspectual composition and the ‘pofective’ in Polish. In Avrutin Sergey, Franks Steven, Progovac Liljana (eds.), *Formal approaches to Slavic linguistics*. — Ann Arbor: Michigan Slavic Publications. — P. 341–373.
- [12] Tatevosov Sergei. 2016. *Verb Classes and a Typology of Actionality* [Glagol’nye klassy i tipologija actsional’nosti]. — Moscow: Yazyki Slavianskoj Kultury.
- [13] Tatevosov Sergei. 2020. On the temporal structure of nonculminating accomplishments // *Linguistics*. — 2020. — Vol. 58, no. 5. — P. 1323–1371.
- [14] Tatevosov Sergei. 2022. *On impossible activities*. Ms., Lomonosov Moscow State University.
- [15] Varasdi Károly. 2014. Making progressives: necessary conditions are sufficient. *Journal of Semantics* — Vol. 31. — P. 179–207.

Distilled Model for Russian News Clustering: much lighter and faster, still accurate

Daria Trofimchuk

National University of Science and Technology

“MISiS”

Moscow, Russia

daryapeshch@gmail.com

Abstract

This paper explores abilities of knowledge distillation for the purposes of News clustering which also can be generalized as an event detection task. We used a BERT-based clustering model as a teacher and tested various student networks based on different architectures (RNN, FFN, convolutional and Transformer-based networks) in order to get a faster lightweight analogue that is more likely to be deployed in real products. We tried two distillation strategies: the first one combined an original loss function from the initial model with a distillation objective, for the second one we used only a specific distillation loss. This approach turned out to be more successful. It let us extend training and validation datasets and gave significantly better results. One of our distilled models scored about 1% lower than the teacher network, but is more than 20 times smaller and 5 times faster by inference.

Keywords: distillation; news clustering; distilled embeddings; pseudo-labeling; distill knowledge from BERT

DOI: 10.28995/2075-7182-2022-21-538-549

Дистиллят модели для кластеризации русскоязычных новостей: точный, но намного легче и быстрее

Дарья Трофимчук

НИТУ «МИСиС»

Москва, Россия

daryapeshch@gmail.com

Аннотация

Данная работа исследует возможности дистилляции знаний для задачи кластеризации новостей, которая также может быть рассмотрена в более общем виде как задача выявления новых событий. В качестве учителя мы использовали модель для кластеризации, основанную на архитектуре BERT, и протестировали различные типы нейросетей-учеников (на базе RNN, сверточных сетей, нейросетей с прямой связью и трансформеров), чтобы получить более быстрый и легковесный аналог, подходящий для использования в реальных продуктах. Мы проверили две стратегии дистилляции: в первом случае мы использовали комплексную функцию потерь, состоящую из функции, которая применялась в исходной модели, и из дистилляционного компонента. В рамках второй стратегии мы оптимизировали только специфическую функцию потерь для дистилляции. Этот подход оказался более продуктивным. Он позволил расширить тренировочный и проверочный наборы данных, что в свою очередь помогло достичь лучших результатов. Одна из протестированных моделей-дистиллятов уступила оригиналу порядка 1 % в качестве, но при этом получилась более чем в 20 раз компактнее и в 5 раз быстрее при инференсе.

Ключевые слова: дистилляция; кластеризация новостей; дистиллированные эмбединги; псевдо-метки; дистилляция знаний из BERT

1. Introduction

As the speed of information spreading boosts, the news clustering problem is becoming more and more challenging. Besides the accuracy of clustering systems, there are two main demands that's role is growing rapidly: optimization of memory/storage and speed-up for deployment of models in real (and mostly real time) applications.

This paper continues the research started in 2020 within the Telegram Data Clustering contest¹ and developed in the Dialogue Evaluation 2021 task on Russian news clustering². Based on results of work shown by contributors, supervisors of this competition stated that almost all of the models used by participants had been slow and extremely parameter-heavy [1]. They suggested distillation as one of promising directions for future studies.

In order to find models with good accuracy/speed/memory trade-offs, we experimented with different student architectures. At the first stage we tested RNN, FFN, convolutional and Transformer-based networks. Then, following results of our preliminary experiments, we proceeded with variations of LSTM-based architecture. As a result, we got a powerful and lightweight model which scored about 1% less than the teacher model, but outperformed it by inference time (1.5 ms versus 7.8 ms for getting one document embedding) and was just 31.3 MB versus 679.3 MB in size.

2. Related Work

Knowledge distillation is one of the blooming techniques for model compression and acceleration. First applications of this approach were shown by Cristian Buciluă and collaborators [2] in 2006, but the general methodology was introduced in 2015 [3] by Geoffrey Hinton et al. The main idea of distillation is to build a small student model which can mimic the large one called a teacher. Since the training process leverages outputs of the original model and doesn't depend on its architecture, almost any kind of a student model could be used for knowledge transfer.

There are different forms of student-teacher relations in distillation which could be described in terms of human beings learning. A student can get knowledge from a single teacher or from a group of them; teacher assistants can also take part in this process. Moreover, students can learn from each other (collaborative learning).

For their purposes distilled models could be roughly divided into two groups. The first one integrates task-specific networks. Besides the studies of Caruana and Hinton the work of Raphael Tang et al. [4] who suggested the way of distilling knowledge from BERT for classification purposes could also be listed as an example of this approach.

The second group includes multi- and even general-task distilled models. They are aimed at obtaining more generalized knowledge during the learning process. For instance, in the paper of Yang Ze et al. [5] this purpose is reached by using multiple teachers to jointly train a single student. Victor Sahn and collaborators went even further and proposed DistilBERT [6], a distilled version of BERT, which has the same backbone as its larger counterparts, can be fine-tuned on a range of tasks but is smaller and performs faster.

Distillation of Transformers has become a usual practice, so nowadays there is a bunch of different tiny versions of them based on knowledge distillation like DistilGPT2³ or Conversational Distil-RuBERT⁴.

It should be noted that distillation can intersect other ideas of model training. For example, in the early paper of Buciluă et al. [2] an ensemble teacher was trained on the relatively small data set and then used to label a large unlabeled data set. A student model was further trained on that labeled data.

This highly relates to pseudo-labeling techniques in self-training and semi-supervised learning. Pseudo-labels usually denote predicted class probabilities used as if they were true labels [7]. In self-training these synthetic labels are produced by a teacher model, and then passed to a student model — just like in distillation methodology described by Buciluă. A formal way to distinguish the two approaches is the size of a

¹ https://contest.com/docs/data_clustering2

² <https://www.dialog-21.ru/evaluation/>

³ <https://huggingface.co/distilgpt2>

⁴ <https://huggingface.co/DeepPavlov/distilrubert-base-cased-conversational>

student network: as mentioned earlier, the main idea of knowledge distillation is to get a smaller model. In self-training the same architectures are often used both for teacher and student.

Pseudo-labeling could be used in various general and downstream tasks of NLP. One of its most common applications now is to handle the problem of insufficient data for pre-training and fine-tuning Transformer-based models. Studies (e.g. [8] and [9]) show that self-training can improve their performance.

3. Original Task, Data and Initial Model

As stated above, this work elaborates and enhances some previous ideas on news clustering. The primary source of data used for the research was the Telegram Data Clustering contest. Its organizers provided all participants with HTML documents without any additional annotations. To make clustering process easier, authors of the next related contest (Dialogue Evaluation 2021) took a number of news documents from this dataset, composed pairs and annotated every pair via a Russian crowdsourcing service Yandex Toloka. Contributors were asked to determine whether two documents describe the same event or not [1]. Final markups included 15K annotated pairs for a training set, 8.5K pairs for a public leaderboard and 8.5K for a private one. The control metrics for the task was F1-score for positive examples in markup (we further denote it as positive F1 or just F1).

It is important to note that a clustering task could be tackled not only as clustering itself but also as a news pair binary classification problem, and in fact some contest solutions using classification overcame clustering-based models. Nevertheless, the clustering approach seems to be more promising for real-life applications, and the reason is computational efficiency. Classification models require pairwise news comparison, while by clustering we can simply obtain a document-level representation and use it for a single neighbor search.

A starting point for this research became a clustering approach proposed by the naergvae team which participated in the Dialogue Evaluation 2021 and described their work in [10]. They trained a BERT-based model which could produce a fixed-size embedding vector for every news document. On this stage a hard triplet loss was used. Output news representations were then passed to the agglomerative algorithm for clustering.

The general idea of this solution were reproduced⁵ by the organizers of the Dialogue Evaluation 2021 Clustering contest. That reproduction is leveraged for the further distillation experiments described in this paper.

4. Methodology and Complex Loss Experiments

For our research we chose the following basic approach: for performance acceleration we should build a lightweight replica of the BERT-based embedder component and pass it to the same unsupervised clustering algorithm. In all our experiments we used an initial embedder model as a teacher.

For preprocessing of documents we took a pre-trained tokenizer from the teacher model. Partly it was an empirical decision, but additionally it was supported by previous distillation experiments which showed [11] that knowledge transfer works better if input embeddings — which depend on input tokens — have the same spaces.

After tokenization we passed batches of data through an embedding layer and pushed sequences of output word vectors to a core part of our student network. Then we used some aggregation function (more on aggregation in section 5) to get a single output vector for every document. In the end we got news vector representations of the same size as produced by a teacher model.

As for training objectives, we had two conceptually different schemes which will be detailed below.

4.1. Complex Loss Distillation

The first approach derives from the original work of Tang et al. [4]. They also used a BERT-based teacher, designed a distilled model for classification purposes and suggested a complex objective combining a traditional cross-entropy loss with a mean-squared-error (MSE) loss between student and teacher logits (distillation loss).

⁵ <https://github.com/dialogue-evaluation/Russian-News-Clustering-and-Headline-Generation>

We tried to adapt this formula to our case. Instead of a cross-entropy loss we used a triplet loss from the initial model (teacher). A resulting formula could be schematically described as follows:

$$L = \alpha \cdot L_{triplet} + (1 - \alpha) \cdot \frac{1}{3} \sum_1^3 L_{distill} \quad (1)$$

where α is a balancing coefficient for two components of the loss and $L_{distill}$ denotes a specific distillation loss which is evaluated for every document in a triplet and then averaged.

The idea of the distillation loss for our case was inspired by [12], [13] and [14]. In these studies student networks distilled the abstraction hidden in teachers by matching their internal representations. Authors implemented additional losses based on affinity of hidden states. Although in our research we don't distill on hidden layers, the final output of both teacher and student models are vector representations which means we could use resembling loss functions.

Thus, for the distillation loss we took a MSE loss between news vector representations produced by teacher and student models. It references Tang et al. [4] and partially Romero et al. [12] and Sun et al. [14] who used different modifications of the MSE. Our distillation loss function was formulated as follows:

$$L_{distill} = L_{MSE} = \frac{1}{n} \sum_1^n (v_{T_i} - v_{S_i})^2 \quad (2)$$

where n denotes dimensionality of vectors v , T and S define teacher and student networks respectively.

Following [4], we took an LSTM-based model as a student for our preliminary experiments. A one directional recurrent network was followed by dropout and two linear layers. The outputs were averaged to get a resulting embedding. We trained our models with early stopping after 2 epochs of non-decreasing validation loss. Unfortunately, their results were unsatisfactory (see Table 1).

It should be noted that the complex loss approach has some crucial weaknesses.

One of them is how triplets for training are formed. There are pivot, positive and negative examples in every triplet, and if a news document has no positive pair in the markup, a pivot document itself is used as a positive example. Moreover, if one document has some positive or/and negative pairs, it reappears in triplets. That leads to inconsistent parameter updates and affects the quality of distillation.

Besides, this approach is strongly dependent on the markup: we can't evaluate the triplet loss without relevant news document pairs. The number of annotated pairs turned out to be not enough for proper training.

<i>Model</i>	<i>a</i>	<i>F1, public LB</i>	<i>F1, private LB</i>
Teacher model		94.6%	94.5%
LSTM + 2 linear layers + dropout	0.5	40.1%	39%
LSTM + 2 linear layers + dropout	0.2	37%	33.4%
LSTM + 2 linear layers + dropout	1	38.6%	34.7%

Table 1: Results of student models trained with the complex loss (with different α coefficients) on the Dialogue Evaluation 2021 data for the Russian news clustering task

4.2. Single Loss Distillation (MSE and Cosine Embedding Loss)

Our next step naturally followed the previous one. We decided to exclude the initial loss from our objective and evaluated difference between news vector representations only. This move let us solve the problem of insufficient training data (and brought us closer to self-learning technique).

For this approach we could use any relevant documents with absolutely no additional annotations. In order to get teacher's pseudo-labels we just needed to pass tokenized texts through the initial model and get news embeddings as an output.

Theoretically we could enrich our dataset by scraping, but we turned to the content of the original Telegram contest instead. There were about 690K raw documents, and after removing news from public and private leaderboards we ended up with almost 650K of them.

We experimented with two different loss functions. Besides the MSE loss, we tried a cosine embedding loss which was also used as one of distillation objectives in [6] and [13] — as an element that tends to align directions of student and teacher vectors. This is formulated as:

$$L_{distill} = L_{cos} = 1 - \cos(v_T, v_S) \quad (3)$$

Since we didn't want to connect the initial model with a student, we leveraged offline distillation methodology, i.e. precomputed teacher's embeddings in advance. Further we were combining them with the relevant tokenized documents and loading this data to our lightweight model.

In order to prove that the switch to a single loss itself is not enough without additional data, we trained a student model on the same examples as in previous experiments but didn't form triplets and applied just the MSE loss. This model formed too many clusters and showed poor results. For comparison we also trained a model with the complex loss function with $\alpha = 0$, which meant that in this case we evaluated only $L_{distill}$, i.e. the same MSE loss. On the contrary, this network didn't form clusters which indicates an average cluster size and 100% recall. It illustrates how different one loss function works for triplets and single documents. Scores of both models (and the model trained with the complex loss with $\alpha = 0.5$) can be examined in Table 2.

<i>Model</i>	<i>Loss</i>	<i>F1, public LB</i>	<i>Recall, public LB</i>	<i>Precision, public LB</i>	<i>Avg cluster size, public LB</i>
Teacher model		94.6%	95%	94.1%	3.23
LSTM + 2 linear layers + dropout	complex, $\alpha = 0.5$	40.1%	29.5%	62.5%	1181.18
LSTM + 2 linear layers + dropout	complex, $\alpha = 0$	63.1%	100%	46.1%	20080
LSTM + 2 linear layers + dropout	single (MSE)	57.7%	41.2%	96.5%	1.85

Table 2: Scores of LSTM-based models trained on the same data (Dialogue Evaluation 2021) but with different loss functions. F1, recall and precision are evaluated for positive news pairs labeled via Yandex Toloka for the Dialogue Evaluation 2021. For comparison we used examples from the public leaderboard. An average cluster size is also evaluated for these news documents.

5. Preliminary Single Loss Experiments

For our preliminary single loss experiments we also started with LSTM-architectures. Then we tried other recurrent neural networks, convolutional, Transformer-based and feed-forward students.

Besides, we applied different aggregation functions. By default we used average aggregation simply averaging hidden states/outputs of a main part of the student embedder (not taking into account padding positions).

We also tested aggregation via simple attention implementation suggested by Raffel and Ellis in [15]. It allows to produce a single fixed-length embedding from an entire sequence by computing an adaptive weighted average of hidden states/output vectors. This implementation could be formulated as follows:

$$e_t = a(h_t), a_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)}, c = \sum_{t=1}^T a_t h_t \quad (4)$$

where a is some learned function, h_t is a hidden state of some token t , T is the length of the sequence and c is a resulting aggregated vector. We will further refer to it as *attentive aggregation*. In our case for a function we used a two layer feed-forward network with ReLU activation. Padding positions were additionally masked with zeros.

One more aggregation function was tested for Transformer-based models, it will be detailed in Section 5.3.

Different student models were trained on GPU with a batch size of 128, 20 epochs each.

5.1. LSTM-based Student Networks

Since the length of news documents is not the same, for proper optimization with recurrent neural networks we used masking via packed padded sequences. Then we aggregated hidden states of the sequence tokens. In some experiments we additionally passed resulting outputs through linear layers. In the end we got news vector representations of the same size as produced by a teacher model.

We tested both one- and bidirectional LSTMs, with or without additional linear layer(s), with or without dropout. We mostly used 128-dimensional token embeddings except experiments where we initialized this layer with pre-trained embeddings from the initial model. The hidden size of LSTM was set for 128. In some cases the attentive aggregation method was used. Results of this work are shown in Table 3.

<i>Model</i>	<i>Loss</i>	<i>F1 public LB</i>	<i>F1 private LB</i>	<i>Size</i>
Teacher model		94.6%	94.5%	679.3 MB
LSTM + 2 linear layers + dropout	MSE	91.4%	91.8%	59.3 MB
LSTM + 2 linear layers + dropout	Cosine	92.4%	92.6%	59.3 MB
Bidirectional LSTM + 2 linear layers + dropout	MSE	92.3%	92.3%	59.9 MB
Bidirectional LSTM + 2 linear layers + dropout	Cosine	93%	92.6%	59.9 MB
Bidirectional LSTM + 1 linear layer	Cosine	93.4%	92.6%	59.6 MB
Bidirectional LSTM + 1 linear layer / <i>attentive aggregation</i>	Cosine	93.8%	93.3%	59.8 MB
Bidirectional LSTM (2 layers) + 1 linear layer	Cosine	93.9%	93.1%	61.1 MB
Bidirectional LSTM (2 layers) + 1 linear layer / <i>attentive aggregation</i>	Cosine	93.9%	93.4%	61.3 MB
Bidirectional LSTM	Cosine	92.7%	92.1%	59.4 MB
Bidirectional LSTM / <i>attentive aggregation</i>	Cosine	93.3%	92.9%	59.5 MB
Bidirectional LSTM + 2 linear layers + dropout / <u>from pretrained</u>	MSE	94%	93.5%	354.2 MB
Bidirectional LSTM + 2 linear layers + dropout / <u>from pretrained</u>	Cosine	94%	93%	354.2 MB

Table 3: Positive F1 scores and size of LSTM-based students in comparison with the teacher model evaluated on the Dialogue Evaluation 2021 data. Student models were trained with a single loss.

As could be seen, student networks using pre-trained token embeddings demonstrate best results, but due to dimensionality of these vectors (768) such models have more parameters and are much heavier than the others. Considering the fact that we were looking for an accuracy/memory trade-off, we decided not to proceed with that type of architecture.

Besides, we noticed that bidirectional models trained with a cosine embedding loss showed better scores, so further we experimented mostly with their variations.

5.2. GRU-based Students

To compare different types of recurrent nets, we also tested GRU-based students. We set the same parameters as for LSTMs and used the most successful configurations from previous experiments. Results of this work could be examined in Table 4.

<i>Model</i>	<i>Loss</i>	<i>F1 public LB</i>	<i>F1 private LB</i>	<i>Size</i>
Teacher model		94.6%	94.5%	679.3 MB
Bidirectional GRU + 1 linear layer	Cosine	93.4%	93.2%	59.4 MB
Bidirectional GRU + 1 linear layer / <i>attentive aggregation</i>	Cosine	93.7%	92.7%	59.5 MB
Bidirectional GRU (2 layers) + 1 linear layer	Cosine	93.5%	93.4%	60.5 MB
Bidirectional GRU (2 layers) + 1 linear layer / <i>attentive aggregation</i>	Cosine	93.8%	93%	60.6 MB
Bidirectional GRU	Cosine	93.3%	92.6%	59.1 MB
Bidirectional GRU / <i>attentive aggregation</i>	Cosine	93.2%	93.1%	59.3 MB

Table 4: Positive F1 scores and size of GRU-based networks evaluated on the Dialogue Evaluation 2021 data. Student models were trained with a cosine embedding loss.

5.3. Transformer-based Student Models

One more architecture used for our experiments was based on encoder layers of Transformers. We utilized their PyTorch implementation⁶. Parameter *src_key_padding_mask* let us mask padding tokens.

The dimension of the encoder feedforward network was set to 1024. We used 128-dimensional token embeddings, the number of encoder layers and self-attention heads in them varied.

We stacked encoder layers and added above them a mapping linear layer if encoder outputs and teacher document embeddings had different sizes. In some cases we also placed ReLU between encoder outputs and a mapping layer. After that we applied one of the aggregation functions.

In addition to the previously described functions we tested an aggregation methodology inspired by the BERT paper. Delvin et al. suggested using a special first token of every sequence token ([CLS]) as the aggregate representation for classification tasks. To get some prediction they pushed it through a linear layer and an activation function [16].

Since our tokenizer also produces [CLS] tokens, we tried to go the same way except that we added one more linear layer after activation. An output of this layer was used as a resulting vector. We called this tactic *CLS-aggregation*. Results of different experiments with Transformers are shown in Table 5.

<i>Model</i>	<i>Loss</i>	<i>F1 public LB</i>	<i>F1 private LB</i>	<i>Size</i>
Teacher model		94.6%	94.5%	679.3 MB
Transformer encoder: 1 layer, 4 heads	Cosine	90.9%	90.4%	61 MB
Transformer encoder: 1 layer, 4 heads, ReLU	Cosine	90.9%	90.7%	61 MB
Transformer encoder: 1 layer, 4 heads / <i>attentive aggregation</i>	Cosine	91.2%	90.3%	61.2 MB
Transformer encoder: 1 layer, 4 heads / <i>CLS-aggregation</i>	Cosine	89.5%	89.3%	61.3 MB

⁶ <https://pytorch.org/docs/stable/generated/torch.nn.TransformerEncoderLayer.html>

<i>Model</i>	<i>Loss</i>	<i>F1 public LB</i>	<i>F1 private LB</i>	<i>Size</i>
Transformer encoder: 1 layer, 4 heads, 64-dimensional token embeddings	Cosine	90%	90.6%	30.4 MB
Transformer encoder: 1 layer, 8 heads, ReLU	Cosine	90.9%	90.3%	61 MB
Transformer encoder: 2 layers, 8 heads	Cosine	90.8%	90.6%	62.3 MB
Transformer encoder: 2 layers, 4 heads	Cosine	90.1%	90.1%	62.3 MB
Transformer encoder: 2 layers, 4 heads / <i>attentive aggregation</i>	Cosine	90.7%	89.7%	62.4 MB
Transformer encoder: 2 layers, 4 heads / <i>CLS-aggregation</i>	Cosine	89.6%	89.7%	62.5 MB

Table 5: Positive F1 scores and size of Transformer-based student models evaluated on the Dialogue Evaluation 2021 data. Models were trained with a cosine embedding loss.

5.4. Other Student Networks

We also conducted some experiments with convolutional, feedforward neural networks (FFN) and RNN-based embedders as students.

For convolutions two strategies were tested, both of them applied combinations of convolutions and pooling-layers. For the first strategy (CNN-1) we used pooling to reduce dimensionality of each sequence to 1. After convolutions and pooling-layers we added a mapping layer in order to transform the size of a resulting vector and make it equal to a teacher embedding.

For the second approach (CNN-2) we gradually increased the number of convolutional filters to get a target size of a resulting vector. The other dimensions were shrunk via pooling.

Our FFN student embedders were constructed from simple linear layers with ReLU activations. For every next layer we doubled the number of output features. At the top of this FFN we put a mapping layer and then applied average aggregation.

Additionally we experimented with a bidirectional RNN. It had the same parameters as LSTM and GRU students. Results of these experiments are reported in Table 6.

<i>Model</i>	<i>Loss</i>	<i>F1 public LB</i>	<i>F1 private LB</i>	<i>Size</i>
Teacher model		94.6%	94.5%	679.3 MB
CNN-1	MSE	88.8%	88.9%	58.4 MB
CNN-2	MSE	84.5%	85.6%	59.1 MB
FFN (3 layers + mapping)	Cosine	91.2%	89.5%	62 MB
FFN (2 layers + mapping)	Cosine	87.9%	87.9%	59.5 MB
Bidirectional RNN + 1 linear layer	Cosine	91.3%	90.7%	58.9 MB

Table 6: Positive F1 scores and size of convolutional, feedforward and RNN-based student networks. Models were trained with a cosine embedding loss and evaluated on the Dialogue Evaluation 2021 data.

6. Search for Best Model

Given the results of preliminary experiments, we turned our attention to three most promising student networks: bidirectional LSTM with 2 recurrent layers, bidirectional GRU with 1 recurrent layer and bidirectional GRU with 2 recurrent layers. In all of them a recurrent part was followed by aggregation and a linear layer (by default average).

In these experiments we used ReduceLRonPlateau scheduler⁷ with a starting learning rate 1e-3 and patience equal to 2. We also used early stopping with activation after 3 epochs of non-decreasing validation loss. Batch size was set to 128.

At first, we tested three models with a cosine embedding loss. For two of them that showed higher scores we experimented with a MSE loss. Since this loss function gave us slightly better results, we used MSE for two experiments with the attentive aggregation.

At the end, we decided to check if reducing a token embedding size would drastically affect the quality of a distilled model. For this purpose we built a bidirectional GRU with 2 recurrent layers and the attentive aggregation but used 64-dimensional token embeddings instead of 128-dimensional. As expected, it almost halved the size of our model. However, results of this student network dropped to some extent.

Comparison of distilled models with the initial one is presented in Table 7.

<i>Model</i>	<i>Loss</i>	<i>F1 public LB</i>	<i>F1 private LB</i>	<i>Size</i>
BiLSTM (2 layers) + 1 linear layer	Cosine	93.7%	93.4%	61.1 MB
BiGRU (2 layers) + 1 linear layer	Cosine	93.7%	93.1%	60.5 MB
BiGRU (1 layer) + 1 linear layer	Cosine	93.3%	92.7%	59.4 MB
BiLSTM (2 layers) + 1 linear layer	MSE	93.6%	93.7%	61.1 MB
BiGRU (2 layers) + 1 linear layer	MSE	93.8%	93.3%	60.5 MB
BiLSTM (2 layers) + 1 linear layer / <i>attentive aggregation</i>	MSE	93.8%	93.6%	61.3 MB
BiGRU (2 layers) + 1 linear layer / <i>attentive aggregation</i>	MSE	93.9%	93.7%	60.6 MB
BiGRU (2 layers) + 1 linear layer / <i>attentive aggregation / 64-dimensional token embeddings</i>	MSE	93.5%	93.5%	31.3 MB

Table 7: Comparison of student models obtained on the last stage of experiments. By default 128-dimensional token embeddings were used. Models were evaluated on the Dialogue Evaluation 2021 data.

⁷ https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLRonPlateau.html

Table 7 shows that results of models are relatively close, so it would be incorrect to state that one student definitely outperforms the rest. A choice of the best distilled model is a matter of discussion, but we suggest that bidirectional GRU-based models (BiGRUs) with 2 recurrent layers and the attentive aggregation are rather useful for practical issues. The BiGRU with 128-dimensional token embeddings gave the highest positive F1 score on two leaderboards. At the same time the network with 64-dimensional embeddings was the smallest in size. Assuming that a small additional drop in quality is not crucial, we would denote the second model as the most successful distilled student.

The scheme of training and inference of a network of this type are shown in Figure 1.

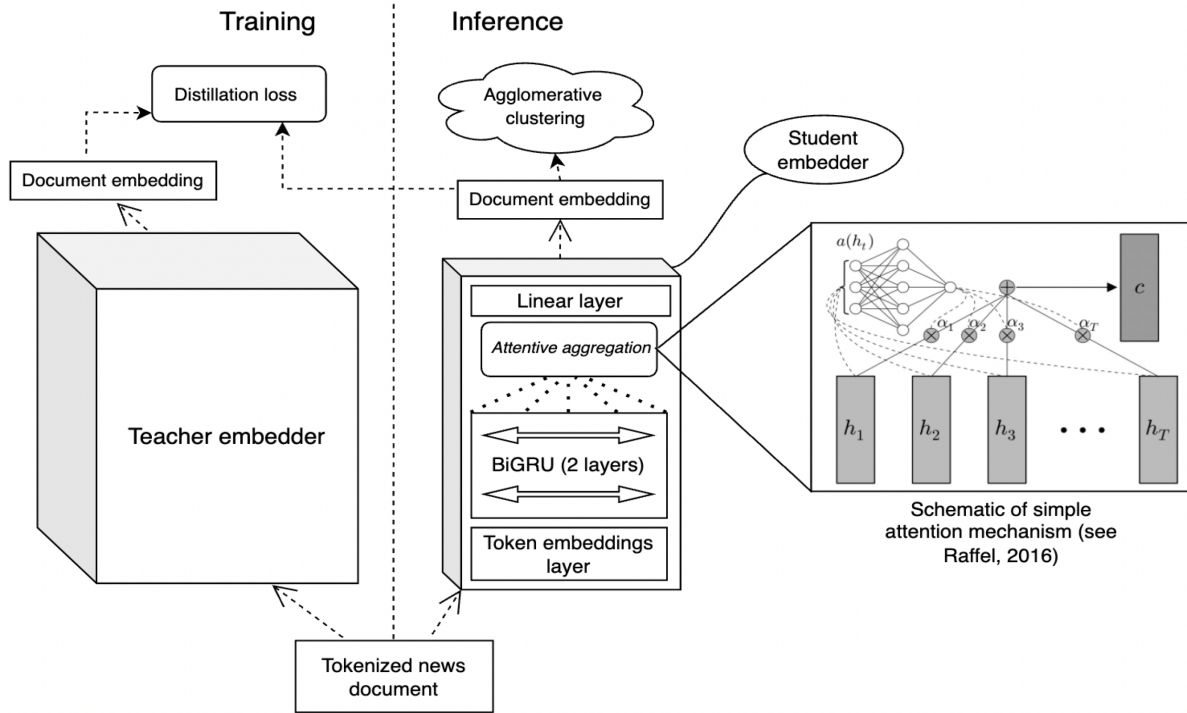


Figure 1: Architecture and training/inference flow of the BiGRU-based distilled model. The scheme of attentive aggregation depicts formula (4) from the paper [15] of Raffel et al.

To emphasize an effect of distillation we also trained the same models on the labeled data with the original triplet loss. Additionally, we fine-tuned on the original task three distilled BERT-based models: Conversational DistilRuBERT, RuBERT Tiny⁸ and RuBERT Tiny-2⁹. Their results can be examined in Table 8.

<i>Model</i>	<i>F1 public LB</i>	<i>F1 private LB</i>	<i>Size</i>	<i>Inference time per document</i>
Teacher Model	94.6%	94.5%	679.3 MB	~7.8 ms
BiGRU Student (128-dimensional token embeddings)	<u>93.9%</u>	<u>93.7%</u>	60.6 MB	~1.5 ms
BiGRU Student (64-dimensional token embeddings)	93.5%	93.5%	<u>31.3 MB</u>	~1.5 ms
BiGRU without distillation (128-dimensional token embeddings)	63.2%	62.9%	60.6 MB	~1.5 ms
BiGRU without distillation (64-dimensional token embeddings)	55.7%	61%	31.3 MB	~1.5 ms
RuBERT Tiny	83.8%	82.7%	45.3 MB	~0.6 ms
RuBERT Tiny-2	90.3%	91.1%	111.7 MB	~0.6 ms
DistilRuBERT	90.7%	91%	406.6 MB	~1.3 ms

Table 8: Comparison of two distilled models (BiGRU Students) with the teacher model and other networks: BiGRU-based models with the same architecture as the students but trained using the original triplet loss without distillation (rows 4 and 5) and distilled BERT models for Russian language fine-tuned on the original Russian news clustering task (rows 6–8). Models were evaluated on the Dialogue Evaluation 2021 data. Reported inference time per document was estimated by averaging inference time for all tokenized documents in the public leaderboard dataset. Models were tested on GPU Tesla P100 PCIe.

7. Conclusion

Thus, we got a student model which is practically as accurate as its teacher but is smaller and faster. For reproducibility we published its training code at our repository¹⁰. Compression by more than 20 times seems to be rather productive — like significant growth of inference speed. This model is much more adaptable for real-life purposes.

Nevertheless, it should be noted that the distillation objective used in our experiment is totally dependent on the original model. This means that such an approach doesn't let us get a student model which could overcome its teacher. The search for a formula that could combine distilled knowledge with some complementary correction work (and wouldn't need extra data labeling) should be a promising path for further studies. Moreover, it would also be efficient to try different distillation methods on the news clustering approaches based on classification solutions.

Acknowledgements

We are grateful to anonymous reviewers for their valuable comments. Additionally, we would like to thank Ilya Gusev for the idea of this work and helpful suggestions during the whole modeling process.

⁸ <https://huggingface.co/cointegrated/rubert-tiny>

⁹ <https://huggingface.co/cointegrated/rubert-tiny2>

¹⁰ <https://github.com/DariaTrofimchuk/distillation>

References

- [1] Gusev Ilya, Smurov Ivan. Russian News Clustering and Headline Selection Shared Task // *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”* (2021), Issue 20. — Moscow, Russia, 2021. — P. 289–302.
- [2] Bucilua Cristian, Caruana Rich, Niculescu-Mizil Alexandru, Model compression // *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* — New York, NY, United States, 2006 — P. 535–541.
- [3] Hinton Geoffrey, Vinyals Oriol, Dean Jeff. Distilling the Knowledge in a Neural Network // *Computing Research Repository*. — 2015. — Vol. arXiv:1503.02531. — version 1. Access mode: <https://doi.org/10.48550/arXiv.1503.02531>.
- [4] Tang Raphael, Lu Yao, Liu Linqing, Mou Lili, Vechtomova Olga, Lin Jimmy. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks // *Computing Research Repository*. — 2019 — Vol. arXiv:1903.12136. — version 1. Access mode: <https://doi.org/10.48550/arXiv.1903.12136>.
- [5] Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, Daxin Jiang. Model Compression with Multi-Task Knowledge Distillation for Web-scale Question Answering System // *Computing Research Repository*. — 2019 — Vol. arXiv:1904.09636. — version 1. Access mode: <https://doi.org/10.48550/arXiv.1904.09636>.
- [6] Sanh Victor, Debut Lysandre, Chaumond Julien, Wolf Thomas. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter // *Computing Research Repository*. — 2019 — Vol. arXiv:1910.01108. — version 3. Access mode: <https://doi.org/10.48550/arXiv.1910.01108>.
- [7] Lee Dong-Hyun. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks // *Workshop on challenges in representation learning, ICML*. Volume 3. Issue 2, 2013.
- [8] Du Jingfei, Grave Edouard, Gunel Beliz, Chaudhary Vishrav, Celebi Onur, Auli Michael, Stoyanov Ves, Conneau Alexis. Self-training Improves Pre-training for Natural Language Understanding // *Computing Research Repository*. — 2020 — Vol. arXiv:2010.02194. — version 1. Access mode: <https://doi.org/10.48550/arXiv.2010.02194>.
- [9] Kuligowska Karolina, Kowalczyk Bartłomiej. Pseudo-labeling with transformers for improving Question Answering systems // *Procedia Computer Science*. Volume 192, 2021. — P. 1162-1169.
- [10] Khaustov S.V., Gorlova N.E., Kalmykov A.V., Kabaev A.S. BERT for Russian news clustering // *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”* (2021), Issue 20. — Moscow, Russia, 2021. — P. 385–391.
- [11] Kim Young Jin, Awadalla Hany Hassan. FastFormers: Highly Efficient Transformer Models for Natural Language Understanding // *Computing Research Repository*. — 2020 — Vol. arXiv:2010.13382. — version 1. Access mode: <https://doi.org/10.48550/arXiv.2010.13382>.
- [12] Romero Adriana, Ballas Nicolas, Kahou Samira Ebrahimi, Chassang Antoine, Gatta Carlo, Bengio Yoshua. FitNets: Hints for Thin Deep Nets // *Computing Research Repository*. — 2015 — Vol. arXiv:1412.6550. — version 4. Access mode: <https://doi.org/10.48550/arXiv.1412.6550>.
- [13] Aguilar Gustavo, Ling Yuan, Zhang Yu, Yao Benjamin, Fan Xing, Guo Chenlei. Knowledge Distillation from Internal Representations // *Computing Research Repository*. — 2020 — Vol. arXiv:1910.03723. — version 2. Access mode: <https://doi.org/10.48550/arXiv.1910.03723>.
- [14] Sun Siqu, Cheng Yu, Gan Zhe, and Liu Jingjing. Patient Knowledge Distillation for BERT Model Compression // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* — Hong Kong, China, 2019 — P. 4323–4332.
- [15] Raffel Colin, Ellis Daniel P. W. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems // *Computing Research Repository*. — 2016 — Vol. arXiv:1512.08756. — version 5. Access mode: <https://doi.org/10.48550/arXiv.1512.08756>.
- [16] Devlin J., Chang M., Lee K., & Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1 (Long and Short Papers) — Stroudsburg, PA, USA, 2019. — P. 4171–4186.

Is neural language acquisition similar to natural? A chronological probing study

Ekaterina Voloshina

AIRI, HSE University

Moscow, Russia

vokat@mail.ru

Oleg Serikov

AIRI, DeepPavlov, HSE University

Moscow, Russia

srkvoa@gmail.com

Tatiana Shavrina

AIRI, SberDevices

Moscow, Russia

rybolos@gmail.com

Abstract

The probing methodology allows one to obtain a partial representation of linguistic phenomena stored in the inner layers of the neural network, using external classifiers and statistical analysis.

Pre-trained transformer-based language models are widely used both for natural language understanding (NLU) and natural language generation (NLG) tasks making them most commonly used for downstream applications. However, little analysis was carried out, whether the models were pre-trained enough or contained knowledge correlated with linguistic theory.

We are presenting the chronological probing study of transformer English models such as MultiBERT and T5. We sequentially compare the information about the language learned by the models in the process of training on corpora. The results show that 1) linguistic information is acquired in the early stages of training 2) both language models demonstrate capabilities to capture various features from various levels of language, including morphology, syntax, and even discourse, while they also can inconsistently fail on tasks that are perceived as easy.

We also introduce the open-source framework for chronological probing research, compatible with other transformer-based models. https://github.com/EkaterinaVoloshina/chronological_probing

Keywords: probing, language acquisition, language modeling, transformers

DOI: 10.28995/2075-7182-2022-21-550-563

Усвоение языка у языковых моделей и человека: хронологическое пробинг-исследование

Екатерина Волошина

AIRI, НИУ ВШЭ

Москва, Россия

vokat@mail.ru

Олег Сериков

AIRI, DeepPavlov, НИУ ВШЭ

Москва, Россия

srkvoa@gmail.com

Татьяна Шаврина

AIRI, SberDevices

Москва, Россия

rybolos@gmail.com

Аннотация

Пробинг-методология позволяет получить представление о явлениях языка, хранящиеся во внутренних слоях нейросети, с помощью внешних классификаторов и статистического анализа.

Предобученные языковые модели на основе трансформерных архитектур широко используются как для задач понимания естественного языка (NLU), так и для задач генерации (NLG), что делает их наиболее часто используемыми для широкого ряда применений. Однако, недостаточно проводился анализ, достаточно ли предобучены модели и содержат ли знания, коррелирующие с теоретическими представлениями о языке.

Мы представляем исследование на основе хронологического пробинга на примере моделей MultiBERT и T5, в котором последовательно исследуем выучиваемую информацию о языке в процессе предобучения моделей на корпусе. Результаты показывают, что 1) лингвистическая информация усваивается уже на ранних этапах обучения 2) обе языковые модели демонстрируют способность фиксировать различные свойства языка на разных уровнях, включая морфологию, синтаксис и дискурс, в то же время они могут не справляться с задачами, которые воспринимаются как простые.

Мы также предоставляем открытый фреймворк для хронологического пробинга, совместимый с языковыми моделями на основе архитектур transformer. https://github.com/EkaterinaVoloshina/chronological_probing

Ключевые слова: пробинг, усвоение языка, языковые модели, трансформеры

1 Introduction

The role of deep learning language models has been increasing in the field of methodology for linguistic research, providing new methods for both diachronic and synchronic studies [Manning2015]. In particular, transformer-based language modeling research has produced a variety of tools that may discover regularities and structures in data, many of which have resulted in practical applications.

In this study, we search for a match between the language competencies of popular language models and compare their results with the levels of a first language learner. As the transformer models are expected to acquire a language during the training process, the probing methodology has shed light on model training success. Probing tasks are usually classification tasks where classes represent different values of a linguistic features, such as a subject number, tree depth, and a connector type. Theoretical representation of language often inquires about the levels of phonetics, morphology, syntax, and discourse/pragmatics to be involved in a probing study ¹.

The main focus of this work is to explore how language models acquire measurable linguistic structures during training. The contributions of our work are the following:

- We propose a methodology for chronological probing, based on checkpoint-wise result comparison during model training². We denote chronological probing as any probing technique that refers to the training history/iterations of the same model.
- We test two models (MultiBERT [Sellam et al.2021] and T5 [Raffel et al.2019]) on existing 12 different probing tasks in morphology, syntax, and discourse and present an analysis of the models' gradual learning of language phenomena, in comparison with the well-known facts about the acquisition of the first language by a child.
- We present the evaluation results for the named models and state that the models tend to learn the linguistic phenomena in a specific order, and some parts of grammar are “acquired” first.

The presented framework and methodology are available open-source under Apache 2.0 license.

2 Related work

Until recent years, the task of learning syntax, which every five-year-old child performs effortlessly, has eluded brute language modeling force. This makes the language models a particular object of study, considered both from the interpretability and modeling language acquisition. As [De Villiers and Roeper2011, p. 119] states, “*computational models of language acquisition must begin and end as an integral part of the empirical study of child language.*”

Following this thesis, we turn our attention to the probing methodology and comparable case studies in the field of language acquisition, focusing on the transformer architectures.

2.1 Probing and approaches to the black box of language modeling

An increasing number of works are devoted to interpreting language models from a linguistic point of view. The quickly advancing field of *probing* received lots of researchers' attention when the hegemony of the large black-box models was set up. Researchers question the extent of the models' “understanding” of the language in probing. They inspect if, and to what limits, the language models' behavior agrees with the insights of the theory of language. Following the hierarchy of language levels (morphology, syntax, discourse) [Dalrymple2001], the probing studies often suppose the experiments related to models' proficiency on a certain level of language.

This line of research typically comes down to analyzing how linguistic structures are represented in a model's knowledge. Such structures represent syntagmatic/paradigmatic mechanisms (how language units combine and alternate, respectively) of language. It is believed [McCoy et al.2020], that rediscovering these structures would help models to get closer to human performance on a variety of tasks.

¹However, some researchers [Embick and Noyer2007, Caha2009] doubt that a language functions as a level system. They suggest that morphology and syntax operate at the same time. Other researchers argue that morphology and syntax are different layers of a language.

²https://github.com/EkaterinaVoloshina/chronological_probing

Probing, in general, considers how interpretable the behavior of the language model wrt the linguistic properties of the data. A huge body of probing studies rely on linear models (e.g., external classifiers [Belinkov2016]) that try to establish the relationship between internal representations from the language model and the desired linguistic phenomena. Thus, the linear correlation is measured between the model's forward pass embeddings and the linguistic properties of the passed data. A sample study [Tenney et al.2019] could measure the strength of correlation between a model's particular layer activations on some word and word's part-of-speech.

Strong correlations have been recorded when comparing the models' forward pass activations with the passed data underlying linguistic structure [Belinkov et al.2017, Tenney et al.2019, Conneau et al.2018a, Hewitt and Liang2019a] using probing methods.

Such a high performance could be misleading. The properties of the model and the properties of the used data impact the resulting score of the correlation probing study. Thus, given only a correlation score, one does not know if it reflects the model's (but not the corpus itself) linguistic informativeness. As a result, several approaches to conducting more reliable studies have been proposed. [Hewitt and Liang2019a, Zhang and Bowman2018, Voita and Titov2020, Pimentel et al.2020].

The probing methodology combining various annotated data is commonly used as the benchmark for language model comparison and evaluation of their generalizing ability. The SentEval toolkit [Conneau and Kiela2018] has led to the popularization of the 10 tasks used to distinguish between random and justified, brittle, and robust results of model training, including different types of architectures. However, analogous research on the same architecture or even the same model is in its early development stage. The first work on probing of neural networks across time was carried by [Saphra and Lopez2018]. The authors showed that first, LSTM acquires syntactic and semantic features and later information structure. [Chiang et al.2020a] looked at the training process of ALBERT and concluded that semantic and syntactic information is acquired during the early steps while world knowledge fluctuates during the training. [Liu et al.2021] showed similar results on RoBERTa: the model shows good results on linguistic probing tasks starting from early stages, and later it learns factual and commonsense knowledge.

Chronological probing could enrich the interpretable documentation of model training in time and thus explore the new aspects of model training and more clearly expose its problems.

2.2 Language acquisition and language models

Language learning is one of the quintessential human traits. First language acquisition(LA), unites both neurocognitive research, psycholinguistics, and computational approaches, focusing on the ability to acquire the capacity to perceive and comprehend language.

Statistical language acquisition Language modeling has formed a branch in language acquisition studies named statistical language acquisition. Various aspects of language, including phonological, syntactic, lexical, morphological, and semantic features, were investigated in terms of statistical patterns children receive with the linguistic input. Recent studies postulating qualitative and quantitative measures of LA include:

- **Morphology and Syntax** Morphology and syntax studies across language acquisition studies are definitely those explored the most. Starting with the poverty of stimulus problem and the argument between innateness and learning of grammar, it has led to typologically various sets of descriptive works and even computational models of the acquisition process. Thus, [Lewis and Elman2001] train a simple RNN to discriminate between grammatical strings that follow the inversion rule and those that do not (e.g., moving the first auxiliary verb such as "Is the man that tall is nice?"). The training data for the study is generated artificially and fails to prove that such a network generalizes on a mixture of diverse syntactic constructions. [Real and Christiansen2005] use bigram models to capture the patterns of auxiliary inversion based on lifelike data from child-directed speech. The model can consistently assign higher probabilities to grammatical strings than ungrammatical strings, which was interpreted as having successfully learned the correct inversion rule. However, as [Kam et al.2008] note, this result is because bigrams such as "who are" are much more frequent than the ungrammatical strings. [Prefors et al.2006] approaches the structure dependency problem

with Bayesian learning and attempts to learn a grammar that could generate additional sentences. The model evaluates and selects between two grammars, a finite state grammar and a context-free grammar constructed by the authors based on a simplified subset of child-directed English. It is worth noting how similar all the problem formulations are to the modern formulations of probing classification problems described below. They are also far from a complete description of the process of mastering grammar.

- **Discourse** The creation of texts, not sentences, with various discourse features, such as competence in speech acts, conversations, speech registers, and extended speaking turns, is more often considered a later stage of speech development. There are no computer models for the assimilation of discursive properties comparable to models for morphology and syntax. However, research in this direction is underway.

In [Ororbia et al.2019], authors examine whether neural language models acquire language better when trained in a multi-modal setting (namely, accompanied with visual contexts) compared to traditional purely textual pre-training. They show that indeed providing models with perceptual context is beneficial for training language models. Authors claim this evidence to correspond with the theory of situated cognition introduced in [Greeno and Moore1993].

In this work, we propose the first methodological step for chronological interpretation of traditional transformer language models in the framework of LA.

3 Experimental setup

3.1 Models

We calculated the accuracy of two transformer models on 12 probing tasks. As we want to know how universal patterns of language acquisition in models are, we experiment with two different transformer architectures: BERT and T5. While BERT has only encoder layers, T5 includes both encoder and decoder layers. Therefore, embeddings from BERT come from the encoder, and T5 embeddings are calculated after going through decoder after encoder.

For this work, we use already published models with available checkpoints. It means that they were trained on different data and computational powers. Moreover, they were trained with different batch sizes (256 for MultiBERT and 32 for T5). However, we follow [Chiang et al.2020b, Liu et al.2021] and measure the training progress in iterations. The further comparison of the two models is indicative only.

MultiBERT [Sellam et al.2021] is based on BERT-base-uncased architecture, and it is the model of the same size (12 layers and embedding size 768). Unlike the original BERT [Devlin et al.2019], it was trained with 25 different seeds. The authors also released checkpoints of the first five models. We use the model with seed 0 in our experiments. MultiBERT was trained on both literary and non-fiction texts. The former comes from BookCorpus [Zhu et al.2015], which includes 11,038 books of 16 different genres. The non-fiction texts are taken from English Wikipedia collected by [Turc et al.2019].

T5-small model is trained within the `t5-experiments` framework.³ and follows the Hugging Face implementation of T5 [Raffel et al.2019]. It consists of 6 layers with 512 embedding size.

Following the previous language modeling works [Devlin2018] [Bojanowski et al.2017], we use the Wikipedia data to train the model. The raw Wikipedia data is provided by The Pile project [Gao et al.2020], contains $\approx 19Gb$ of expository prose texts of various domains, and is treated as a language modeling dataset of reasonably well quality.

Baseline As a baseline, we use the method described in [Hewitt and Liang2019b]. We train logistic regression on top of embeddings of models mentioned above with shuffled class labels.

3.2 Probing tasks

Probing tasks come from several datasets published earlier: SentEval [Conneau et al.2018b], Morph Call [Mikhailov et al.2021], DisSent [Nie et al.2019], DiscoEval [Chen et al.2019], and BLiMP [Warstadt et al.2020]. The class balance of first eight tasks is illustrated with Figure 4 in Appendix. We choose these tasks to make our results comparable to other works on probing.

³<https://github.com/yurakuratov/t5-experiments>

Task	Sentence examples	Labels
Subject number	<i>Her employer had escaped with his wife for several afternoons this summer.</i>	NN
	<i>Your Mackenzie in-laws have sordid reputations few decent families wish to be connected with .</i>	NNS
Person	<i>So I still can recomend them but prepare pay twice as much as they tell you initially .</i>	has a person marker
	<i>The service was friendly and fast , but this just does nt make up for the lack - luster product .</i>	does not have a person marker
Tree depth	<i>We have done everything we can for her .</i>	11
	<i>Alvin Yeung of Civic Party</i>	3
Top constituents	<i>Did it belong to the owner of the house ?</i>	VBD_NP_VP_.
	<i>How long before you leave us again ?</i>	WHNP_SQ_.
Connectors	<i>He 'd almost forgotten about that man . Sarah had somehow brought him back , just as she had his nightmares .</i>	but
	<i>I let out a slow , careful breath . Felt tears sting my eyes .</i>	and
Sentence position	<i>Quneitra Governorate (/ ALA-LC : “ Muḥāfaat Al-Qunaytrah “) is one of the fourteen governorates (provinces) of Syria . The governorate had a population of 87,000 at the 2010 estimate . Its area varies , according to different sources , from 685 km ² to 1,861 km ² . It is situated in southern Syria , notable for the location of the Golan Heights . The governorate borders Lebanon , Jordan and Israel .</i>	1
	<i>The bossom and the part of the xhubleta covered by the apron are made out of crocheted black wool . The bell shape is accentuated in the back part . The xhubleta is an undulating , bell-shaped folk skirt , worn by Albanian women . It usually is hung on the shoulders using two straps . Part of the Albanian traditional clothing it has 13 to 17 strips and 5 pieces of felt .</i>	4
Penn Discourse Treebank	<i>Solo woodwind players have to be creative,they want to work a lot</i>	Pragmatic Cause
	<i>The U.S. , along with Britain and Singapore , left the agencyl, its anti-Western ideology , financial corruption and top leadership got out of hand</i>	List
Discourse Coherence	<i>Within the fan inlet case , there are anti-icing air bosses and probes to sense the inlet pressure and temperature .’, ‘High speed center of pressure shifts along with fin aeroelasticity were major factors . At the 13th (i.e .’, ‘the final) compressor stage , air is bled out and used for anti-icing . The amount is controlled by the Pressure Ratio Bleed Control sense signal (PRBC) . The “ diffuser case “ at the aft end of the compressor houses the 13th stage .</i>	a text is not coherent
	<i>This experience of digital circuitry and assembly language programming formed the basis of his book “ Code : The Hidden Language of Computer Hardware and Software ” . Petzold purchased a two-diskette IBM PC in 1984 for \$ 5,000 . This debt encouraged him to use the PC to earn some revenue so he wrote an article about ANSL.SYS and the PROMPT command . This was submitted to PC Magazine for which they paid \$ 800 . This was the beginning of Petzold ’s career as a paid writer . In 1984 , PC Magazine decided to do a review of printers .</i>	a text is coherent

Table 1: Examples of tasks

Task	Acceptable sentence	Unacceptable sentence
Transitive	<i>The pedestrians question some people.</i>	<i>The pedestrians wave some people.</i>
Passive	<i>Tracy isn't fired by Jodi's daughter.</i>	<i>Tracy isn't muttered by Jodi's daughter.</i>
Principle A c command	<i>This lady who is healing Charles wasn't hiding herself.</i>	<i>This lady who is healing Charles wasn't hiding himself.</i>
Adjunct Island	<i>Who does John leave while alarming Beverly?</i>	<i>Who does John leave Beverly while alarming?</i>

Table 2: BLiMP Minimal pairs examples

As we want to show another perspective on language acquisition, we balance classifier probing tasks with BLiMP tasks. As BLiMP only covers morphology and syntax, all discourse-based tasks are evaluated with a classifier.

The datasets from Benchmark of minimal linguistic pairs (BLiMP) have a structure different from other tasks: every task includes pairs with minimal differences to illustrate one of the grammatical features of English. One sentence of the pair is grammatical, whereas another one is unacceptable. We chose four BLiMP tasks for our experiments: transitive and passive verbs, Principle A of C command, and Island effects. For the first two tasks, pairs have different verbs, where only one verb is transitive or can be used in a passive form. These tasks are categorized as morphological (see Table 2).

The second two tasks reflect syntactic effects on English. The Principle A task shows the use of reflexives. According to [Chomsky1981], a reflexive should have a local antecedent.

The task on Island effects tests a model's sensibility to syntactic order. Island is a structure from which a word can not be moved [Ross1967]. The sentence is unacceptable if a word is moved out of an island.

The tasks from other datasets are summarized below:

- **Subject number** (SentEval): this task is supposed to show how models acquire morphology. It is a binary classification task with labels NNS and NN (plural and singular number, respectively). The classifier should decide on a sentence class based on the number of sentence subjects.
- **Person** (Morph Call): this task is also morphological. It is a binary classification with labels 0 and 1, which signifies if a subject has a person marker or not.
- **Tree depth** (SentEval): this task contains six classes, which stands for the depth of the syntactic tree of a given sentence. Hence, this task is meant to show the level of syntax acquisition.
- **Top constituents** (SentEval): this multiclass task belongs to the group of syntactic tasks. The aim is to choose a class that includes constituents located right below the sentence (S) node.
- **Connectors** (DisSent): this dataset includes pairs of sentences originally connected with one of 5 prepositions, and the task is to choose the omitted preposition. It is supposed to show how models catch discourse relations.
- **Sentence position** (DiscoEval): this dataset contains sequences of 5 sentences, and the first sentence is placed in the wrong place. Therefore, the aim is to detect the original position of these sentences. This task is also meant to show models' accuracy in discourse.
- **Penn Discourse Treebank** (DiscoEval): the task is based on Penn Discourse Treebank annotation. The aim is to choose the right discourse relation between two discourse items from Penn Treebank.
- **Discourse coherence** (DiscoEval): this task is a binary classification with classes 1 and 0. Class 1 means that the given paragraph is coherent, and class 0 should be assigned to paragraphs with shuffled sentences.

3.3 Probing Methods

Sentence embedding classification: Token embeddings from transformer models are turned into sentence embedding via mean pooling. Then logistic regression classifies embeddings' sentences. This method is used with tasks from SentEval and Morph Call.

Positional sentence classification: For the Sentence Position task, we used the following method. First, we count sentence embeddings as described above. Then the difference between the first embedding and the other is calculated pair-wisely. The first embedding and its differences with others are concatenated and put as input to logistic regression.

Sentence embedding concatenation classification: We concatenated sentence embeddings for other discourse tasks, which were calculated as the mean of token embeddings. The concatenated sentence embeddings served as inputs for logistic regression.

Masking tasks: The probing task is based on the idea of masking language modeling. In a sentence, each word is masked, and then its probability is summed with other words' probabilities. The probability of an acceptable sentence should be higher than the probability of an unacceptable sentence. This method is for use for all tasks from BLiMP.

4 Results

4.1 Results of MultiBERT

The results of the experiments with the MultiBERT-base model are summarized in Figure 1. The model shows the best results on Subject Number and Person tasks. The classification of PDTB relations, Tree depth, and Principle A acceptability are performed with the worst accuracy.

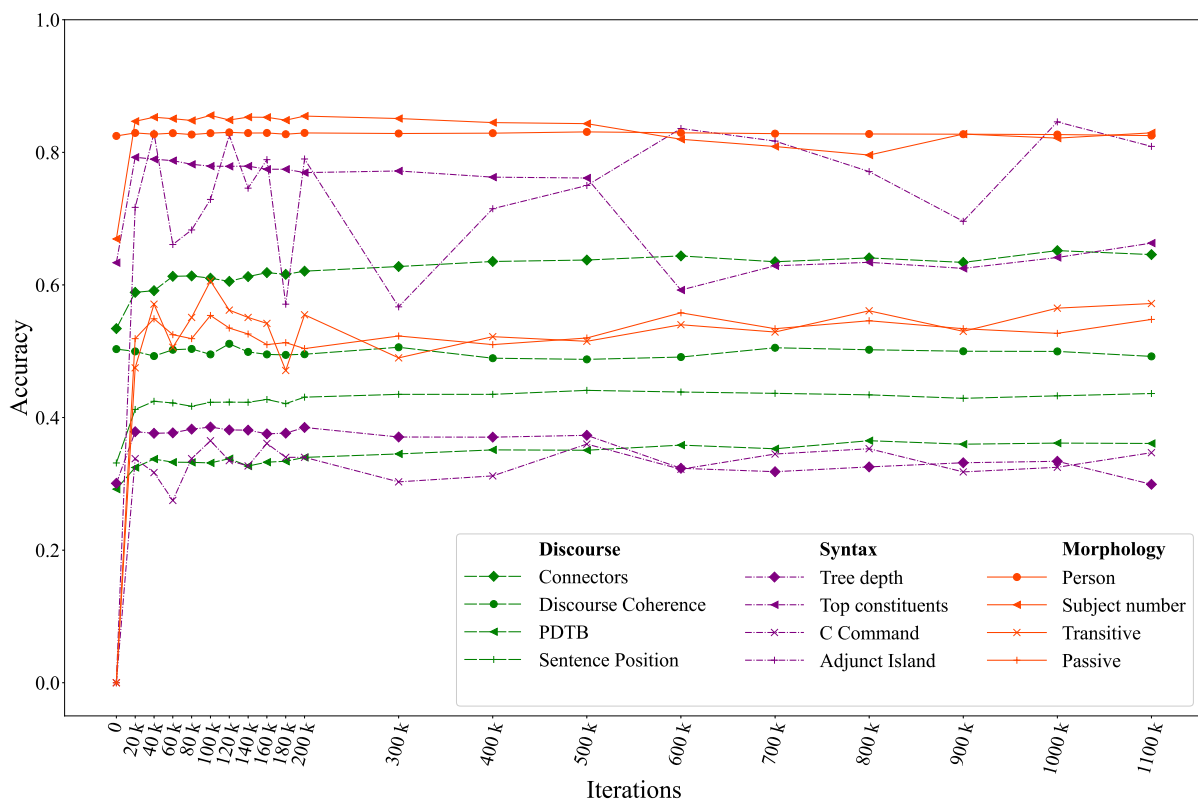


Figure 1: Comparison of MultiBERT results on different tasks. How to read this figure: from left to right, on the X axis, we see results of intermediate evaluation on the task during model pre-training. Each iteration is equal to 25,600,000 sentences for MultiBERT and 3,200,000 sentences for T5. The Y-axis shows the accuracy metric on the tasks. Tasks are shown in the legend in different colors. As we can see, in the process of model pre-training, there already is a gradual increase in accuracy in tasks related to morphology (shown in orange) in the early stages. The information in the model embeddings stabilizes fairly quickly and remains stable from the 20,000th training step. The same can not be said for tasks related to syntax (shown in purple): their quality remains unstable and fluctuates quite a lot during pre-training. Discourse tasks (green) remain stable at a low-quality level from the start and tend to improve the metrics very slowly.

As seen from Figure 1, accuracy of models stop changing after 600,000 iterations. However, there is a significant difference between tasks from BLiMP and other datasets. For example, the performance

on the Adjunct Island task remains unstable during the whole period of iterations. Another difference between these tasks lies in the quality of the models. It is illustrated with tasks grouped as “morphological”: Subject Number and Person tasks, which use logistic regression on MultiBERT embeddings, are solved much better than Transitive and Passive verbs. However, as follows from the plot, it is hard to group tasks based on the absolute value of accuracy.

The changing dynamics provide another perspective. From this point of view, all tasks grouped as “discourse” show a similar feature: unlike others, their quality does not fluctuate but rather slightly grows across the training time. On other tasks, models increase the quality during the first iterations. “Syntactic” tasks tend to change even during later iterations. However, it is not a strict rule, and some tasks show similar behavior to “morphological” ones.

4.2 Results of T5

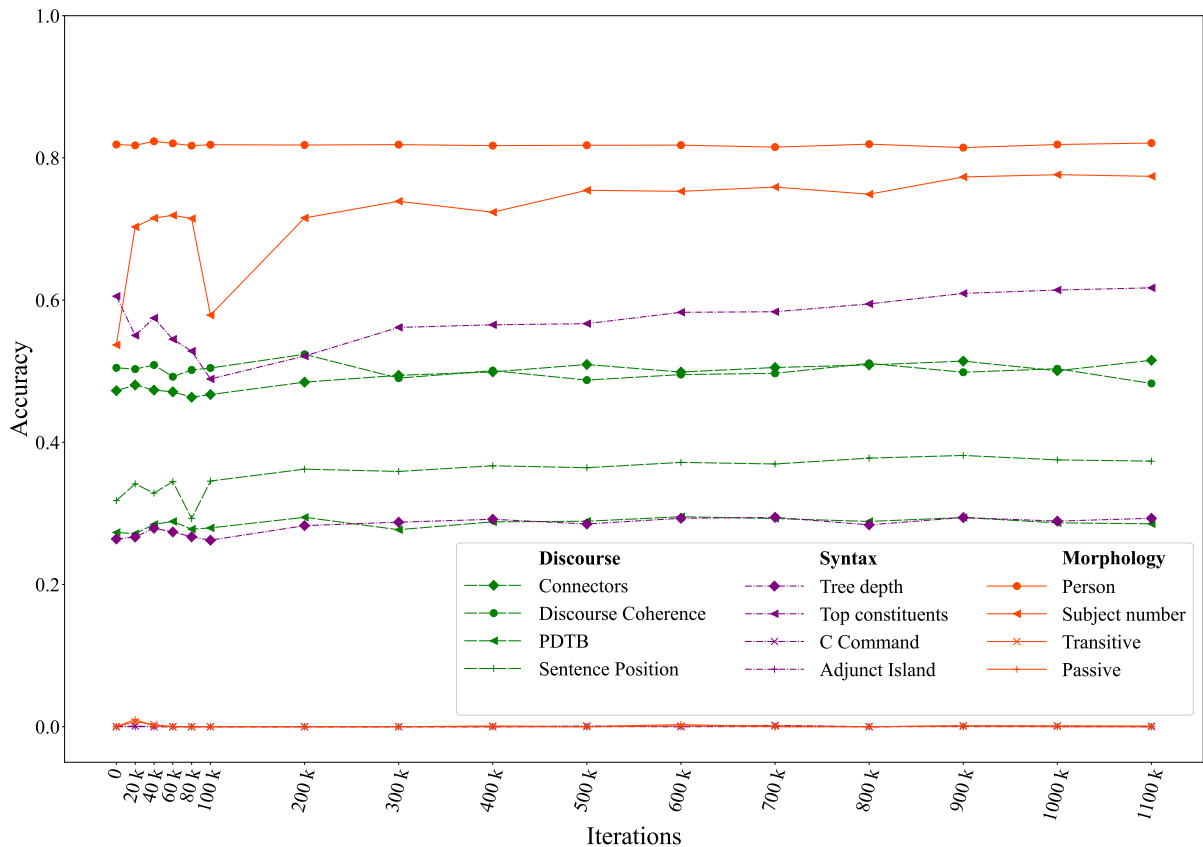


Figure 2: Comparison of T5 results on a different task. How to read this figure: from left to right, on the X axis, we see results of intermediate evaluation on the task during model pre-training. Each iteration is equal to 25,600,000 sentences for MultiBERT and 3,200,000 sentences for T5. The Y-axis shows the accuracy metric on the tasks. Tasks are shown in the legend in different colors. As we can see, in the case of the T5 model, the task quality seems to be more stable from the beginning, with a few exceptions like subject number classification. Most of the tasks show the slow yet gradual growth of the metrics, but somehow not the verb transitivity classification.

Due to the architecture, the significant difference in T5 results is the zero-close quality on BLiMP datasets. Except for these tasks, the quality of T5 is similar to MultiBERT. The best performance is on the Person task, and the worst quality is shown on PDTB relation classification and Tree depth.

Unlike MultiBERT, we first used the available checkpoints of T5 with a step of 100,000 iterations. Then we trained a new model on the same resources and texts, but it might have a better initialization, which affected the final results.

Similar to MultiBERT, discourse tasks show almost no significant change and slow growth, whereas the model increases its results on syntactic and morphological tasks during the first 100,000 iterations.

4.3 Comparison of models

We described the surface results of models' performance and now can deep into more detailed results. The results described above should be considered relative. To illustrate how much information models acquire during these iterations, we compare them to final models. As the process of training T5 was not finished, we compared this model with the original T5. As seen from Figure 3, MultiBERT scores are close to the results of the final checkpoint. Hence, there is no need to look at later iterations. The comparison with the original T5 shows that the model we use performs worse due to the smaller resources it was trained on. Therefore, the difference in quality should not be explained by the difference in architecture.

However, we should consider that some tasks are performed with the same quality as embeddings with shuffled labels (Discourse coherence and Person). Moreover, T5 does not perform much better than the Penn Discourse Treebank relations baseline. Consequently, models encounter difficulty with discourse tasks.

Furthermore, MultiBERT and T5 show similar learning trajectories on several tasks, such as Connectors and Sentence Position tasks. Another key feature shared by the two models is the termination of increases between 500,000 and 600,000 iterations. Despite the fact that models vary in size and training process, they show some similarities in probing tasks. Hence, the acquisition generally does not depend on the model architecture.

5 Discussion

Our results show that linguistic information is acquired fast, before 600,000 training iterations. It corresponds to results of other researchers [Chiang et al.2020a, Liu et al.2021] that independently showed similar results on a fast acquisition of linguistic features. However, discourse is not fully acquired by the end of the observed training period compared to the baseline results. The difference between syntactic and morphological tasks is insignificant. It correlates with ideas in morphosyntax. Although we can not prove that morphology and syntax are regarded as the same layer in models, we can make a less strict statement that models acquire all grammatical units simultaneously.

BLiMP gives another perspective on the process of acquisition. MultiBERT results remain unstable for a longer period than similar tasks with classifiers. It might indicate the difference between two different approaches to probing. However, from the linguistic point of view, BLiMP includes more difficult linguistic feature cases, while SentEval tasks test more basic knowledge. Hence, it could explain worse results.

T5 architecture does not allow to use of this dataset in the same way as for MultiBERT since Masking Language Modeling and T5 generation are different tasks. We leave for further research an adaptation of this dataset for T5.

5.1 Human language acquisition results

Many of the linguistic features used in probing tasks have been well studied in terms of their promptness and ease of acquisition by English speakers. First of all, it concerns morphology and syntax. Markers such as a person, number, and gender are acquired very early by children: before age two [Clark2017]. Of course, in languages besides English, the acquisition of these features varies: if the feature is marked consistently with one affix and no morphological alternation, children seem to find it easier to acquire. It is shown that the earlier mastery of case marking is present in languages like Hungarian and Turkish but not in German or Serbo-Croatian [Slobin1985].

Discursive features are acquired by children much later. Studies like [Pearson2003] show that child's texts become more complex and decontextualized with age. Also, texts produced by children gradually progress in achieving more cohesion through "referential and semantic links that bridge across sentences; they achieve coherence through a global hierarchical structure". The discourse in these conversations

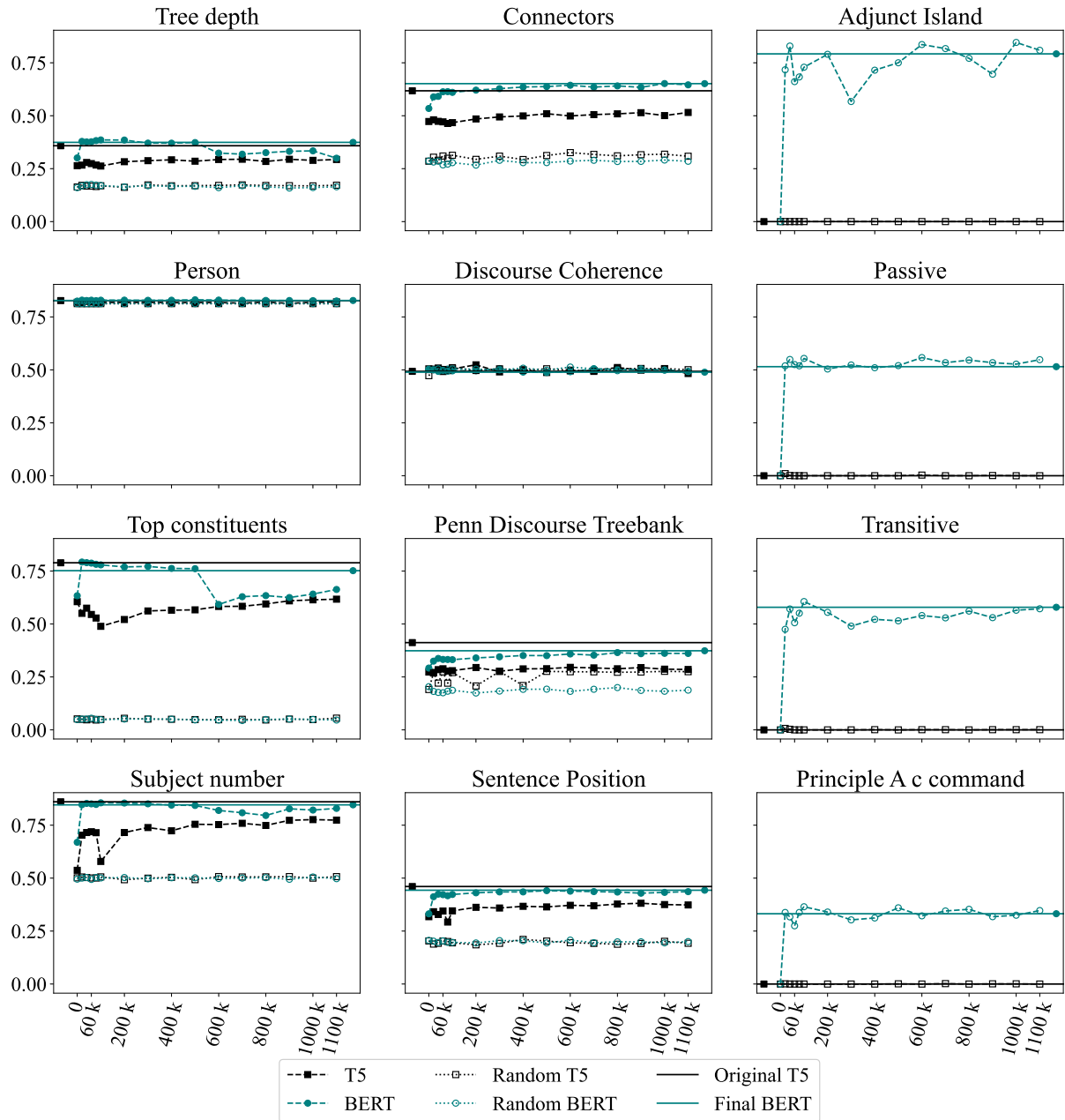


Figure 3: Performance on models on different tasks. The detailed task-wise comparison shows the difference in T5 and MultiBERT training results. The models are compared to the final available checkpoints of the models (*Original T5*, *Final BERT*) and with the random baseline.

between toddlers is tentative when neither side can be reliably significant. A longitudinal study of dialogues between two little girls aged 4 to 6, [McTear1985] traced the emergence of more and greater thematic continuity in their conversation as utterances began to play the dual role of responding to the preceding utterance as well as enabling further conversation. However, [Dorval et al.1984] showed that second-graders (eight-year-olds) were almost as likely to give unconditioned responses as conditionals, with no significant improvement seen until fifth grade.

Regarding the requirements that the sphere of language acquisition imposes on children, one can very carefully assess the limit in which the language models under consideration lie in terms of their abilities: their embeddings correspond to the level of language proficiency in a child under 11 years old.

6 Conclusion

Encoder and encoder-decoder language models have increased importance in tasks requiring understanding the natural language. The probing methodology we presented allows analyzing the changes within the language model during training, from epoch to epoch. The overall results of the work show that:

- T5 does not give any results on BLiMP due to the generation algorithm. Most tasks show that T5 acquires basic morphological and syntactic features and some discourse features.
- MultiBERT shows results close to the trained model starting from 100,000 iterations. MultiBERT does not improve its quality on some discourse tasks compared to randomly labeled embeddings. However, it could be said that MultiBERT acquires each level to some extent.
- Both T5 and MultiBERT demonstrate comparable results regarding the quality of the language level acquisition. As we can not distinguish the factors between these results (whether this is the model's architecture, the training corpora, or both), we present these results 'as is' for the researchers that use them in the downstream tasks.
- Recording such results during training could save a lot of computational resources and time for interpreting the results, including downstream tasks. There are understandable context length limitations that prevent, for example, learning the discourse tasks. However, the results of the T5 model compared to the random embeddings on some tasks were unexpected or lower than expected.
- As the results show, the easiest tasks for the models tend to be morphology and syntax-related. These language level results are correlated and show a similar learning trajectory. Unlike morphology and syntax tasks, results on discourse-based tasks tend to be low, therefore, there is not enough evidence to claim that discourse has been learned.
- Using language acquisition as a trace can benefit in comparing general human language ability and modern language modeling methods. Drawing a border from above on the results on discursive tasks, we can say that in the current realities, the models do not pass the bar that 8-year-old children pass.

We welcome both NLP and language acquisition research communities to reproduce the experimental setup and use the presented approach while training other architectures and contribute to formulating the better and more complex tasks correlated with language learning.

References

- Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. // *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, P 1–10, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.
- Yonatan Belinkov. 2016. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, P 1–12.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Pavel Caha. 2009. The nanosyntax of case.
- Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. // *Proc. of EMNLP*.
- Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020a. Pretrained language model embryology: The birth of albert. *arXiv preprint arXiv:2010.02480*.
- Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020b. Pretrained language model embryology: The birth of ALBERT. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 6813–6828, Online, November. Association for Computational Linguistics.

- Noam Chomsky. 1981. Lectures on government and binding (dordrecht: Foris). *Studies in generative grammar*, 9.
- Eve V Clark. 2017. Morphology in language acquisition. *The handbook of morphology*, P 374–389.
- Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $\&!#^*$ vector: Probing sentence embeddings for linguistic properties. // *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 2126–2136, Melbourne, Australia, July. Association for Computational Linguistics.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018b. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.
- Mary Dalrymple. 2001. *Lexical functional grammar*. Brill.
- Jill G De Villiers and Thomas Roeper. 2011. *Handbook of generative approaches to language acquisition*, volume 41. Springer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Jacob Devlin. 2018. Multilingual bert.
- Bruce Dorval, Carol O Eckerman, and Susan Ervin-Tripp. 1984. Developmental trends in the quality of conversation achieved by small groups of acquainted peers. *Monographs of the Society for Research in Child Development*, P 1–91.
- David Embick and Rolf Noyer. 2007. Distributed morphology and the syntax/morphology interface. *The Oxford handbook of linguistic interfaces*, 289324.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- James G Greeno and Joyce L Moore. 1993. Situativity and symbols: Response to vera and simon.
- John Hewitt and Percy Liang. 2019a. Designing and interpreting probes with control tasks. // *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 2733–2743, Hong Kong, China, November. Association for Computational Linguistics.
- John Hewitt and Percy Liang. 2019b. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.
- Xuân-Nga Cao Kam, Iglia Stoynezhka, Lidiya Tornyova, Janet D Fodor, and William G Sakas. 2008. Bigrams and the richness of the stimulus. *Cognitive science*, 32(4):771–787.
- John D Lewis and Jeffrey L Elman. 2001. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. // *Proceedings of the 26th annual Boston University conference on language development*, volume 1, P 359–370. Citeseer.
- Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*.
- Christopher D Manning. 2015. Last words: Computational linguistics and deep learning.
- R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. // *Proceedings of the Third Black-boxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, P 217–227, Online, November. Association for Computational Linguistics.
- Michael McTear. 1985. *Children's conversation*. B. Blackwell.

- Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. 2021. Morph call: Probing morphosyntactic content of multilingual transformers. // *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, P 97–121, Online, June. Association for Computational Linguistics.
- Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 4497–4510.
- Alexander Ororbia, Ankur Mali, Matthew Kelly, and David Reitter. 2019. Like a baby: Visually situated neural language acquisition. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 5127–5136, Florence, Italy, July. Association for Computational Linguistics.
- Barbara Zurer Pearson. 2003. Language acquisition: Discourse, narrative and pragmatics. *Disertasi. USA: Department of Communication Disorders*.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. *CoRR*, abs/2004.03061.
- Amy Prefors, Terry Regier, and Joshua B Tenenbaum. 2006. Poverty of the stimulus? a rational approach. // *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.
- Colin Raffel, Noam 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 preprint arXiv:1910.10683*.
- Florencia Reali and Morten H Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29(6):1007–1028.
- John Robert Ross. 1967. Constraints on variables in syntax.
- Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.
- Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D’Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, et al. 2021. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.
- Dan I Slobin. 1985. Crosslinguistic evidence for the language-making capacity. *The crosslinguistic study of language acquisition*, 2:1157–249.
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv e-prints*, P arXiv:1905.06316, May.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.
- Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 183–196, Online, November. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohanney, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. // *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, P 359–361, Brussels, Belgium, November. Association for Computational Linguistics.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. // *Proceedings of the IEEE international conference on computer vision*, P 19–27.

Appendix

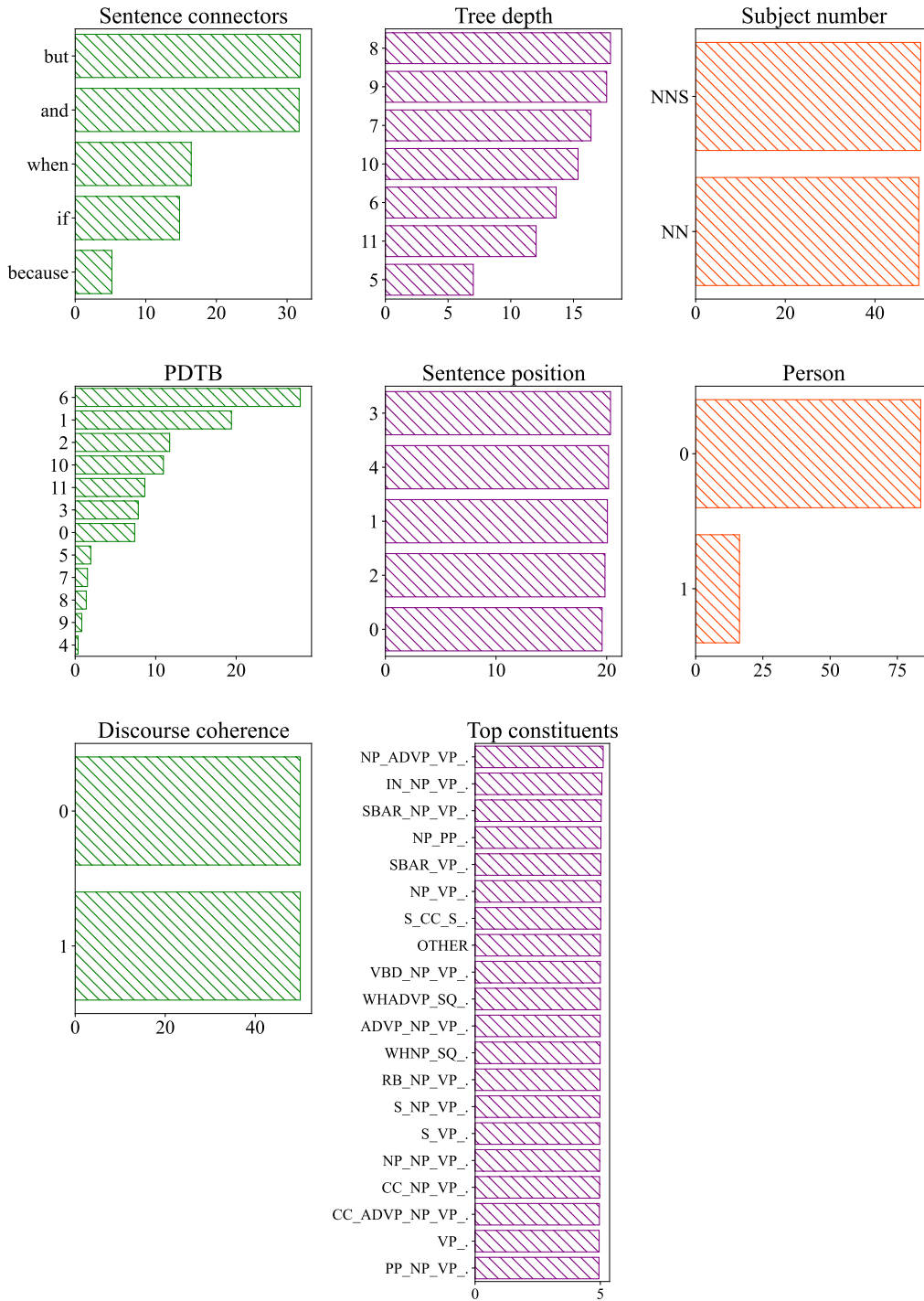


Figure 4: The class balance of datasets

Collocation2Text: Controllable Text Generation from Guide Phrases in Russian

Vychegzhanin S. V.
Vyatka State University,
Kirov, Russia
vychegzhaninsv@gmail.com

Kotelnikov E. V.
Vyatka State University,
Kirov, Russia
kotelnikov.ev@gmail.com

Abstract

Large pre-trained language models are capable of generating varied and fluent texts. Starting from the prompt, these models generate a narrative that can develop unpredictably. The existing methods of controllable text generation, which guide the narrative in the text in the user-specified direction, require creating a training corpus and an additional time-consuming training procedure. The paper proposes and investigates *Collocation2Text*, a plug-and-play method for automatic controllable text generation in Russian, which does not require fine-tuning. The method is based on two interacting models: the autoregressive language ruGPT-3 model and the autoencoding language ruRoBERTa model. The idea of the method is to shift the output distribution of the autoregressive model according to the output distribution of the autoencoding model in order to ensure a coherent transition of the narrative in the text towards the guide phrase, which can contain single words or collocations. The autoencoding model, which is able to take into account the left and right contexts of the token, “tells” the autoregressive model which tokens are the most and least logical at the current generation step, increasing or decreasing the probabilities of the corresponding tokens. The experiments on generating news articles using the proposed method showed its effectiveness for automatically generated fluent texts which contain coherent transitions between user-specified phrases.

Keywords: text generation; GPT; BERT

DOI: 10.28995/2075-7182-2022-21-564-576

Collocation2Text: Управляемая генерация текста по направляющим выражениям на русском языке

Вычегжанин С. В.
Вятский государственный
университет,
Киров, Россия
vychegzhaninsv@gmail.com

Котельников Е. В.
Вятский государственный
университет,
Киров, Россия
kotelnikov.ev@gmail.com

Аннотация

Крупные предобученные языковые модели способны создавать разнообразные и связанные тексты. Эти модели, отталкиваясь от заправки, генерируют повествование, которое может уходить в непредсказуемом направлении. Существующие методы управляемой генерации текста, направляющие повествование в тексте в заданном пользователем направлении, требуют создания обучающего корпуса и выполнения трудоемкой процедуры дополнительного обучения. В статье предложен и исследован plug-and-play метод автоматической управляемой генерации текстов на русском языке *Collocation2Text*, не требующий процедуры дообучения (fine-tuning). Метод основан на взаимодействии двух моделей: авторегрессионной языковой модели ruGPT-3 и автокодирующей языковой модели ruRoBERTa. Идея метода заключается в смещении выходного распределения авторегрессионной модели в соответствии с выходным распределением автокодирующей модели для обеспечения связанного перехода повествования в тексте к направляющему выражению, которое может содержать одиночные слова или словосочетания. Автокодирующая модель, способная учитывать левый и правый контекст токена, «подсказывает» авторегрессионной модели, какие токены являются наиболее и наименее логичными на текущем шаге генерации, повышая или понижая вероятности соответствующих токенов. Эксперименты по генерации новостных статей с использованием предложенного метода показали его эффективность для автоматического создания гладких текстов, содержащих связанные переходы между заданными пользователем выражениями.

Ключевые слова: генерация текстов; GPT; BERT

1 Introduction

Natural language generation (NLG) is a field of computational linguistics that deals with the construction of computer systems which can generate understandable texts in English or other human languages [8, 29]. At an early stage of development, language generation was carried out using rule-based models, hidden Markov model, as well as shallow models trained on sparse multidimensional features. Due to large neural network language models, in particular, models based on the Transformer architecture [33], trained on dense vector representations, automatically generated texts showed a new level of quality. Modern neural network models such as GPT-2 [28] and GPT-3 [3] can generate impressively realistic human-like texts.

NLG has a wide range of applications, including generation of answers to user questions in dialogue systems (chatbots) [31] and question-answer systems [6], generation of stories [1], poetry [22], news articles [32], product and service reviews [7].

Making text generation controllable is an important fundamental issue in NLG. For example, when generating stories, it is necessary to control the storyline or the completion of the story. Controllable Text Generation (CTG) is the task of generating a natural language text with given attributes [26]. Topic, sentiment, keywords, entities, events, etc. can be considered as attributes.

Soft and hard controls can be applied over text generation models. Soft control aims at ensuring, e.g., the desired sentiment or general topic of the generated text. The aim of hard control is to ensure that some explicit constraints, e.g., specific keywords are contained in the text. Figure 1 shows an example of CTG with hard control when generating a story with keywords from the storyline in a given order.

Prompt	The student needed a laptop for study.
Storyline	store → bought → installed → happy → project
Generated text	The student needed a laptop for study. He came to the computer store . The student chose and bought a new laptop. He installed the necessary software. The student is happy to do a study project using a new laptop.

Figure 1: Example of controllable story generation with hard control

The problem of most existing models with soft [4, 14] and hard [5, 36] controls lies in the need to create a training corpus and perform a training procedure. Creating such a corpus and additional training of the model is difficult, expensive and time consuming. This paper proposes to overcome this problem by developing a plug-and-play method that can be applied to pre-trained large language models. It should be noted that currently there is a lack of research on controllable text generation in Russian, thus, the proposed method was tested on the Russian language models.

The idea of the method is to shift the output distribution of the autoregressive language model (ruGPT-3) according to the output distribution of the autoencoding language model (ruRoBERTa). The method provides a coherent transition of the narrative in the text to the guide phrase. It is a plug-and-play method, i.e., it can be used with any autoregressive and autoencoding language models that have vocabularies built using the same encoding algorithm. Within our study, the experiments on generating news articles proved the effectiveness of the proposed method for creating a coherent text from the list of guide phrases.

The contributions of this paper are the following:

- we propose a method of controllable text generation *Collocation2Text* that generates texts according to a user-specified list of guide phrases, which can be either single words or collocations;
- we apply this method to the Russian language;
- we conduct experiments with article generation in order to confirm the effectiveness of the proposed method.

2 Previous Work

This section discusses the existing methods for text generation in relation to the story generation task, which is of the main interest in our study. A story is a description of real or imaginary characters and events generated to achieve one or more goals, e.g., to entertain or educate [1]. The difficulty of the neural network story generation task consists in generating a coherent and fluent story that is much longer than a short input user-specified prompt. Story generators can be classified into three categories:

- structural models which generate a structured story by dividing the story into slots following a given scheme [12, 16];
- planning-based models which generate a story as a chain of causally connected events to pursue a final goal [9, 13];
- machine learning models which learn the conditional probability distribution between story events from a story corpus [18, 34].

Currently, the greatest success in automatic story generation has been achieved by machine learning models, in particular, deep neural networks. A number of previous studies in this direction should be noted. Jain et al. [11] used a recurrent neural network with an attention mechanism and gated recurrent units. As input, the neural network received standalone textual descriptions describing an event or scene and converted them into coherent summaries. However, according to the results of the experiments, the summaries obtained were not completely semantically related to the input descriptions. To overcome this disadvantage, Fan et al. [5] decomposed story generation into two stages. During the first stage, they generated the story premise representing the structure of the story using the convolutional language model. At the second stage, the sequence-to-sequence model was used to create a story that followed the premise.

Peng et al. [24] developed a framework to control story ending valence (happy or sad ending) and a storyline based on a recurrent neural network generation model with an attention mechanism. The experiments showed that through introducing storylines the coherence of the story is improved compared to uncontrolled generation models.

Yao et al. [36] proposed another hierarchical story generator that combines storyline planning and text generation. During the learning stage, for each story from the corpus, a storyline was built from the most important words in each sentence using the RAKE algorithm [30]. Then, the storyline was converted to a text using the sequence-to-sequence model. It was proved that with explicit storyline planning the generated stories are more diverse and coherent than those generated without creating a full plan.

The examples of texts generated by the GPT-2 model [28] showed that large pre-trained language models are able to generate texts similar to human-written texts. These models have been actively used by researchers to generate stories. Dathathri et al. [4] proposed a plug-and-play method that trains an external attribute discriminant model. The gradients from the attribute model “push” the hidden activation of the pre-trained language model on the Transformer architecture to guide the target text generation.

Pascual et al. [23] proposed another plug-and-play method of controllable text generation, which can be applied to the existing autoregressive language model without additional training. The authors created a strategy for controlled decoding, in which the generated text contains words from the given guide word sequence. The idea of the method is that in the process of generating the next word the output distribution of the language generation model is shifted to the semantic space of the guide word. The degree of similarity between the tokens of the language model and the guide word is calculated as the cosine measure in the vector space word2vec [21] or GloVe [25]. The method not only encourages the guide word to occur explicitly, but also encourages the model to generate the appropriate context for the guide word to occur.

Recently, significant progress in the control of autoregressive models has been made using prompt-tuning. For example, Li and Liang [17] proposed prefix-tuning, which keeps language model parameters frozen but optimizes a small continuous task-specific vector. Qin and Eisner [27] proposed to learn a mixture of soft prompts. Liu et al. [19] designed an optimized and adapted implementation of deep prompt tuning for generation.

Our study is consistent with [23]. The difference of our method is that the output distribution of the autoregressive language generation model is not shifted to the guide word in the word2vec or GloVe vector space, but is summed with the output distribution of another autoencoding language model. In addition, the proposed method allows using single words and collocations in the guide sequence.

3 A method of controllable text generation

The proposed *Collocation2Text* method can be applied to any autoregressive language model for which the probability of a sequence of linguistic units (characters, tokens, words, sentences) $X = \{x_1, \dots, x_n\}$ is decomposed using the chain rule:

$$p(X) = \prod_{i=1}^n p(x_i | x_1, x_2, \dots, x_{i-1}) = \prod_{i=1}^n p(x_i | X_{<i}). \quad (1)$$

We consider models p that assign a probability to all sequences X in the space of strings $Y(V, n_{max}) = V^{n_{max}}$, where V is the model vocabulary and n_{max} is the maximum sequence length.

The task of text generation based on the autoregressive language model is to decode sequences of linguistic units from the distribution p . In the generation process, a score function and a decoding algorithm are of the most importance. The score function assigns some degree of significance for generating the resulting text to each sequence of linguistic units. The score function depends on the sequence probability and can be modified according to the requirements for the generation method. Formally, the score function of the autoregressive language model can be defined as a map from strings generated under the model vocabulary to a real number and written in the form $score_{alm}(\cdot | X_{<i}): V^i \rightarrow \mathbb{R}$. For such models, the default score function is usually the log-probability:

$$score_{alm}(\cdot | X_{<i}) = \log p(\cdot | X_{<i}). \quad (2)$$

A decoding algorithm is a class of algorithms that decode a text according to the score function. Examples of such algorithms are beam search [20] and nucleus sampling [10]. We apply top- K sampling, which consists in choosing a token from the K tokens with the largest values of the score function.

Autoregressive models (e.g., GPT-3) are trained to predict the next token in a sequence by the previous context. For such models, text generation is a natural application. However, their disadvantage in the task of controllable text generation according to given guide phrases is that in the process of predicting the next token the context following this token is not taken into account. In contrast, autoencoding (masked) language models (e.g., BERT) are trained to predict masked tokens in the text and reconstruct the original text. Such models take into account the left and right contexts of the predicted token. We combine the advantages of the two classes of models as follows. The autoencoding language model, given the left and right contexts of the token, helps the autoregressive model change the output distribution, increasing the probability of more suitable tokens and decreasing the probability of less suitable tokens for the coherent transition to the guide phrases.

In this paper, we propose the *Collocation2Text* method, which modifies the autoregressive model score function (2) by shifting its values based on the output distribution of the autoencoding model. We consider a probabilistic language generator p , a guide phrase $W = \{w_1, \dots, w_m\}$ which should occur in the generated text and a sequence of tokens $X = \{x_1, \dots, x_{i-1}, x_i\}$. We propose to modify the score function for the current token x_i as follows:

$$score'_{alm}(x_i | X_{<i}, W) = score_{alm}(x_i | X_{<i}) + score_{mlm}(x_i | X_{<i} \cup W), \quad (3)$$

where $score_{mlm}(x_i | X_{<i} \cup W)$ is a score function for the masked language model, $X_{<i} \cup W = \{x_1, \dots, x_{i-1}, w_1, \dots, w_m\}$.

In equation (3), the output score is calculated as the sum of the score of the autoregressive model based on the left context of token x_i ($X_{<i}$) and the score of the autoencoding model based on the left ($X_{<i}$) and right (W) contexts of token x_i .

In order to increase the probability of the first token w_1 from the guide phrase W in the text, the value of the score function for w_1 is modified as follows

$$score'_{alm}(w_1 | X_{<i}, W) = score'_{alm}(x_K | X_{<i}, W) + \lambda_i \cdot \alpha_i \cdot \Delta_i, \quad (4)$$

where λ_i is a hyperparameter indicating the strength of the score shift of the token $w_1 \in W$ at the i -th generation step, α_i and Δ_i are parameters determined by equations:

$$\alpha_i = \frac{s_{w1} - s_{min}}{s_{max} - s_{min}}, \quad (5)$$

$$\Delta_i = s_{max} - s_K, \tag{6}$$

where $s_{w_1} = score'_{alm}(w_1|X_{<i}, W)$ is the value of the score function for the first token w_1 of the guide phrase W ; $s_{min} = \min(score'_{alm}(x_i|X_{<i}, W))$ is the minimum value of the score function; $s_{max} = \max(score'_{alm}(x_i|X_{<i}, W))$ is the maximum value of the score function; $s_K = score'_{alm}(x_K|X_{<i}, W)$ is the value of the score function for the last token on the list of top- K tokens in the top- K sampling decoding strategy.

Thus, we put the token w_1 in the list of top- K tokens. The position of w_1 on the top- K list depends on the initial relative position α_i of this token.

Figure 2 shows a scheme of increasing the probability of the first token w_1 from the guide phrase W at the i -th step of the top- K sampling decoding strategy according to equation (4).

The coefficient λ_i determines the speed of the appearance of the guide phrase in the text. As λ_i increases, the first token of this phrase rises in top- K . The value of the coefficient λ_i increases at each i -th step of the generation process according to the equation:

$$\lambda_i = \lambda_0(i - i_n), \tag{7}$$

where λ_0 is the initial value of the shift; i_n is the number of the step at which the previous guide phrase appeared.

The proposed method ensures the interaction between neural network models of two types. The autoencoding model, which is able to take into account the left and right contexts of the token, “tells” the autoregressive model which tokens are the most logical at the current decoding step, shifting the values of the score function according to equation (3). Additionally, at each decoding step, the value of the score function for the first token of the guide phrase W is increased and thus the probability of choosing this token is increased. After occurring in the text, the token w_1 is replaced by the whole phrase W .

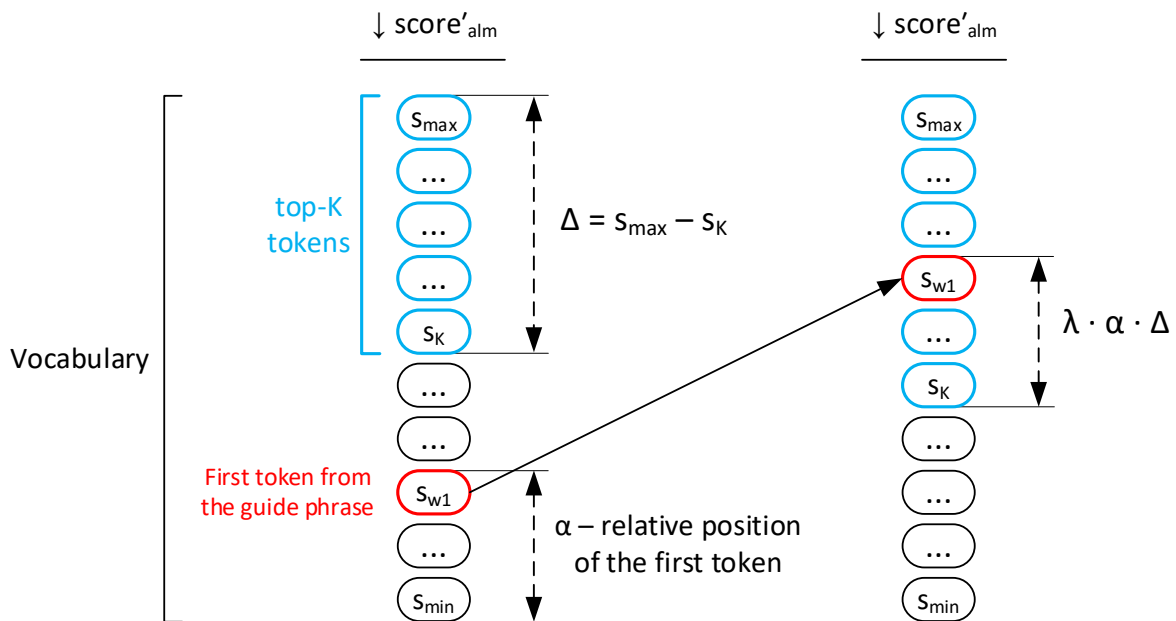


Figure 2: Scheme for increasing the probability of the first token of the guide phrase according to equation (4)

To illustrate the proposed method, we consider a text at an i -th generation step and a guide phrase, separated by a special $\langle \text{mask} \rangle$ token (Figure 3).

Мэр рассказал о планах благоустройства городских парков. Сегодня в городе <mask> очередная пресс-конференция

The mayor told about the plans for the improvement of city parks. Today in the city <mask> another press conference

Figure 3: Text example at the i -th generation step (Russian and English versions)

In Figure 3, the sentence used as a prompt for the autoregressive model is highlighted in blue, the guide phrase is highlighted in orange. By the i -th generation step, the autoregressive model has already generated the phrase “Сегодня в городе” (“Today in the city”). At the i -th step, as input, the model receives the text located to the left of the <mask> token and returns the distribution $score_{atm}$ over the tokens of its vocabulary. The autoencoding model takes the whole text as input, including the <mask> token and the guide phrase, returning the distribution $score_{mlm}$ over the tokens of its vocabulary. Then the distributions returned by the models are summed according to equation (3), and the value of the score function for the first token of the guide phrase is also increased according to equation (4). The resulting output distribution is subjected to the top- K sampling decoding algorithm that selects a token from the model vocabulary to be placed instead of the <mask> token. In the next step, the <mask> token moves one position to the right. The guide phrase is inserted until it occurs in the text.

4 Experimental Setup

In the experiments, ruGPT3Large¹ (760 million parameters) was used as an autoregressive language model, which is a Russian-language adaptation of the GPT-2 model [28], and ruRoBERTa-large² (355 million parameters) was used as an autoencoding language model.

To check whether the word generated by the ruGPT3Large model is on the list of guide phrases, we compared the lemma of this word and the lemma of the first word of the guide phrase. The morphological analysis was performed using the *pymorphy2* library [15].

In the experiments, news articles were generated according to a storyline containing from three to six guide phrases, which were single words or collocations. We chose the news domain since methods for generating news articles are widely used in computer journalism, e.g., to provide journalists with first draft news and templates [2]. In addition, the study of text generation methods in relation to news is an important direction of NLG because it allows evaluating the ability of language models to create disinformation [32]. Table 1 shows ten prompts with guide phrases used in the experiments.

The following news generation strategies were compared:

- uncontrolled language generation, i.e., using only the base ruGPT3Large model (ruGPT3 strategy);
- shifting the token scores according to equation (3) without increasing the probability of choosing the first token of the phrase W by the model p (ruGPT3 + ruRoBERTa strategy);
- shifting the token scores according to the equation (3) with increasing the probability of choosing the first token of the phrase W according to equation (4) (ruGPT3 + ruRoBERTa + $\lambda\alpha\Delta$ strategy).

Top- K sampling with the parameter $K = 10$ was used for decoding. The length of the generated sequence was 90 tokens (excluding the prompt).

The quality of the generated texts was evaluated using the measures of perplexity, repetition, and success rate [4, 23]. Perplexity (PPL) is calculated as exponential average of the negative logarithmic probability per token in the language model. A separate ruGPT3Medium³ model (350 million param-

¹ https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2.

² <https://huggingface.co/sberbank-ai/ruRoberta-large>.

³ https://huggingface.co/sberbank-ai/rugpt3medium_based_on_gpt2.

ters) was used to compute the perplexity. Repetition score (Rep) [35] calculates the proportion of repeated 4-grams in the text. Success rate (SR) calculates the proportion of guide phrases which occurred in the generated text.

5 Results and Discussion

The vocabularies of the ruGPT3 and ruRoBERTa models differ from each other: the size of both is 50,257 tokens, and the intersection of these vocabularies contains 35,215 tokens. Thus, the proportion of common tokens in model vocabularies is 70.1%. Both models use BBPE tokenization (Byte-level Byte-Pair-Encoding) as the tokenization strategy. The differences in the vocabularies are due to the differences in the training corpora used to pretrain the models. The training corpus for the ruGPT3 had a total size of over 600 Gb. It included a huge collection of Russian literature, Russian and English Wikipedia, public sections of Pikabu⁴, a complete collection of materials from the popular science portal 22century.ru and the banking portal banki.ru, as well as the Omnia Russica corpus⁵. The training corpus for ruRoBERTa had a size of 250 Gb. It was created from the ruGPT3 training corpus, from which the English texts were removed. Examples of ruGPT3 and ruRoBERTa tokenization are shown in Figure 4.

Text: В этом году конкурс в вузы на ИТ-направления сильно вырос.

ruGPT3: ['В', 'этом', 'году', 'конкурс', 'в', 'ву', 'зы', 'на', 'ИТ', '-', 'на', 'прав', 'ления', 'сильно', 'вырос', ':']

ruRoBERTa: ['В', 'этом', 'году', 'конкурс', 'в', 'в', 'узы', 'на', 'И', 'Т', '-', 'направ', 'ления', 'сильно', 'вырос', ':']

Figure 4: An example of ruGPT3 and ruRoBERTa tokenization

Table 2 presents the average values of perplexity, repetition, and success rate, calculated for news articles generated from 10 prompts and guide phrases presented in Table 1. We generated 10 samples of news per prompt. Thus, the average values of the measures for each strategy were calculated using 100 samples.

In Table 2, the values of SR show that the ruGPT3 + ruRoBERTa + $\lambda\alpha\Delta$ strategy ensures the occurrence of more than 96% of guide phrases in the text, while in the texts generated by the ruGPT3-such phrases are absent. When using the ruGPT3 + ruRoBERTa strategy, the guide phrases occurred in relatively few texts, as indicated by the value SR = 13.63%.

Analyzing the Rep values, we can conclude that the strongest Rep = 61.35% when using the ruGPT3 + ruRoBERTa strategy. The cyclic repetition of the same fragments was observed in all the texts. The ruGPT3 + ruRoBERTa + $\lambda\alpha\Delta$ strategy has the least repetition, which does not exceed 4%.

⁴ <https://pikabu.ru>.

⁵ <https://omnia-russica.github.io>.

#	Prompt	Guide phrases
1	В пятницу в драматическом театре состоялась премьера спектакля. On Friday, the play premiered at the Drama Theatre.	новое прочтение, игра актеров, красивые декорации, реакция зрителей, успех new reading, play of the actors, beautiful scenery, reaction of the audience, success
2	В пятницу в драматическом театре состоялась премьера спектакля. On Friday, the play premiered at the Drama Theatre.	произведение классической литературы, ошибка режиссера, освещение, провал спектакля work of classical literature, director’s mistake, lighting, performance failure
3	Все билеты на очередной этап гонок «Формула-1» были раскуплены задолго до соревнования. All tickets for the next racing event of the Formula 1 had been sold out long before the competition began.	опытные гонщики, мощные двигатели, визг тормозов, рев толпы, восторг зрителей experienced racers, powerful engines, screeching brakes, roaring crowd, delighted spectators
4	В этом году конкурс в вузы на ИТ-направления сильно вырос. This year, the acceptance rate for IT specialties in universities has grown significantly.	количество выпускников школ увеличилось, интересная специальность, перспективы хорошей работы, высокая зарплата the number of school graduates has increased, interesting speciality, good job prospects, high salary
5	Стоимость автомобилей в автосалонах выросла на 25% с начала года. The cost of cars in showrooms has increased by 25% since the beginning of the year.	новые автомобили, средняя комплектация, долгое ожидание, отсутствие комплектующих new cars, medium equipment, long wait, lack of components
6	После снятия ограничительных мер посещаемость кинотеатров увеличилась. After lifting the restrictive measures, cinema attendance increased.	сильная эпидемия, длительный карантин, рост на 10% severe epidemic, long quarantine, 10% growth
7	Родители и школьники дают разную оценку дистанционному образованию. Parents and schoolchildren assess distance education differently.	вспышка коронавируса, обучение в дистанционном формате, родители высказали недовольство, ученики стали позже вставать outbreak of coronavirus, distance learning, parents expressed dissatisfaction, schoolchildren get up later
8	Мэр рассказал о планах благоустройства городских парков. The mayor told about the plans for the improvement of city parks.	очередная пресс-конференция, планы, программа озеленения города, новые клумбы, красивый город another press conference, plans, program of planting greenery in the city, new flowerbeds, beautiful city
9	В издательстве выходит новая книга известной детской писательницы. A new book by a famous children’s writer is coming out.	интересный сюжет, история о волшебстве, магические превращения, яркие иллюстрации, первые положительные отзывы, большой тираж interesting plot, story about magic, magical transformations, bright illustrations, the first positive reviews, large circulation
10	Ученые опубликовали результаты масштабного исследования об изменении климата. Scientists have published the results of a large-scale study on climate change.	повышение температуры, разрушение озонового слоя, скорость таяния ледников, жизнь на Земле, угроза человечеству rise of temperature, ozone depletion, rate of glacier melt, life on Earth, threat to humanity

Table 1: List of prompts with guide phrases used in experiments

Generation strategy	λ_0	\downarrow PPL \pm Std	\downarrow Rep, %	\uparrow SR, %
ruGPT3	–	8.5 ± 2.7	14.53	0.00
ruGPT3 + ruRoBERTa	–	6.9 ± 3.5	61.35	13.63
ruGPT3 + ruRoBERTa + $\lambda\alpha\Delta$	0.1	20.6 ± 5.6	4.00	96.93
	0.3	18.9 ± 4.9	2.56	99.40
	0.5	18.3 ± 4.0	3.62	99.67

Table 2: Values of quality measures for generating news articles. Lower values of PPL and Rep correspond to a higher quality model (\downarrow), higher values of SR correspond to a better model (\uparrow)

The minimum PPL obtained for ruGPT3 + ruRoBERTa is equal to 6.9 (the lower the perplexity, the better the model). However, given the high Rep value for ruGPT3 + ruRoBERTa, we cannot define this strategy as of high quality because the texts contain a large number of repeating fragments. Therefore, without an additional increase of the probability of the first token from the guide phrase, it is impossible to generate good texts. For the ruGPT3 + ruRoBERTa + $\lambda\alpha\Delta$ strategy, the perplexity is greater than for ruGPT3. Such an increase in perplexity compared to the base ruGPT3 model is due to the fact that the shift in the values of the score function (2) in order to guide generation along the given guide phrases is “unnatural” for the model. This causes more “surprise” of the model to the tokens observed in the text.

Figure 5 for the example shown in Figure 3 represents graphs of score functions for ruGPT3 ($score_{ruGPT3}$) and ruRoBERTa ($score_{ruRoBERTa}$) for the top-500 common tokens – candidates for the <mask> token, sorted in descending order of scores in the ruGPT3.

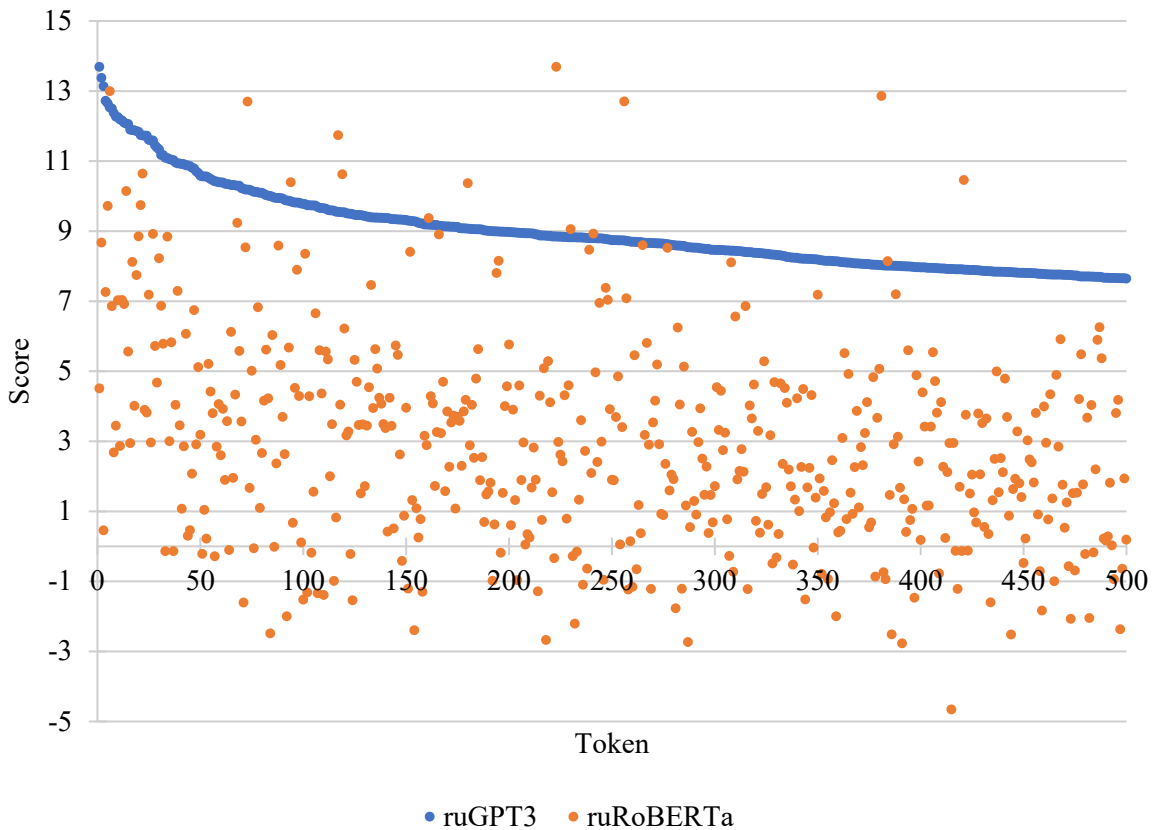


Figure 5: Graphs of the scores for ruGPT3 and ruRoBERTa for the example in Figure 3. The abscissa shows the serial numbers of tokens in the top-500

Figure 5 shows that with a monotonous decrease of the function $score_{ruGPT3}$ in the graph the values of the function $score_{ruRoBERTa}$ change randomly. The tokens that are more logical for ruGPT3 as those to continue the text based on the left context are not as logical for ruRoBERTa given left and right contexts. Therefore, function (3), which sums the values of the functions $score_{ruGPT3}$ and $score_{ruRoBERTa}$, is a compromise solution for determining the logic of the token at the current generation step.

Table 3 shows the top-10 tokens that are candidates instead of the <mask> token in Figure 3 with the largest values of the score functions for ruGPT3, ruRoBERTa and ruGPT3 + ruRoBERTa.

ruGPT3		ruRoBERTa		ruGPT3 + ruRoBERTa	
Token	Score	Token	Score	Token	Score
работают	13.692	прошла	13.692	проходит	25.526
работает	13.377	проходит	12.995	пройдет	22.885
благо	13.136	состоялась	12.857	прошла	22.537
уже	12.720	состоится	12.702	будет	22.371
идет	12.648	пройдет	12.701	идет	22.369
проходит	12.531	пройдёт	11.737	проводится	22.210
в	12.501	будет	10.643	работает	22.049
действуют	12.375	-	10.625	планируется	21.486
более	12.269	была	10.458	состоится	21.436
действует	12.254	началась	10.394	пройдёт	21.286

Table 3: Top-10 tokens with the highest values of score functions for the example in Figure 3

Analyzing Table 3, we can notice that the top-10 ruGPT3 tokens do not reflect the tendency of the narrative in the text to shift towards the guide phrase “очередная пресс-конференция” (“another press conference”). On the contrary, the top-10 ruRoBERTa tokens provide a logical transition to the guide phrase. After summing the functions $score_{ruGPT3}$ and $score_{ruRoBERTa}$ the top-3 tokens include the words “проходит” (“is being held”), “пройдет” (“will be held”) и “прошла” (“was held”), which, on average, for both models seem to be the most suitable for both continuing the left context and for a coherent transition to the right context. The result of the method at the i -th step for the example in Figure 3 is the next phrase: “Сегодня в городе проходит” (“Today another press conference is being held in the city”)⁶. In the next step, the <mask> token is moved one position to the right, and the guide phrase is inserted again until it appears in the text. Thus, the proposed method provides a coherent transition to the guide phrase in the text.

Table 4 shows the average computation time for $score_{ruGPT3}$ and $score_{ruRoBERTa}$, as well as the running time of the top- K sampling algorithm in the process of generating one text from the input data from Table 1. The time was calculated when the algorithm was executed on the server with GPU RTX A6000 48 Gb.

Table 5 shows the examples of texts generated using the developed method for different values of λ_0 . Other examples are given in the repository⁷.

Time of computation $score_{ruGPT3}$, sec	Time of computation $score_{ruRoBERTa}$, sec	Time of top- K sampling, sec	Total time, sec
6.777	5.587	0.843	13.207

Table 4: Calculation time of $score_{ruGPT3}$, $score_{ruRoBERTa}$ and the running time of the top- K sampling algorithm when generating one text from the input data from Table 1

⁶ In Russian, due to specific word order, the phrase “another press conference” is followed by “is being held”.

⁷ <https://github.com/iccreamz/Collocation2Text>.

λ_0	Guide phrases	Generated text
0.1	<p>очередная пресс-конференция, планы, программа озеленения города, новые клумбы, красивый город</p> <p>another press conference, plans, program of planting greenery in the city, new flowerbeds, beautiful city</p>	<p>Мэр рассказал о планах благоустройства городских парков. Очередная пресс-конференция с журналистами состоялась после того, как депутаты Городской думы приняли решение о выделении в этом году средств для приведения в порядок центральных городских парков и о создании новых зон отдыха и развлечений. - Все планы, которые были озвучены сегодня на заседании Городской думы, были поддержаны депутатами - и это только программа озеленения города. - На площади Ленина появятся новые клумбы и цветники - по инициативе депутатов. - Будет большой и красивый город –</p> <p>The mayor told about the plans for the improvement of city parks. Another press conference with journalists took place after the deputies of the City Duma decided to allocate funds this year to put the city’s central parks in order and create new recreation and entertainment areas. - All the plans that were announced today at the meeting of the City Duma were supported by the deputies - and this is only a program of planting greenery in the city. - New flowerbeds and gardens will appear in Lenin Square - at the deputies’ initiative. - There will be a big and beautiful city -</p>
0.3	<p>интересный сюжет, история о волшебстве, магические превращения, яркие иллюстрации, первые положительные отзывы, большой тираж</p> <p>interesting plot, story about magic, magical transformations, bright illustrations, the first positive reviews, large circulation</p>	<p>В издательстве выходит новая книга известной детской писательницы. Ее интересный сюжет, яркие персонажи и увлекательная интрига никого не оставят равнодушными! Это - настоящая детская история о волшебстве, доброте и волшебных превращениях в мире сказок и магические превращения, происходящие в нашем мире, где все мы являемся частью чего-то большого, прекрасного и удивительного! Очень интересные яркие иллюстрации! И первые положительные отзывы! Это очень не большой тираж! Но это то, что нужно маленьким детям, которым еще не довелось прочитать книгу.</p> <p>A new book by a famous children’s writer is coming out. Its interesting plot, bright characters and fascinating intrigue will not leave anyone indifferent! This is a real children’s story about magic, kindness and magical transformations in the world of fairy tales and magical transformations taking place in our world, where we are all part of something big, beautiful and amazing! Very interesting bright illustrations! And the first positive reviews! This is not a very large circulation! But this is what young children who have not yet had a chance to read a book yet need.</p>
0.5	<p>новое прочтение, игра актеров, красивые декорации, реакция зрителей, успех</p> <p>new reading, play of the actors, beautiful scenery, reaction of the audience, success</p>	<p>В пятницу в драматическом театре состоялась премьера спектакля. И снова это было новое прочтение пьесы. Игра актеров и режиссера в спектакле не уступает профессионализму постановщиков. Очень понравились красивые декорации, костюмы и реакция зрителей. Первый успех спектакля - это очень хороший старт для молодого театра. А впереди уже новые премьеры. И, конечно, очень интересно наблюдать за тем, как театр развивается. Смотреть фото с концерта можно здесь.</p> <p>On Friday, the play premiered at the Drama Theatre. And again it was a new reading of the play. The play of the actors and the director in the performance is not inferior to the professionalism of the directors. I really liked the beautiful scenery, costumes and the reaction of the audience. The first success of the performance is a very good start for the young theatre. And there are new premieres ahead. And, of course, it is very interesting to watch how the theater develops. You can see photos from the concert here.</p>

Table 5: Examples of texts generated by the proposed method. The prompts are highlighted in blue, the guide phrase – in red. The spelling of the generated texts is given using the original variant of the model

6 Conclusion

The proposed *Collocation2Text* method allows generating texts according to the user-specified list of guide phrases that represent the text storyline. Guide phrases can be single words and collocations. The method combines the advantages of autoregressive and autoencoding language models. The first model can well predict the next token in the sequence by the previous context, and the second model – the missing word in the text. We applied this method to the Russian language. The experiments on the news articles generation showed that the method ensures the occurrence of a high proportion of guide phrases in the text (more than 96% in SR), provides a low repetition of the text (less than 4% in Rep), which is even less than the repetition of the basic generative language model. However, this increases the text perplexity. The analysis of the tokens with the highest value of the score function showed that the method contributes to a coherent transition between the previously generated text and the guide phrases that are supposed to occur in the text.

Ethical Considerations

Information reliability in automatically generated texts is determined by the information reliability used to pre-train the models. It is known that large language models, including ruGPT-3, are capable of generating texts that cannot be distinguished from human-written texts. The method proposed by us can generate texts containing false and unreliable information. The use of guide phrases that contain factually incorrect content will result in the generation of incorrect or false information that can be used to harm, for example, to spread personalized misinformation. Our intention is to make social communications easier. We hope that this method will be used in socially positive applications. Providing open access to such methods will help to develop ways to detect them.

Acknowledgements

This work was supported by Russian Science Foundation, project № 22-21-00885, <https://rscf.ru/en/project/22-21-00885>.

References

- [1] Alhussain A. I., Azmi A. M. Automatic Story Generation: A Survey of Approaches // *ACM Computing Surveys*. — 2021. — Vol. 54, I. 5. — P. 1–38.
- [2] Anantrasirichai N., Bull D. Artificial intelligence in the creative industries: a review // *Artificial Intelligence Review*. — 2022. — Vol. 55. — P. 589–656.
- [3] 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.
- [4] 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. — Vol. arXiv: arXiv:1912.02164. — Access mode: <https://arxiv.org/abs/1912.02164>.
- [5] Fan A., Lewis M., Dauphin Y. Hierarchical neural story generation // *Computing Research Repository*. — 2018. — Vol. arXiv:1805.04833. — version 2. Access mode: <https://arxiv.org/abs/1805.04833>.
- [6] Fu Y., Feng Y. Natural Answer Generation with Heterogeneous Memory // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. — 2018. — P. 185–195.
- [7] Garbacea C., Carton S., Yan S., Mei Q. Judge the Judges: A Large-Scale Evaluation Study of Neural Language Models for Online Review Generation // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. — 2019. — P. 3968–3981.
- [8] Gatt A., Krahmer E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation // *Journal of Artificial Intelligence Research*. — 2018. — Vol. 61. — P. 65–170.
- [9] Haslum P. Narrative Planning: Compilations to Classical Planning // *Journal of Artificial Intelligence Research*. — 2012. — Vol. 44. — P. 383–395.
- [10] 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.
- [11] Jain P., Agrawal P., Mishra A., Sukhwani M., Laha A., Sankaranarayanan K. Story Generation from Sequence of Independent Short Descriptions // *Computing Research Repository*. — 2017. — Vol. arXiv:1707.05501. — version 2. Access mode: <https://arxiv.org/abs/1707.05501>.

- [12] Jhala A., Young R.M. Intelligent Machinima Generation for Visual Storytelling // *Artificial Intelligence for Computer Games*. — 2011. — P. 151–170.
- [13] Kartal B., Koenig J., Guy S.J. User-driven narrative variation in large story domains using monte carlo tree search // *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. — 2014. — P. 69–76.
- [14] Keskar N.S., McCann B., Varshney L., Xiong C., Socher R. CTRL – A Conditional Transformer Language Model for Controllable Generation // *Computing Research Repository*. — 2019. — Vol. arXiv: arXiv:1909.05858. — Access mode: <https://arxiv.org/abs/1909.05858>.
- [15] Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // *Analysis of Images, Social Networks and Texts*. — 2015. — P. 320–332.
- [16] 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.
- [17] 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 (Volume 1: Long Papers)*. — 2021. — P. 4582–4597.
- [18] Li Z., Ding X., Liu T. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training // *Proceedings of the 27th International Conference on Computational Linguistics*. — 2018. — P. 1033–1043.
- [19] Liu X., Ji K., Fu Y., Du Z., Yang Z., Tang J. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks // *Computing Research Repository*. — 2021. — Vol. arXiv: arXiv:2110.07602. — Access mode: <https://arxiv.org/abs/2110.07602>.
- [20] 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.
- [21] Mikolov T., Chen G., Corrado G., Sutskever I., Dean J. Efficient estimation of word representations in vector space // *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*. — 2013. — P. 1–12.
- [22] Oliveira H.G. A Survey on Intelligent Poetry Generation: Languages, Features, Techniques, Reutilisation and Evaluation // *Proceedings of the 10th International Conference on Natural Language Generation*. — 2017. — P. 11–20.
- [23] 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.
- [24] Peng N., Ghazvininejad M., May J., Knight K. Towards controllable story generation // *Proceedings of the First Workshop on Storytelling*. — 2018. — P. 43–49.
- [25] Pennington J., Socher R., Manning C.D. Glove: Global vectors for word representation // *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. — 2014. — P. 1532–1543.
- [26] Prabhunoye S., Black A.W., Salakhutdinov R. Exploring Controllable Text Generation Techniques // *Proceedings of the 28th International Conference on Computational Linguistics*. — 2020. — P. 1–14.
- [27] Qin G., Eisner J. Learning how to ask: Querying LMs with mixtures of soft prompts // *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. — 2021. — P. 5203–5212.
- [28] 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/>.
- [29] Reiter E., Dale R. *Building applied natural language generation systems* // Cambridge University Press, Cambridge, UK, 2000.
- [30] Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents // *Text Mining: Applications and Theory*. — 2010. — P. 3–20.
- [31] Santhanam S., Shaikh S. A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems – Past, Present and Future Directions // *Computing Research Repository*. — 2019. — Vol. arXiv: arXiv:1906.00500. — Access mode: <https://arxiv.org/abs/1906.00500>.
- [32] Shu K., Li Y., Ding K., Liu H. Fact-Enhanced Synthetic News Generation // *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*. — 2021. — P. 13825–13833.
- [33] 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.
- [34] Wang T., Wan X. T-CVAE: Transformer-based conditioned variational autoencoder for story completion // *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. — 2019. — P. 5233–5239.
- [35] 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.
- [36] 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.

Keywords in retrieval of prosodic data

Yanko T. E.

Institute of Linguistics,
Russian Academy of Sciences /
1 bld. 1 Bolshoy Kislovsky
Lane, 125009 Moscow,
tanya_yanko@list.ru

Abstract

Data retrieval can be based on the keywords found in the orthographic records of sound speech. This paper is aimed at analyzing the communicative meanings, compositions of meanings expressed by prosody, and the structure of segmental material aligned with prosodic variations. The meanings and their compositions expressed by prosody can have segmental counterparts, e.g., lexemes, or some other linguistic units. These counterparts, in turn, are used as keywords for retrieval. The results of searching are used as a material for analyzing the syntactic structure of the communicative components of sentences, such as themes and rhemes of statements, and the components of some specific illocutions (dreams, recollections, explanations), and combinations of illocutions with discourse incompleteness. The data source for this analysis is the Russian National Corpus (Multimodal sub-corpus Murko), Prosodically Annotated Corpus of Spoken Russian (SpokenCorpora.ru), and video-hosting Youtube. As the instrument, the computer system for sound speech analyzing Praat is used. The paper is illustrated throughout with pitch contours of sound records.

Keywords: sound speech; illocutions; data retrieval; keywords

DOI: 10.28995/2075-7182-2022-21-577-585

Метод поиска просодических данных по ключевым словам

Янко Т. Е.

Институт языкознания РАН
/ 125009, Москва, Большой
Кисловский пер. 1 стр. 1
tanya_yanko@list.ru

Аннотация

К поиску информации в массиве звучащих данных может быть применен метод ключевых слов. В работе анализируются коммуникативные значения, их композиции, которые выражаются просодией, и структура сегментного материала, несущего эти значения. Значения и их композиции, выражающиеся просодически, могут иметь свои сегментные корреляты, лексемы и другие единицы языка. Эти корреляты и используются в качестве ключевых слов поиска. Результаты поиска служат материалом для анализа синтаксической структуры коммуникативных компонентов предложений, таких, как темы и ремы сообщений, компоненты особых типов иллокуций (мечтаний, воспоминаний, обоснований), а также композиции иллокуций с дискурсивной незавершенностью. В качестве источника звучащего материала использован Мультимодальный подкорпус Национального Корпуса русского языка НКРЯ, корпус «Рассказы о сновидениях и другие корпуса звучащей речи» (SpokenCorpora.ru) и видео-хостинг Youtube. Инструментальный анализ данных проведен с помощью компьютерной системы анализа звучащей речи Praat. Работа иллюстрирована графиками изменений частоты звуковых данных.

Ключевые слова: звучащая речь; поиск информации; иллокуции; ключевые слова

1 Введение

В настоящее время исследователями просодии активно создаются просодически аннотированные звучащие корпуса. Для русского материала это корпус «Русский мультимедийный курс» [7] и «Рассказы о свидениях и другие корпуса звучащей речи» [9]. Аннотирование материала предусматривает ручную просодическую и иную разметку орфографической записи текста. Единицам текста приписываются частотные (в терминах частоты основного тона), паузальные, долготные и другие характеристики. Это создает возможность для поиска по движениям тона (подъемам, падениям и их комбинациям), длине пауз и другим параметрам.

Между тем при анализе просодии часто возникает необходимость поиска не по просодическим показателям, а по тем значениям, которые просодия выражает. Эти значения могут выражаться лексически, т.е. через сегментные единицы языка. Соответственно, мы предлагаем заменить поиск просодических маркеров значений на поиск по сегментным единицам языка: ключевым словам (или, дескрипторам). Ключевое слово — это традиционный термин поиска информации. Через ключевые слова и логические связи формулируется поисковый запрос, множество ключевых слов — это т.н. поисковый образ текста или массива текстов. В качестве поискового образа текста может быть использован и сам текст. Примеры ключевых слов в смысле подхода к поиску, предлагаемого здесь, мы приведем ниже, потому что коммуникативные значения представляют собой весьма абстрактные смыслы.

В соответствии с советом анонимного рецензента этой работы, мы должны отмежеваться от другого способа поиска информации в звучащем массиве речи, тоже называемого поиском по ключевым словам. При этом способом ключевыми словами объявляются собственно носители акцентных пиков на акцентной кривой [18] или слова, обозначающие эмоции [13], ибо слова, выражающие эмоции, по [13], могут иметь просодическое выражение. Таким способом через анализ просодической кривой автоматически индексируются транскрипты звучащих текстов. В дальнейшем это обеспечивает информационный поиск в массиве транскриптов. При нашем же подходе речь идет только о словах, непосредственно выражающих коммуникативные значения или косвенно на них указывающих. Коммуникативные значения формируют предложение как речевой акт и отвечают за встраивание речевого акта в связный нарратив или диалог, а применение метода ключевых слов происходит не в процессе индексирования текста, а на этапе формирования запроса.

Как могут быть использованы результаты поиска по ключевым словам, связанным с прагматическими значениями? Возьмем результаты поиска. Перед нами просодический материал с заданными заранее прагматическими параметрами речевых актов. Это фрагмент рабочего корпуса для проведения научного анализа: анализа синтаксиса речевых актов, семантики и способов наложения акцентной кривой на сегментный материал. Так, например, результаты поиска могут не только указать на словоформу — носитель определенного движения тона, но и представить достаточный материал для определения общих принципов выбора акцентоносителя коммуникативно значимой просодической структуры.

Таким образом, в этой работе предлагается рабочий метод оптимизации формулирования информационного запроса. Подбор ключевых слов предлагается производить методом интроспекции и с использованием традиционных для информационного поиска тезаурусных связей. При поиске используются ключевые слова, коррелирующие с коммуникативными значениями, а также их синонимы, родовидовые значения и ассоциаты. В дальнейшем результаты поиска подвергаются лингвистическому анализу, поэтому в работе, кроме обоснования методики поиска, говорится и о конкретных эпизодах использования его результатов.

Приведем пример. Пусть рассматривается просодический контур предложений, которые выражают воспоминания говорящего. Воспоминание — это особый иллокутивный тип предложения. Говорящий погружается в свой внутренний мир и воссоздает картины прошлого: они возникают перед его мысленным взором. Для выражения этого значения просодия предусматривает специальное средство. В русском языке это конструкция типа ИК-6 (по Е.А. Брызгуновой [1]), совершающаяся с существенным растяжением ударного слога словоформы-акцентоносителя.

При этом поиск планируется проводить не по ИК-6 (включая ее тональные варианты), а по ключевому слову *вспоминаю*, его синонимам и ассоциатам. Рассмотрим тональный рисунок ИК-6 с растяжкой:

1) *Мы заехали на поля-янку, на березовую, погуля-яли там* [9].

Говорящий вспоминает, какой ему снился сон. Акцентоноситель здесь и ниже выделен жирным шрифтом. Ключевых слов воспоминания в этом примере и в его контексте нет. Соответствующее значение выражается только просодией. Обратимся к другому примеру. Герой-вертолетчик, чудом оставшийся в живых после Чернобыльской аварии, вспоминает, как он тушил пожар в ядерном реакторе:

2) *И вот я завис над этим самым реа-актором, на высоте двести ме-етров, пятьсот рентген зашкаливает* [14].

Последний пример получен уже в результате поиска по ключевому слову «воспоминания», которое значилось в названии записи. В результате поиска в массиве данных, полученных по ключевым словам *воспоминания, мемуары, вспоминаю, помню*, возникают данные для синтаксического анализа сегментного материала с целью определения принципов выбора словоформы-акцентоносителя. В примерах 1) и 2) это словоформы *полянку, погуляли, реактором* и *метров*. Как мы увидим ниже, выбор этих словоформ в качестве носителя акцента ИК-6 (с растяжкой) не случаен, он согласуется с определенными принципами, которые были выделены ранее для предложений другого иллокутивного типа и с другим типом акцента [15-16].

При поиске по ключевым словам мы в целом придерживаемся традиционной структуры значений, отмечаемых в словарной статье информационно-поискового тезауруса [12]. Семантические отношения между заглавным ключевым словом и единицами, находящимися с ним в синонимических, родовидовых, ассоциативных (т.е. имеющих референцию к той же ситуации, что и заглавное ключевое слово, ср. *покупка* vs. *продажа*) отношениях и в отношении *часть* vs. *целое* нацелены, как известно, на увеличение полноты поиска информации. В качестве ключевых слов мы используем не только однословные именные группы и именные группы с зависимыми, как это делается традиционно при информационном поиске, но и языковые единицы других типов вплоть до короткого предложения. Соответственно, ассоциативную связь по сравнению с традиционным тезаурусным подходом мы понимаем расширительно, как имеющую ту же референцию, что и исходное ключевое слово.

В разделе 2 рассматривается вопрос *Что случилось?* как ключевое слово, его аналоги, просодии ответа на него или обоснования, почему такой вопрос задается, а также выбор словоформы-акцентоносителя в ответе. В разделе 3 мы возвращаемся к воспоминаниям и показываем, что воспоминания и мечты формируют акцентоноситель по принципам, аналогичным тем, которые дает ответ на вопрос *Что случилось?*. В разделе 4 рассматриваются ключевые слова, называющие жанры пресс-конференции и интервью. Это жанры, демонстрирующие композицию дискурсивной незавершенности с недостаточно исследованной ранее на этот счет иллокуцией *да-нет*-вопроса.

2 Вопрос *Что случилось?* как ключевое слово

Вопрос *Что случилось?* и его аналоги (*В чем дело?*; *Что происходит?*; *Что такое?*; *Ты что?*; *Ты чего?*; *Ты чё?*; *А что?*; *А чё?*; *А ты чё?*) имеют в качестве ответа или в качестве объяснения, почему задается этот вопрос, иллокуцию обоснования. Рассмотрим пример: *Дядя Вася Денисюк // Говорит: //– Послушай, друг, // Ты скажи мне, // что случилось: // На лице твоём испуг! // – Дядя Вася Денисюк, – // Отвечает старый друг, – // Ты скажи мне, // что случилось: // У тебя же дым вокруг!* (Л. Каминский). Дядя Вася задает вопрос *что случилось?*, потому что видит испуг на лице своего друга, а его друг задает вопрос *что случилось?*, потому что видит дым в квартире у дяди Васи. Ответы и объяснения, почему вопрос задается, имеют две существенные особенности: 1) они выражают иллокуцию обоснования; 2) они имеют нерасчлененную комму-

никативную структуру, потому что состоит из одной ремы. Методологическая ценность анализа нерасчлененных предложений состоит в том, что они дают материал для определения акцентоносителя в коммуникативной составляющей, имеющей в качестве синтаксического сегментного материала предложение в целом (S). Обратимся к просодии обоснования и рассмотрим пример 3).

3) — *Что случилось?* — *У Максим Ивановича **мама** умерла* [8].

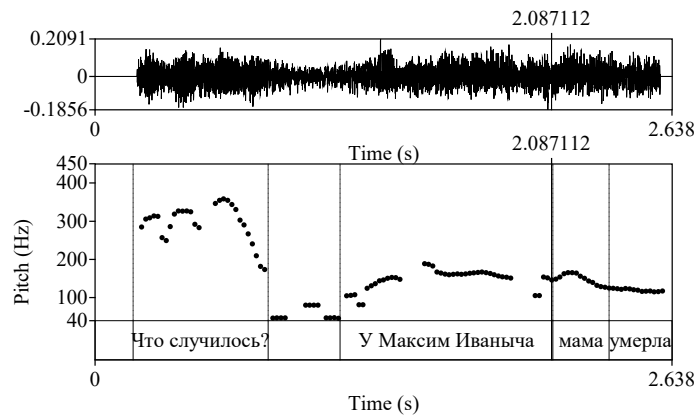


Рисунок 1: График изменения частоты (нижняя панель) в примере 1).

Онтологическая ситуация, в которой фигурирует пример 3), состоит в том, что первый говорящий слышит телефонный звонок и задает вопрос *Что случилось?* второму говорящему, ответившему на звонок. Второй говорящий сообщает о смерти матери общего коллеги. Рисунок 1 говорит о том, что акцентоноситель в ответной реплике примера 3) — словоформа *мама*. Она несет акцент ИК-2 (по Е. А. Брызгуновой [1]). Всего в ответной реплике четыре фонетических слова: *у Максим, Ивановича, мама и умерла*. Возникает вопрос, какому принципу подчиняется выбор акцентоносителя в предложениях обоснования, таких, как ответная реплика в примере 3).

Рассмотрим информационную (с точки зрения членения на данное vs. новое) и семантико-синтаксическую структуру предложения. Все элементы события, о котором говорится в ответе, — Максим Иванович, его мама и факт ее смерти — соотносятся с новой информацией. С семантико-синтаксической же точки зрения в предложении выделяется именная группа *у Максим Ивановича* как внешний посессор (EP), или второй участник ситуации (A2), именная группа *мама* — как субъект (A1) и финитный глагол — как предикат (P).

Для определения иерархии компонентов предложения как претендентов на роль акцентоносителя продолжим рассмотрение ответов на вопрос типа *Что случилось?*.

4) — *А че ты в тулупе-то парисься?* — ***Наручники мешают*** [8].

5) — *В чем дело?! Что?! Кто приказал?* — ***Дезинфекция** же в области. Деревесный жук все **деревья** поел* [8].

6) — *Да ты че?! — На Арбате **танки** стоят* [8].

7) ... *ты чего реवेशь?* — ***Пела** божественно* [8].

8) — *Что с вельветом?* — *Вельвет **берут*** [8].

Все результаты поиска здесь не приводятся, приведенные примеры позволяют проиллюстрировать следующие приоритеты, наблюдаемые при выборе акцентоносителя.

(1) Первый участник ситуации A1 (подлежащее) имеет приоритет перед финитным глаголом P, что иллюстрирует пример 4), ср. *Наручники (A1) мешают (P)*.

(2) Второй участник ситуации A2 имеет приоритет перед первым участником A1 и перед финитным глаголом P, о чем говорит пример 5), ср. *Древесный жук (A1) все деревья (A2) поел (P)*.

(3) Основные участники ситуации имеют приоритет перед внешним посессором (пример 3) *У Максим Ивановича мама умерла*) и локативным обстоятельством L (примеры 5) и 6), ср. *Дезинфекция же в области (L) и На Арбате (L) танки (A1) стоят (P)*.

(4) Финитный глагол P имеет приоритет перед обстоятельством образа действия (M), ср. пример 7): *Пела (P) божественно (M)*.

(5) Если синтаксически приоритетная группа соотносится с известным O (старым, данным, предупомянутым), она отдает свое право на роль акцентоносителя группе, следующей за ней в иерархии, т.е. такой, у которой при выборе акцентоносителя меньший вес. Так, в примере 8) акцентоноситель — глагол в личной форме при том, что прямое дополнение A2, в соответствии с п. (2) имеет больший вес, чем глагол. Дополнение A2 *вельвет* теряет свое право на роль акцентоносителя, т.к. в ответе оно заимствовано из вопроса: *Вельвет (O) берут (P)*. Здесь мы исходим из предположения, что в ответе на вопрос типа *что случилось?* могут быть местоименные, предупомянутые и другие компоненты, соответствующие известной информации и что при отсутствии просодических показателей темы это не влияет на нерасчлененность предложения. Иначе говоря, в составе ремы в определенных условиях могут быть известные компоненты.

Таким образом, можно наблюдать следующие приоритеты иерархии: $O < M < P < L < A1 < A2$. Наивысший приоритет в этой (пока неполной) иерархии у второго участника ситуации, ср. прямое дополнение в примере 5). Эти приоритеты совпадают с теми, которые были получены методами интроспекции и эксперимента [15-16], однако корпусный анализ говорит о неполноте результатов, достигнутых ранее, и необходимости проведения дальнейшего анализа. Некоторые из этих приоритетов были выделены и на английском материале [2], [3: 208], [6].

Добавим также, что, если приоритетная группа имеет внутреннюю структуру (так, имя может иметь атрибутивные зависимые, оно может быть встроенным в сочиненную или числовую группу, оно может представлять собой имя и отчество, имя и фамилию, имя и прозвище, титул и имя), акцентоноситель в такой группе выбирается в соответствии с локальными принципами, по сумме представленных в работах [4: 146]; [10: 203–206]; [11]; [5: 202]. Так, в примере *На базаре шесть гусей продал* [8]) в приоритетном актанте *шесть гусей*, который имеет структуру числовой группы, по локальному правилу выбирается акцентоноситель *гусей*, а в примере *...я его на собственные деньги купил* [8] в приоритетной именной группе *собственные деньги* — словоформа *деньги*. Локальные правила применяются к именным группам рекурсивно вплоть до достижения терминального узла.

Цель этого раздела состояла в том, чтобы продемонстрировать возможность разработки звучащего массива исследования путем использования ключевого слова *что случилось* и его аналогов.

3 Воспоминания и мечты

Для анализа предложений, отражающих погружение говорящего в его внутренний мир, были использованы ключевые слова *мечтаю, вспоминаю, помню, воспоминания*. Вернемся к примерам 1) и 2), рассмотренным во Введении. В примере 1) имеются два отдельных акцентоносителя *полянку (A2)* и *погуляли (P)*, а в примере 2) акцентоносители — это словоформы *реактором (A2)* и *метров*. Эти акцентоносители соответствуют иерархии приоритетов, полученной в разделе 2. Акцентоноситель *метров* соответствует локальному правилу выбора акцентоносителя в

числовой группе *двести метров*. Рассмотрим еще несколько примеров. Пример 9) служит результатом поиска в Мультимедийном подкорпусе НКРЯ по ключевому слову запроса *помню*.

9) *Я даже помню: доро-о-жка (A1) была-а (P) [8].*

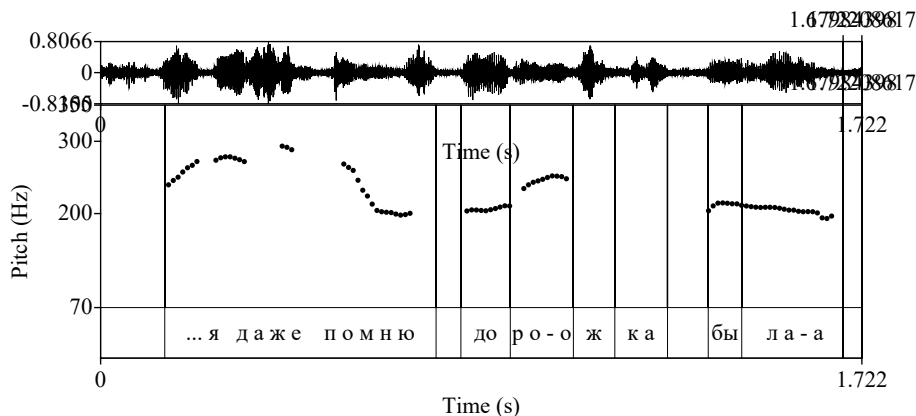


Рисунок 2: График изменения частоты в примере 9)

Тонотрама (нижняя панель) говорит о том, что на акцентоносителе словоформе *дорожка* фиксируется пологий подъем типа ИК-6, который сопровождается растянутой артикуляцией ударного слога. Заударный слог и последующая словоформа характеризуются плавным понижением тона. Перед нами просодическая структура, маркирующая воспоминания. Ср. также примеры 10)-11), отражающие мечты говорящего.

10) *Знаешь, я мечтаю, чтобы ты приехал на неде-ельку (A2), мы б с тобой погуля-яли (P), сходили бы в кино-о (A2) [8].*

11) *Она (O) уже седа-ая (P), наверное, теперь... За-амужем (P) давно (M). А все-таки мечтаю, хоть голос ее услышать [8].*

Примеры 9)-11) также подтверждают гипотезу о том, что принципы выбора акцентоносителя, полученные на примере ответов на вопрос типа *что случилось?*, применимы и к предложениям погружения говорящего в его внутренний мир.

Итак, дан краткий анализ просодии и принципа выбора акцентоносителя в предложениях с иллюкативной силой мечтаний и воспоминаний. Показано, что подбор звучащего материала для анализа можно осуществить через ключевые слова *помню, вспоминаю, мечтаю*. Отметим также, что просодический контур, условно названный нами просодией воспоминания и мечты, имеет более широкое толкование: соответствующей просодией оформляются любые речевые акты, связанные с имитацией мыслительной деятельности, которая понимается говорящим как протекающая во времени. Говорящий вспоминает, мечтает, размышляет, подбирает ускользящее слово, ищет подходящие примеры для того, чтобы слушающий лучше его понял, говорящий «шевелит мозгами».

4 Дискурсивная незавершенность и да-нет-вопрос

Дискурсивная незавершенность — одно из коммуникативных значений, имеющее определенные просодические показатели. Незавершенность, говорит о том, что текущий речевой акт позиционируется говорящим как неконечный в ряду речевых актов. Говорящий дает понять слушающему, что он еще не готов передать слово и собирается продолжить речь. Пример 12) демонстрирует незавершенность в речевом акте сообщения. Просодическими показателями незавершенности служит подъем частоты на акцентоносителях незавершенности *неофициальное, маленькая, называется и обозначили*.

12) Орлинка — это название **неофициальное**, потому что бухта **маленькая**, официально она никак не **называется**, ее так **обозначили**, чтоб хоть как-то [8].

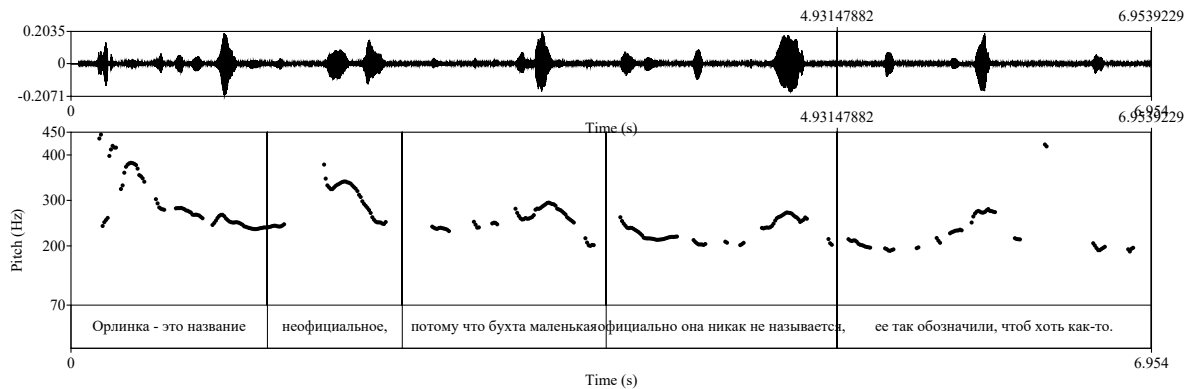


Рисунок 3: График изменения частоты в примере 12).

В работе [17] показано, что незавершенность совместима со всеми основными типами речевых актов: императивом, вокативом и вопросом с вопросительным словом. Все эти типы речевых актов способны присоединять показатель незавершенности. Единственным типом речевого акта, который оставался не исследованным окончательно, был *да-нет*-вопрос. Очевидно, что сочетаемость *да-нет*-вопроса с незавершенностью затруднена. Причиной этого представляется то, что показатель *да-нет*-вопроса, акцент типа ИК-3, это также и наиболее частый показатель незавершенности. *Да-нет*-вопрос, в котором иллюкуция вопроса передается частицей *ли*, в качестве *да-нет*-вопроса из рассмотрения исключается, потому что в вопросе с *ли* иллюкутивная сила выражается не просодически, а сегментно: частицей *ли*. Соответственно, вопросы с *ли* легко комбинируются с незавершенностью.

В настоящее время исследование корпусного материала говорит о том, что сочетаемость *да-нет*-вопроса с незавершенностью возможна. Анализ дискурсивной ситуации, в которой представлено образование вопроса в режиме незавершенного дискурса, свидетельствует о том, что за вопросом, за которым в наиболее естественном случае должен следовать ответ, иногда может следовать и другой вопрос или обоснование того, почему этот вопрос задается. Будет ли при переходе от одного вопроса к вопросу, следующему за ним, или повествовательному предложению выражаться значение незавершенности? Это не бессмысленный вопрос, потому что незавершенность, вообще говоря, может оставаться и не выраженной просодически.

В каких же ситуациях может быть задано более одного вопроса подряд? Мы предполагаем, что наиболее благоприятным дискурсивным контекстом для серии вопросов служит контекст пресс-конференции. Когда журналист получает слово, он, как правило, делает попытку задать максимальное количество вопросов, приличествующее ситуации, потому что второй раз слова ему уже не дадут. Заметим также, что вопрос, имеющий продолжение, которое нуждается в указании на незавершенность, должен содержать достаточно объемный сегментный материал, чтобы акцентам, выражающим отдельно вопрос и незавершенность, было где разместиться.

Кроме пресс-конференции более двух вопросов подряд может встретиться и в жанре интервью, когда интервьюеру естественно подготовить ряд вопросов заранее. Кроме того, может быть рассмотрен материал раздела «вопросы» на научных конференциях и лекциях. Все эти жанры послужили источником ключевых слов для поиска *да-нет*-вопросов, имеющих указание на то, что текущий вопрос — не последнее предложение в речи. Наиболее плодотворный результат дал анализ материалов пресс-конференций:

13) **Можете** рассказать вообще, есть ли такие, кто хочет приобрести пятьдесят процентов в капитале **Открытие**? Или, может быть, **всю** группу целиком? И какие шансы у IPO **Открытие**? [youtube]

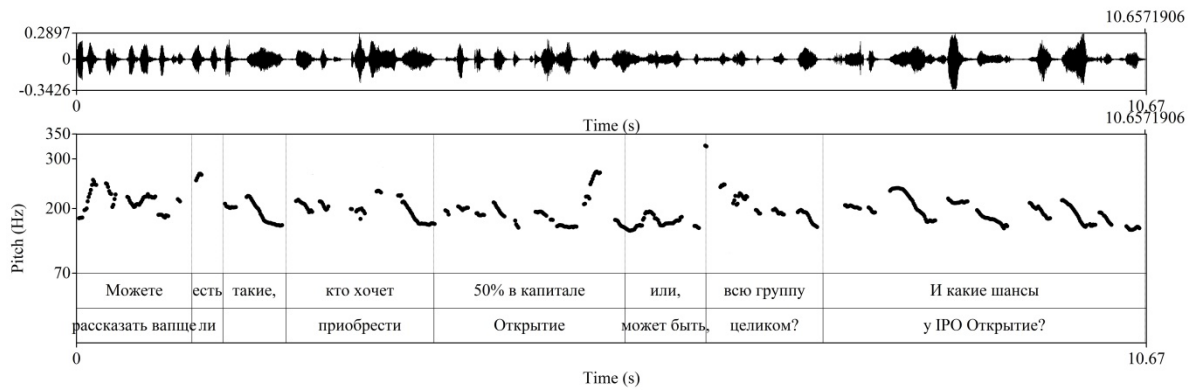


Рисунок 4: График изменения частоты в примере 13).

В примере 13) журналистка задает три вопроса подряд. Первый вопрос содержит подъем, маркирующий иллокуцию *да-нет*-вопроса на акцентоносителе вопроса словоформе *можете*, и подъем, маркирующий незавершенность на акцентоносителе незавершенности словоформе *Открытие* (первая встречаемость этой словоформы). Второй вопрос содержит только акцентоноситель вопроса, несущий подъем. Это словоформа *всю*. Последний вопрос с вопросительным словом *какие* завершает серию вопросов. Второй и третий вопрос указаний на незавершенность не содержат. Пример 13) весьма многословный, что, как известно, затрудняет анализ, но именно те контексты, которые содержат обширный сегментный материал, и составляют материал для анализа совместимости *да-нет*-вопросов и незавершенности.

Аналогично в примере 14) имеется подъем вопроса на словоформе *правильно* и подъем незавершенности на словоформе *получалось*:

14) *Правильно* я понимаю, что дружественными путями не *получилось*? И тогда решили действовать *агрессивно*? [youtube].

Всего в нашей коллекции около двух десятков примеров, аналогичных 13) и 14). Большинство примеров получено путем поиска по ключевому слову *пресс-конференция*. Несколько примеров найдено в жанре интервью с использованием соответствующего ключевого слова. Примеров достаточно для того, чтобы показать, что *да-нет*-вопрос совместим с просодией незавершенности.

5 Заключение

Показано, что поиск в массиве звучащих данных может быть проведен с использованием метода ключевых слов. Материалом для поиска служат транскрипты звучащих текстов, заголовки в орфографической записи и рубрики массива, в котором хранится текст. В качестве иллюстрации метода были использованы ключевые слова, обозначающие иллокутивную силу искомого речевого акта, и слова, связанные с обозначением искомой иллокутивной силы синонимическими, родовидовыми отношениями и отношениями, которые могут указывать на соответствующую искомой иллокутивной силе коммуникативную ситуацию. Последний тип отношений мы трактуем как ассоциативную связь в терминологии информационно-поискового тезауруса.

Кроме демонстрации метода поиска звучащего материала по ключевым словам в работе дается анализ проблем, для решения которых результаты поиска послужили рабочим материалом. Так, для анализа иллокутивной силы обоснования было использовано ключевое слово *что случилось*, его синонимы и ассоциаты. Результаты поиска позволили проанализировать принципы выбора словоформы-акцентоносителя в предложениях нерасчлененной коммуникативной структуры, которые как раз и представлены иллокуцией обоснования. Кроме того, анализ ответов на вопрос типа *что случилось?* позволяет выявить синтаксические приоритеты, которые действуют при выборе акцентоносителя в русском языке. Это типологически значимая языковая шкала, которая применима не только к русскому языку.

Далее. Для анализа выбора акцентоносителя в иллокуциях, соответствующих погружению говорящего в его внутренний мир, использованы ключевые слова *помню, вспоминаю, воспоминания, мечтаю*. Результаты поиска позволяют проанализировать просодическую структуру предложений, отражающих мыслительную деятельность, которая понимается как продолжающаяся во времени, и способы наложения просодической кривой на сегментный материал, в частности, выбор словоформы-акцентоносителя.

И наконец, поиск по ключевым словам *пресс-конференция* и *интервью* дает материал для анализа *да-нет*-вопроса в контексте дискурсивной незавершенности, потому что жанры, обозначенные этими ключевыми словами, способны представить вопрос в контексте других вопросов, следующих за ним, или объяснений того, почему этот вопрос задается, т.е. в контексте незавершенного дискурса.

Теоретическую значимость предложенного метода мы видим в разработке толкований прагматических значений, имеющих просодическое выражение, и в сопоставлении полученным толкованиям их лексических коррелятов, способных служить ключевыми словами запроса. Создание толкований фактически представляет собой разработку понятийного аппарата и соответствующего метаязыка для описания коммуникативно-просодического уровня языка.

References

- [1] Bryzgunova E. A. Intonation [Intonatsija] // Russian Grammar [Russkaja grammatika]. Vol. 1, Nauka, Moscow, 1982. P. 103-118.
- [2] Enkvist N. E. Marked focus: functions and constraints // Greenbaum, S. et al. (Eds.) Studies in English Linguistics for Randolph Quirk. London: 1979. P. 134–152.
- [3] Halliday M. A. K. Notes on transitivity and theme in English. Part 2 // Journal of Linguistics. 1967. Vol. 3. P. 199–244.
- [4] 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. P. 103-118.
- [5] Kodzasov S. V. Phrase accentuation laws [Zakony frazovoi aktsentuatsii] // Prosodicheskie stroi russkoi rechi. 1996.
- [6] Schwarzschild R. GIVENness, AvoidF, and other constraints on the placement of accent // Natural Language Semantics. 1999. 7. 141–177.
- [7] Russian multichannel discourse [Russkij mul'tikanal'nyj diskurs] (Multidiscourse 2020): <<http://multidiscourse.ru>>
- [8] Russian National Corpus [Nacional'nyj korpus russkogo jazyka]: <www.ruscorpora.ru>.
- [9] Prosodically Annotated Corpus of Spoken Russian “Night dream stories” [Rasskazy o snovidenijah i drugie korpusa zvuchashhej rechi]: <<http://spokencorpora.ru>>.
- [10] Russian Grammar [Russkaja grammatika]. Vol. 2, Nauka, Moscow, 1982.
- [11] Svetozarova N. D. Accent-rhythmical innovations in Russian spontaneous speech [Aktsentno-ritmicheskie innovatsii v russkoi spontannoj rechi]. Problemy fonetiki, I. 1993.
- [12] STATE STANDARD GOST 7.25-2001 SIBID. Thesaurus for information retrieval. The principles of development, structure, and presentation. The Forth of November 2001. [GOST 7.25-2001 SIBID. Tezaurus informacionno-poiskovyj odnojazychnyj. Pravila razrabotki, struktura, sostav i forma predstavlenija ot 04 sentjabrja 2001 - docs.cntd.ru]: <<http://docs.cntd.ru>>.
- [13] Toivanen J., Seppänen T. Prosody-based search features in information retrieval. Proceedings of FONETIK, Stockholm, Finland. 2002. Pp. 105-108.
- [14] Video-hosting youtube: <<https://www.youtube.com>>.
- [15] Yanko T. E. The communicative structure with a non-inherent theme [Kommunikativnaia struktura s neingerentnoi temoi]. Nauchno-tehnicheskaja informatsija, 1991. 2 (7).
- [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. Speech acts within the structure of discourse: the incompleteness cues in sound corpora [Rechevye akty v strukture svjaznogo diskursa: pokazateli nezavershennosti po dannym korpusov zvuchashhej rechi] // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2018). [Komp'juternaja lingvistika i intellektual'nye tehnologii]. Issue 17 (24). [Vyp. 17 (24)]. Moscow: RGGU Publ. 2018. P. 791-802.
- [18] Lezhenin Iu., Zhuikov A., Bogach N., Boitsova E., Pyskin E. PitchKeywordExtractor: Prosody-based Automatic Keyword Extraction for Speech Content // Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 265–269.

Historical Text Corpora and the Conclusiveness of Linguistic Analysis

Anton Zimmerling

Pushkin State Russian Language
Institute, Moscow, Russia;
Institute of Linguistics, Russian
Academy of Science, Moscow, Russia
fagraey64@hotmail.com

Abstract

I discuss the methodology and conclusiveness of the corpus-based historical linguistics and analyze two formal models predicting the language-internal variation in Early Old Russian syntax. Linguistic models claiming a rigid distribution of grammatical features like \pm overt realization of agreement markers activate hidden corpus characteristics such as profiles of text genres, chronology, vector of change, \pm impact of L2, \pm presence of supra-dialect features. In this case they can be valued and checked on text samples, where genre features are stable, while location and time vary.

Keywords: historical text corpora; conclusiveness; predictive linguistic models; Old Russian language

DOI: 10.28995/2075-7182-2022-21-586-593

Исторические корпуса текстов и лингвистическая доказательность

Циммерлинг А. В.

Государственный институт русского языка
имени А. С. Пушкина, Москва, Россия
Институт языкознания РАН, Москва, Россия
fagraey64@hotmail.com

Аннотация

В статье обсуждается доказательность лингвистических исследований, опирающихся на исторически корпуса текстов? и проверяются две формальные модели, предсказывающие распределение показателей согласования в древнерусском языке. Формальные лингвистические модели проверяемы в той степени, в которой они имплицитно опираются на скрытые характеристики корпусов, в том числе на профили жанров текста. Модельное представление о наличии в древнерусском языке нулевой связки 3 л. подтвердилось, однако тезис о том, что употребление ненулевых связок перфекта 3л. в XII в. всегда отражает влияние второго языка автора, опровергается на выборке текстов, представляющих разные древнерусские диалекты.

Ключевые слова: исторические корпуса текстов; доказательность; предиктивные лингвистические модели; древнерусский язык

1 Variation in Historical Text Corpora

Variation of grammatical parameters is a normal state for all natural languages, including old languages as they are reflected in historical text corpora. Linguistic models aim at predicting the distribution of language features in corpora and text collections. Parametric models generally have a greater predictive force, since they explain a visible diversity of features by different settings of the same constraint. E.g., predicative instrumental case (Instr_{PRED}) in Modern Russian is a case form assigned to predicative nominals or adjectives by the clausal subject. Standard accounts of Russian grammar, cf. [Bailyn 2012] fail to explain sentences like ⁷мне_{DAT} стало_{PAST.3SG.N} известным_{INSTR.PRED}, [CP что Вы в городе], ‘I got to know that you are in the city,’ lit. ‘to-me became known, that...’ accepted by a minor part of Russian speakers. Such sentences are puzzling, since they lack any nominative subject. A parametric approach explains this variation and helps to find the invariant. Both groups of Russian speakers have the subject control parameter. The feature that varies across speakers is not the correlation between Instr_{PRED} and subjects, but the status of sentential arguments like [CP что Вы в городе]: the minority of speakers licenses them as assigners of Instr_{PRED}, while the majority does not [Циммерлинг 2018]. An extreme case of language-internal variation can be described in terms of dialectal divergences, if a split in parameter values between different groups of speakers is consistent [Henry 1998].

Historical corpora pose an extra problem: it is not always clear how to classify the author’s idiom in a situation of language-internal variation. Assume that there are two expressions α and β that can be used in the same context, but only part of the speakers licenses them both. Let us also assume that the corpus has a system of genre tags and all texts representing genre G1, $T \in \{T_{G1}\}$ only display the option α , while all texts representing genre G2, $T \in \{T_{G2}\}$ display both α and β . This is exactly what one finds in Old Russian and historical corpora like RNC Old Russian corpus https://ruscorpora.ru/new/search-old_rus.html. The options α and β are values of the parameter that licensed either a zero form of the 3rd person BE-auxiliary in the Old Russian perfect construction ($\alpha = \emptyset^{3P}$), or a non-zero form ($\beta = \text{есть}_{3SG}$, суть_{3PL} , есть_{3DU}) in this construction. G1 can be identified with Old Russian birch bark letters, graffiti and other non-bookish texts. At present moment, only birch bark letters are included in RNC and placed in a separate section of RNC Old Russian corpus <https://ruscorpora.ru/new/search-birchbark.html>, so that G1 can be associated in RNC with a single text genre. G2 are Old Russian bookish texts of the XI—XIII centuries. The texts from the RNC Old Russian corpus largely overlap with G2, with one exception: “Russkaja Pravda” has a characteristic G1 grammar and is a non-bookish text that should be classified with birch bark letters, if the demarcation of sub-corpora were solely based on the G1 vs G2 distinction. The prediction is that texts from G1 only have the option $\text{пришел}_{SG.M} \emptyset^{3P}$ ‘<he> has come’ with the zero copula, while the texts from G2 use both the zero copula and the overt 3rd person copulas $\text{пришел}_{SG.M} \emptyset^{3P} \sim \text{пришел}_{SG.M} \text{есть}_{3SG}$. It is tempting to assume that the actual Old Russian usage was associated with G1, while overt 3rd person copulas in the perfect construction one finds in G2 were a tribute to the written tradition [Хабургаев 1978; Зализняк 2008: 236]. The idea that the speakers modified their grammar and added an extra option $\langle \alpha \rangle \Rightarrow \langle \alpha, \beta \rangle$, when they switched from colloquial speech (G1) to bookish texts (G2) is natural and almost trivial. A question however arises, whether this intuition can be checked empirically on corpora, since the absence of β is a definitional property of G1 and T_{G1} and the presence of β is a definitional property of G2 and T_{G2} .

The option β is not the only characteristics of G2. Semi-formally, G2 contains hybrid Old Russian texts, heavily influenced by a different, though closely related language, Old Church Slavonic. The notion of hybrid text can be formalized. The basic hypothesis is that the first language, L1 (= Old Russian, as reflected in G1) is suppressed in G2 by the second language, L2 (= Old Church Slavonic), while the texts from G1 are monolingual. Again, there is a methodological question, how one can prove that all speakers of Old Russian invariably switched to L2, when they used the option β and wrote down the 3rd person perfect form with overt copulas as $\text{пришел} \text{есть}$, $\text{пришли} \text{суть}$. A related methodological issue is whether the negative evidence from G1, where the option β is not displayed is conclusive and one can establish for sure that L1 Old Russian lacked any overt copulas in the 3rd person perfect form. For the sake of simplicity one can assume that the sets $\{T_{G1}\}$ and $\{T_{G2}\}$ are closed, since the possibilities of adding new texts to historical corpora are limited. The actual situation is different: recently discovered

birch bark letters are added to T_{G1} , while T_{G2} can be expanded by other bookish texts or alternative versions of the texts already included in Old Russian corpora. Regarding the illustration provided, the discovery of overt 3rd person perfect auxiliaries and sentences like *приишель есть* in a recently added birch bark letter will falsify the initial hypothesis that G1 completely lacks the option β , while the discovery of a bookish text, whose author generalized the use of zero copula in sentences like \emptyset^{3P} *приишель*_{SG.M} will challenge the initial hypothesis that the option β is visible in all texts from G2. The chosen model can rule out these exceptions as insignificant by claiming that the author of that birch letter reverted to G2/L2 grammar, while the author of that bookish text reverted to G1/L1 grammar, but such explanations are apparently circular and not conclusive.

I argue that historical corpus linguistics still has some tools to check similar hypotheses concerning the variation in a dead language, but one has to narrow the search and first establish a group of texts relevant for checking each grammatical feature.

2 Old Russian Grammar and Two Linguistic Models

Old Russian perfect is an analytical construction consisting of an active *l*-participle discriminating number and gender, and an auxiliary/copula (inflected present tense indicative BE-form) expressing the person-and-number agreement with a subject controller, if any. The grammar of Old Russian perfect is described by two linguistic models proposed by Andrej A. Zaliznjak (1993, 2008) and developed by other scholars.

2.1 Zaliznjak's Model A

The first hypothesis of Zaliznjak, henceforth — ‘Zaliznjak-A’ claims that the agreement markers used in the perfect construction were clitics. The 1st-2nd person perfect auxiliaries were clausal-second enclitics, i.e. enclitics taking the second position in clause, while Old Russian lacked overt 3rd person perfect clitics and replaced them with a zero 3rd person auxiliary [Zaliznjak 1993: 285].

- (i) <Zaliznjak-A> Old Russian agreement markers in the perfect construction were clitics with a fixed distribution. Such markers developed only in the 1st-2nd persons, but not in the 3rd person, where a zero auxiliary (copula) with feature ‘3rd person’ was used instead. Therefore overt agreement markers in the 3rd person were redundant.

Although Zaliznjak himself did not postulate a zero 3rd person BE-copula for Old Russian, this step made in [Zimmerling 2020; Циммерлинг 2021: 26-27] seems inevitable, since the conclusion that colloquial Old Russian completely lacked overt 3rd person perfect auxiliaries is based on the assumption that the absence of an overt agreement marker signals the agreement feature ‘3rd person’¹. The tag ‘colloquial Old Russian’ used by Zaliznjak or its synonym ‘Old Russian 1, OR-1’ introduced in [Zimmerling 2020] and the tag ‘G1’ used in this paper, refer to the same thing — the Old Russian idiom without overt 3rd person perfect auxiliaries. The difference is that OR-1 is an aprioristic model notion that is abstracted from with any collections of real texts, while G1 is a set of texts representing some specific genres, where (i) holds true.

2.2 Zaliznjak's Model B

The second hypothesis of Zaliznjak, henceforth — ‘Zaliznjak-B’ predicts the syntax of overt 3rd person BE-copulas, including 3rd person perfect auxiliaries in G2, i.e. hybrid bookish Old Russian texts. The

¹ A descriptive alternative — to postulate a defective morphological present tense paradigm for the Old Russian copula BE instead of recognizing the zero copula $\emptyset^{BE.PRES.3P}$ is not satisfactory, since Old Russian retained overt forms of the full verb BE.

prediction is that 3rd person BE-copulas form in G2 a complementary distribution with clausal subjects [Зализняк 2008: 240, 258-259].

- (ii) <Zaliznjak-B> Old Russian agreement markers including 3rd person BE-auxiliaries tend to form a complementary distribution with the agreement controllers (grammatical subjects).

I henceforth call the idiom associated with G2 ‘Old Russian-2’, or ‘OR-2’.

2.3 Past Tenses in OR-1 and OR-2

There is a sharp split between OR-1 and OR-2 regarding the use of past tenses. Hybrid texts from G2 have the OR-2 grammar and use simple past forms (aorist and imperfect) along with perfect. The texts from G1 that have the OR-1 grammar either generalize the use of perfect as the only extant past tense form or use aorist and imperfect in insignificant quantities, cf. [Zaliznjak 2004: 142]. Thus, OR-2 is an idiom which 1) has aorist and imperfect, 2) licenses overt 3rd person perfect auxiliaries.

2.4 Old Russian Dialects

There are four major groups of Old Russian dialects: a) Southern; b) North-Western (Novgorod and Pskov), c) Western (Smolensk and Polotsk), d) Great Eastern (e.g., Suzdal). Dialectal markers can be found both in texts from G1 and in texts from G2, though in the latter group they are often undesirable. That means that the presence/absence of overt 3rd person auxiliaries can also depend on the dialect, although this scenario has not been proven so far.

3 Analysis

I check models A and B on a sample of five non-anonymous G2 Old Russian texts of the XII century representing three dialects and measure three features: 1) the ratio of perfect and simple past tenses; 2) the presence of overt 3rd person perfect auxiliaries; 3) the correlation between overt 3rd person perfect auxiliaries and overt clausal subjects². The control group of texts includes four G2 XIV-XV century texts of comparable genre.

3.1 The Main Group

The main group includes five non-anonymous author texts representing three dialects and seven idioms. “Daniil’s Pilgrimage” and “Vladimir Monomakh’s Edification” are South Russian texts dated by the first quarter of the XII century. Three texts come from Novgorod: “Kirik’s Inquiries” are dated by ca. 1132-1156 A.D., “The Edification of Elija-John” is dated by 1166 A.D., while “Antonius’ Pilgrimage” is dated with ca. 1200. Two texts represent the genre of *khozhenie*, i.e. pilgrimage, two texts — the genre of *pouchenie*, i.e. edification. “Kirik’s Inquiries” are written in form a dialogue between church hierarchs and ordinary clerics who ask them concerning various issues in canon law. Its main part is written by Kirik, while two supplements are written by Savva and Elija.

Four of these texts are currently added to RNC Old Russian corpus as separate entries, the data from “Monomakh’s Edification” was in March 2022 only available as part of the larger compilation

² An anonymous reviewer argues that from the corpus linguistics perspective, modeling based on one parameter may lead to overgeneralization. In this paper we however used three parameters — 1) \pm overt 3rd person perfect auxiliaries, 2) \pm complementary distribution of auxiliaries and agreement controllers, 3) ratio of the perfect clauses from the total amount of three past tense forms — for measuring the variation in XII century Old Russian grammar. The analysis confirmed that these three parameters are independent.

(“Laurentius’ Chronicle”)³. For the remaining four texts, RNC tells the number of words and sentences, but not the number of clauses, which is the relevant parameter for the processing of perfect clauses. The texts from RNC Old Russian corpus are not annotated syntactically. “Monomakh’s Edification” consists of several parts. I ignored the last one — “The Lenten Prayer”, since it does not contain any perfect constructions. Other four texts from main group were processed and annotated in their entirety. I also checked an extra version of “Kirik’s Inquiries”, which differs from the version included in RNC.

3.2 The Control Group

The control group includes two Novgorod texts from the mid XIV century — “Stefan’s Pilgrimage” and “The Discourse about Paradise” by Vasiliy Kalika — and two Suzdal texts from the mid XV century — “The Pilgrimage to the Florence Council” and “A Note about Rome”. Three of them represent the genre of *khozhenie*, while Vasiliy Kalika’s text is comparable to Elijah-John’s “Edification”.

3.3 Syntactic Annotation

The text fragments containing the perfect construction in the 3rd person were extracted and parsed in order to check model B. 3rd person perfect clauses were assigned the value ‘+SUBJ’ if they contained an overt subject expressed by the accented noun/pronoun or NP, otherwise they were assigned the value ‘-SUBJ’, even if the subject position is reconstructed in a larger syntactic context.⁴

3.4 Disambiguation of Contexts and Semantic Annotation

In order to check model A, all contexts containing overt 3rd person perfect auxiliaries were annotated semantically. Standard perfect auxiliaries, including 3rd person auxiliaries *ecmb*¹, *cymb*¹, *ecma*¹ are forms lacking the phrasal accent: they are placed clause-internally and do not bring about existential or verificational meanings. Existential /verificational 3rd person auxiliaries *ecmb*², *cymb*², *ecma*² are accented words with a free linear distribution. They can take clause-initial position, adjoin to initial proclitics *u*, *a*, *нѣ* or stand clause-internally. Sentences with standard non-existential & non-verificational forms *ecmb*¹, *cymb*¹, *ecma*¹ were assigned the value ‘PERFECT I’, sentences with existential or verificational *ecmb*², *cymb*², *ecma*² were assigned the value ‘PERFECT II’. Some contexts with clause-internal 3rd person perfect auxiliaries are two-way ambiguous between PERFECT I and PERFECT II. PERFECT I is a standard construction with person-and-number agreement, where the alternation $\emptyset^{3P} \sim ecmb^1, cymb^1, ecma^1$ is not semantically-driven. PERFECT II is possible only in the 3rd person and shows number, but not person agreement: *ecmb*², *cymb*², *ecma*² bring about the diagnostic existential or verificational semantics and do not alternate with \emptyset^{3P} .

3.5 The Ratio of Perfect vs Simple Past

All perfect clauses in 1-3 persons and all clauses containing aorist and imperfect forms were counted. The analytic plus perfect forms and other copular constructions with *l*-participles were not considered. The ratio of perfect clauses (P) from the total amount of three past tense forms (A & I & P), $k_{PERF} = P/A + I + P$ was measured in order to estimate, whether P had a prominent status in the idiom of each author.

³ By May 2022, the situation has changed: Monomakh’s “Edification” and his “Lenten Prayer” were added to RNC Old Russian corpus as separate entries, while the entire “Laurentius’ Chronicle” is no longer indexed as a part of RNC.

⁴ At this stage, syntactic annotation was done manually. The assignment of the values ‘+SUBJ’ and ‘-SUBJ’ according to the principle outlined in the section 3.3. requires a modification of the existing parsers.

3.6 Direct Quotations

The texts included several direct quotations from the Bible. In case these quotations included overt 3rd person perfect auxiliaries, the corresponding examples were considered borrowings from Old Church Slavonic and excluded from the sample.

4 Results and Discussion

4.1 Model A and Overt 3rd Person Auxiliaries

The results of this study partly falsify Zaliznjak's Model A. The zero BE-form \emptyset^{3P} is the main option for all authors, but two South Russian authors use overt 3rd person perfect auxiliaries in non-existential and non-verificational contexts in the beginning of the XII century. Daniil and Monomakh have 21 unambiguous examples of PERFECT I with *ecmb*¹, *cymb*¹, *ecma*¹ in clause-internal positions. Three Novgorod authors from the second half of the XII century have only 9 examples with overt 3rd person auxiliaries: the contexts are ambiguous between PERFECT I and PERFECT II readings. In the body of "Kirik's Inquiries", overt 3rd person perfect auxiliaries are found only in the main part written by Kirik himself: his followers Savva and Elija do not use them in their supplements. XIV-XV century authors lack overt 3rd person perfect auxiliaries completely. The overall ratio of perfect clauses (k_{PERF}) in a text plays no role for the choice of overt 3rd person auxiliary. The figures are shown in tab. 1⁵.

	DAN	MON	KIR _{KO}	KIR _{AU}	S & E _{KO}	S & E _{EAU}	EL-JO	ANT	STEF	VAS	FLOR	ROME
Date	ca. 1104 – 1106	ca. 1097 – 1117	ca. 1132 – 1156				1166	ca. 1200	1347-1349	1347	1438-1439	1438
k_{PERF}	20.3%	20.3%	13.2%	11.4%	17.4%	16.9%	67.4%	56.9%	4.4%	22.6%	26.5%	71.4%
Dialect	Chernigov	Perejaslavl	Novgorod								Suzdal	
Overt 3 rd person marker	18	3	5	3	0	0	2	2	0	0	0	0
\emptyset^{3P}	72	17	12	9	13	11	13	80	6	15	45	7
%	22.5%	15%	41.4%	25%	0%	0%	13.3%	2.4%	0%	0%	0%	0%

Tab. 1. 3rd person perfect auxiliaries by XII century authors compared to XIV–XV century texts.

These figures suggest that overt 3rd person auxiliaries in PERFECT I were initially an optional feature of early South Russian dialects, i.e. an inherited feature by the G1 / OR-1 grammar of some XII century authors and not only a learned feature of G2/L2 grammar. In the beginning of the XII century Daniil still could select overt 3rd person auxiliaries in 22, 5% of PERFECT I clauses, but this variant of PERFECT I was rapidly declining.

4.2 PERFECT II as an L1 FEATURE

PERFECT II was a construction, where accented auxiliaries *ecmb*², *cymb*², *ecma*² did not alternate with \emptyset^{3P} . Daniil regularly used it in existential contexts and combined *ecmb*², *cymb*² with the *l*-participle of the verb *быти* 'be' 31 times in sentences like (1): this peculiar combination of present and past BE-forms indicates a change of state and/or a canceled result.

⁵ The abbreviations read: DAN = Daniil, MON = Monomakh, KIR = Kirik, S & I = Savva and Elija, EL-JO = Elija-John, ANT = Antonius, STEF = Stefan, VAS = Vasiliy Kalika, FLOR = "Pilgrimage to the Florence Council", ROME = "A Note about Rome", KO = "Novgorodskaja Kormchaja 1282 A.D.", AU = "Autonomous version of 'Kirik's Inquiries'".

- (1) **И есть**²_{3SG} **был**_{AUX.PAST.SG.M} **град**_{NOM.SG.M} Тивириада великъ велми (Daniil, LXXXII).
'And the city of T. **was once** very big <but **deteriorated** since that.>'

XII century Novgorod authors have 9 ambivalent sentences with overt 3rd person perfect auxiliaries in clause-internal positions: they should be rather interpreted as PERFECT II, than as PERFECT I. None of the Novgorod examples shows the combination *есть*² + *быль*, which is another relic preserved in the dialect of Daniil.

4.3 Model B and Clausal Subjects

The retrieved data does not fully support Zaliznjak's Model B. South Russian authors of the XII century (Daniil, Monomakh) do not have complementary distribution of overt 3rd person PERFECT I auxiliaries and grammatical subjects. A vast majority of PERFECT I clauses has the + SUBJ feature. In Daniil's text, a filled subject slot has no impact on the choice of the non-zero 3rd person marker: the expectancy is 20% both for + SUBJ and for - SUBJ clauses. The data from XII century Novgorod is scanty, but compatible with the prediction of Model B that the expectancy of overt 3rd person marker increases in – SUBJ perfect clauses.

	DANIIL		MONOMAKH		KIRIK _{КО}		KIRIK _{АО}		ELIJA-JOHN		ANTONIUS	
DATE	ca. 1104 – 1106		ca. 1097 – 1117		ca. 1132 – 1156		ca. 1132 – 1156		1166		ca. 1200	
DIALECT	Chernigov		Perejaslavl		Novgorod							
	+ SUBJ	-SUBJ	+ SUBJ	-SUBJ	+ SUBJ	-SUBJ	+ SUBJ	-SUBJ	+ SUBJ	-SUBJ	+ SUBJ	-SUBJ
OVERT 3 rd PERSON MARKER	15	3	2	1	1	4	0	3	0	2	1	1
∅ ^{3P}	60	12	15	2	7	5	6	3	4	9	50	30
%	20%	20%	11.8 %	33.3%	12.5%	44.4%	0%	50%	0%	18.8%	2%	3.2%

Tab. 2. Grammatical subjects and overt 3rd person perfect auxiliaries in the XII century Russian

The contrast between South Russian and Novgorod authors of the XII century captured by tab. 2 can be explained by two factors. One of them is geographical. Different Old Russian dialects could have different settings of the parameter (ii) predicting the distribution of grammatical subjects in Old Russian idioms with overt 3rd person perfect auxiliaries. Another one is chronological. During the XII century some Russian dialects lost overt 3rd markers in PERFECT I. In both cases, Zaliznjak's Model B must be adjusted to specific data.

4.4 PERFECT I as an L1 FEATURE

Two relevant characteristics of G2/ OR-2 — the use of the simple past and overt 3rd person perfect auxiliaries — are displayed differently in the sample. Aorist and imperfect are chronologically stable and displayed uniformly in all dialects, since it is a learned L2 feature imported by all XII — XV centuries OR-2 texts. Meanwhile, PERFECT I with overt 3rd person perfect auxiliaries is a feature that was adapted only by those XII authors who had in their L1 grammar.

5 Conclusions

We discussed two linguistic models predicting language-internal variation in Old Russian syntax and checked them in a historical corpus consisting of several XII century author texts representing three dialects and seven different idioms. Linguistic models claiming a rigid distribution of language features or clustering of linguistic parameters activate many hidden text/corpus characteristics such as profiles of text genres, chronology and vector of change, ± impact of L2, ± presence of supra-dialect features,

etc. In this case they can be valued and checked on text samples, where the genre features are relatively stable, while location and time vary.

Acknowledgements

I am grateful to the anonymous reviewers for the valuable comments. The sole responsibility for the shortcomings is mine. This paper has been written with support from the project "Parametric description of the languages of Russian Federation" realized at Pushkin State Russian Language Institute.

References

- Henry, Alison. Dialect variation, optionality and the learnability guarantee // *Linguistica Atlantica*. — 1998. — Vol. 20. — P. 51–71.
- [1] Khaburgaev, Georgi A. The Fate of the Auxiliary Verb in Old Slavic Analytic Forms in Russian [Sud'ba vspomogatel'nogo glagola drevnih slavjanskih analiticheskikh form v russkom jazyke] // *The Bulletin of Moscow University. Series 9. Philology*. — 1978. — N 2 [Vestnik Moscovskogo universiteta. Ser. 9. Filologija. — 1978. — N 2].
- [2] Zaliznjak, Andrej A. On the Language of Birch Bark Letters [K izucheniju jazyka berestjanyh gramot] /Janin, Valentin L., Zaliznjak, Andrej A. Novgorod Birch Bark Letters. From the Excavations 1984-1989 [Janin V.L., Zaliznjak A.A. Novgorodskie gramoty na bereste. Iz raskopok 1984–1989 godov. Moscow: Nauka, 1993. — P. 191–319.
- [3] Zaliznjak, Andrej A. Old Novgorod Dialect. 2nd ed. [Drevnenovgorodskij dialekt. 2-e izdanie. — Moscow: Jazyki slavjanskoj kul'tury, 2004.]
- [4] Zaliznjak, Andrej A. Old Russian Enclitics [Drevnerusskie enklitiki]. — Moscow: Jazyki slavjanskoj kul'tury, 2008.
- [5] Zimmerling, Anton V. (2018). Two Dialects of Russian Grammar: Corpus Grammars and Formal Models [Dva dialekta drevnerusskoj grammatiki: korpusnye grammatiki i model'] *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2018"* [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2018"], Moscow, pp.818–830.
- [6] Zimmerling, Anton V. (2020). Zero forms in morphological paradigms: the verb "BE" in Russian // *Computational linguistics and intellectual technologies*, Issue 19 (26). Proceedings of the international conference "Dialogue 2020" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2020", Moscow, pp. 795 - 810].
- [7] Zimmerling, Anton V. From Integral Frameworks to Aspective Description [Ot integral'nogo k aspektivnomu]. — Moscow-St.Petersburg: Nestor-Istoria, 2021.

References

- [1] Henry, Alison. Dialect variation, optionality and the learnability guarantee // *Linguistica Atlantica*, 1998. — Vol. 20. — P. 51–71.
- [2] Zimmerling, Anton V. Zero forms in morphological paradigms: the verb "BE" in Russian // *Компьютерная лингвистика и интеллектуальные технологии*. Вып. 19 (26). Труды международной конференции «Диалог 2020» [Computational linguistics and intellectual technologies, Issue 19 (26). Proceedings of the international conference "Dialogue 2020", pp. 795–810].
- [3] Зализняк А. А. К изучению языка берестяных грамот /Янин В.Л., Зализняк А.А. Новгородские грамоты на бересте. Из раскопок 1984–1989 гг. — Москва: Наука, 1993. — P. 191–319.
- [4] Зализняк А. А. Древненовгородский диалект. 2-е изд. Москва: Языки славянской культуры, 2004.
- [5] Зализняк А. А. Древнерусские энклитики. Москва: Языки славянской культуры, 2008.
- [6] Хабургаев Г. А. Судьба вспомогательного глагола древних славянских аналитических форм в русском языке // *Вестник Московского Университета. Сер. 9. Филология*. — 1978. — № 2.
- [7] Циммерлинг А. В. Два диалекта русской грамматики: корпусные грамматики и модель // *Компьютерная лингвистика и интеллектуальные технологии*. Труды международной конференции «Диалог 2018», Москва, с. 818–830 [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2018", Moscow, pp. 818–830].
- [8] Циммерлинг А. В. От интегрального к аспективному. — Москва-Санкт-Петербург: Нестор-История, 2021.

Human Communicative Responses to Different Modes of Gaze Management by the Robot

Zinina A. A.

Kurchatov Institute / Moscow, Russia
Russian State University for the
Humanities / Moscow, Russia
Moscow State Linguistic University /
Moscow, Russia
zinina_aa@nrcki.ru

Kotov A. A.

Kurchatov Institute / Moscow, Russia
Russian State University for the
Humanities / Moscow, Russia
Moscow State Linguistic University /
Moscow, Russia
kotov@harpia.ru

Zaidelman L. Y.

Kurchatov Institute / Moscow, Russia
Russian State University for the
Humanities / Moscow, Russia
zaydelman_ly@nrcki.ru

Arinkin N. A.

Kurchatov Institute / Moscow, Russia
Russian State University for the
Humanities / Moscow, Russia
arinkin_na@nrcki.ru

Abstract

We investigated communicative reactions of people ($N = 46$), while telling stories to two companion robots, who reacted differently to the human gaze (head turning). In response to a human gaze the “aversive” robot averted its gaze away from the user, while the “responsive” robot, lifted its head and showed a responsive gaze. We found that users with high level of emotional intelligence prefer the gaze responsive robot and better recognize the difference between the robots. Thus, these users constitute the core group for the technology. In this paper, we further examine behavioral patterns of people in the experiment situation: (a) shift of attention to the story; (b) shift of attention to the robot; (c) joint attention. We also distinguish the communicative reactions of people, mainly from the core group, to the aversive and responsive gazes of the robots: positive responses to gaze contact and negative responses to gaze aversion. We show that for some users the responsive gaze behavior of the robot may serve as positive feedback, increasing the number of human iconic gestures, while telling a story to the responsive robot, and decreasing the number of iconic gestures in a story to the aversive robot.

Keywords: Multimodal communication, human computer interaction, emotional computing, companion robots, gaze contact, eyes contact

DOI: 10.28995/2075-7182-2022-21-594-602

1. Introduction

Human gaze has multiple functions and is controlled by different systems, first of all, by attention. On the one hand, a human turns his eyes to an object that attracts his attention and tries to orient himself by switching his gaze between objects in the environment [17; 20]. At the same time, eye movements can support both voluntary (endogenous) attention, caused by internal stimuli, and involuntary (exogenous) attention, imposed by the external situation [19]. On the other hand, gaze is an expressive mean: with the help of gaze a subject has the possibility to intentionally or spontaneously indicate to the interlocutor the object of his attention, to direct the gaze of the interlocutor – to participate in the situation of joint attention, or to demonstrate meaningful mimic patterns: to turn eyes, to *shoot* with eyes, thus, informing the interlocutor a negative or positive evaluation. The semiotic functions of nonverbal actions, including the direction of gaze, were considered in the classic work by Adam Kendon [10]. The connection between attention and the communicative functions of the gaze is studied by O.A. Fedorova, who with the help of

modern methods (using eyeglasses-trackers) describes the strategies of distribution of visual attention of participants of natural communication: *general*, *context-dependent* and *individual* [5]. The competition of the cognitive systems for eye control has not only a perceptual but also an important communicative function [1; 18; 16]. Eye movements are an iconic sign indicating the communicative intentions or states of the subject. At the same time, this sign is extremely ambiguous: a person can interpret a side-gaze of an interlocutor (a) as *thoughtfulness* or *thinking* or (b) as *inattention* or attention shift to another object [11; 12]. The eyelid and eyebrow system, controlled by muscles or action units [2] can also directly serve the attention mechanisms (such as squinting or opening eyes when looking at an object), or express cognitive and emotional states (such as opening eyes and raising eyebrows to express a surprise).

Coordinated simultaneous gaze at an object is described as the effect of *joint attention* [6], where the attention as a cognitive function is shared between two participants in communication [4]. Joint attention is one of the basic mechanisms of child's development in the aspects of speech and social skills. As a cognitive function, it is much more than simultaneous gazing at an object – it involves shared intentional attitude toward the world [9] and includes the skills of (a) detecting the attention of another person, (b) attention management, (c) social coordination, and (d) theory of mind. Modeling of joint attention is an extremely promising area for effective human-machine interaction, and is widely investigated in real communications with robotic and virtual agents [3]. The use of artificial agents provides greater ecological validity as compared to classical protocols using a screen with pictures or schematic eyes – the gaze-cueing paradigm, for example, as in [4; 8].

We use the situation of interaction with the experimental robot companion F-2 to position a person in a certain pattern of communicative attention distribution by the robot interlocutor. The distribution of the person's attention, his communicative reactions, and a description of his experience after the experiment can serve to evaluate the person's response to different modes of gaze control by the robot.

2. Robotic responsive gaze experiment

We conducted an experiment in which we investigated whether different robot's responsive gaze could, on the one hand, change the attractiveness of the robot to the user and, on the other hand, cause certain behavior in users. We expected, that people with high emotional intelligence would be more sensitive to the robot with responsive gaze – they would prefer this robot and may demonstrate the responsive behavior in gestures and gaze, as compared to their interaction with gaze aversive robot. Within the experiment, subjects had to tell two robots 6 stories following a list of pictures by Herluf Bidstrup (Fig. 1). Each picture was represented as a stack of cards in random order. Before the experiment, the F-2 robots were introduced to the subjects. The main purpose of the study – to investigate the effect of the robot's responsive gaze – was not reported to the respondents. They were told that the developers were trying to train the robot to follow the story narrated by a human. After the experiment each participant filled out a questionnaire, describing his/her experience with the robots. We also used the *EmIn* – Emotional Intelligence Test [13; 14] to evaluate the level of emotional intelligence of the participants. The test consists of 46 questions; answers are scored and contribute to 4 scales, where the most important scale, following the hypothesis of the study, is *Understanding of Emotions of Others*.

2.1. Experimental conditions

The robots responded to human communication in two different ways. When a person looked away from the robot (e.g., looked aside or looked at the table to follow the cards), each of the robots looked down – at the table with the cards. When a person looked at the robot, the first robot (marked with a square) looked back: raised its head, opened its eyelids, and raised its eyebrows, while the second robot (marked with a triangle) demonstrated side gaze, looking left or right in random order (Table 1).

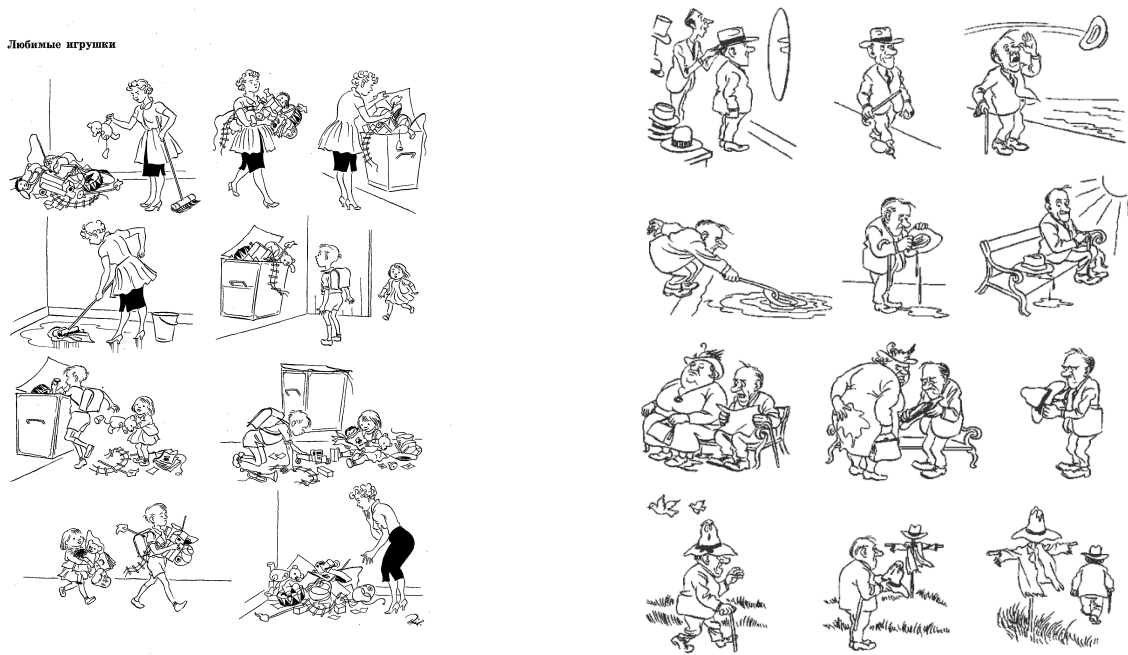


Figure 1: Example of stimulus material – a graphic story by Herluf Bidstrup. Each story was represented as a stack of separate cards in random order.




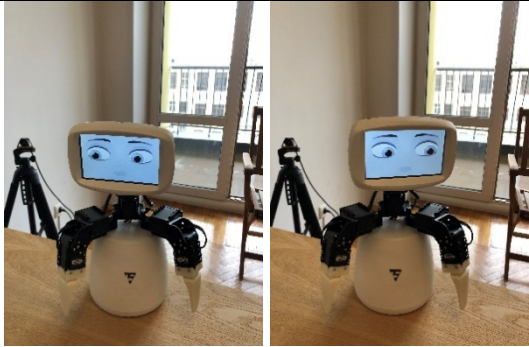
	Before or after the user's gaze	During the user's gaze
Gaze-responsive robot (marked with a square)		
Gaze-aversive robot (marked with a triangle)		 <p>Left- or right-side gazes are selected randomly</p>

Table 1. Robots' reactions to human gaze: gaze responsive and gaze aversive behavior

2.2. Participants

A total of 46 subjects participated in the experiment (mean age of 27 years, 33 females). Three main attention zones were suggested in the experiment settings: participants could place cards on the table or show them to the left or right robot. Subjects could communicate with the left and right robot from the same position at the table, so the robots constantly maintained their respective behavior (Fig. 2).



Figure 2: Experimental situation: (a) from the perspective of a participant and (b) front view

2.3. Robot movement and gaze recognition system

The robot control system controlled the robot's movement during the entire experiment. During greetings and speech utterances, the robot's head and arms were controlled by gestures coordinated with its speech production. During listening to a story, the robot's arms were controlled by the inactivity imitation component: random automanipulations or imitations of breathing have been performed. The vector of the user's attention was automatically identified by the orientation of the user's face and was recognized by a specially developed computer vision component based on OpenCV. This system had been preferred to an eye-tracker as a possible "built-in" solution for emotional companion robots, allowing us to avoid any calibration procedure, and thus, maintaining more natural communication. When the system had been recognizing a change in the user face vector, it had been translating messages like 'person looks_at egocentric(robot)' or 'person looks away'. The first message caused the "responsive" robot to look back at the user and the "aversive" robot – to look away. When the user's gaze was moved away from a robot, this robot showed a downward gaze on the table.

3. Results

We found that subjects' emotional intelligence influences their preference of the robot, as well as their ability to recognize the difference in their gaze behaviors. Participants with high scores on the *Understanding of Emotions of Others* scale are better at recognizing the difference of gaze patterns demonstrated by the two robots ($p < 0.05$, Mann-Whitney U-test). In the group that noticed the difference, the level of emotional intelligence was 47.8 (SD 7.8), in the group that did not notice the difference, the level of emotional intelligence was 38.7 (SD 10.4). The number of people who have correctly identified the difference between two robots was: 8% (1 of 12 persons) among people, who have preferred the gaze avoiding robot, 23% (5 of 21 persons) among people, who have evaluated the robots equally, and 69% (9 of 13 persons) among people, who have preferred the gaze responsive robot. It means that people, recognizing the difference between gaze patterns of the robots significantly often ($p < 0.01$, Spearman correlation) prefer the robot with gaze responsive behavior. The results in more

details are described in [21]. In our further studies of people communicative behavior, we include all the participants, but consider as the core group people with high emotional intelligence, able to distinguish robot's gaze patterns. As we expect, these people can become the core user group for future companion robots.

3.1. Human communicative reactions to robot behavior

The vast majority of the subjects (63.0%, $n = 29$) reported that they noticed a difference between the robots, however, only a third (32.6%, $n = 15$) were able to describe this difference accurately – as a difference in the direction of the robots' gaze. Some subjects did notice the difference in the robots' movements, but interpreted the robots' actions differently. For example, one of the subjects noted that the gaze avoiding robot was *following the card with his gaze* – this could happen when the subject looked at the robot and showed the picture (moved it aside of his body), and the robot reacted to the gaze and moved its gaze sideways, as if following the picture. Another subject interpreted the robot's head-down action as an appropriate reaction of the robot to the sad events in a story. While these actions constitute some exceptions, we can identify some more general patterns human gaze distribution when interacting with robots with different types of gaze behavior.

(1) **Shift of attention to the story (side-gaze).** Most subjects (67.3%, $n = 31$) could not accurately describe the difference between the robots. Many subjects concentrated on the subtask *construct a story from the cards*, they told stories without showing any gestures and only occasionally looking at the robots (Fig. 3). In turn, the automatic system could not register a sufficient number of incoming gazes and the number of robot's reactions and differences between the robots decreased accordingly: both robots looked down in front of them. This communicative behavior could be caused by several reasons: (a) the subjects were drawing parallels to the typical *construct a story from the cards* school task and believed that their creativity would be evaluated, (b) the subjects were involved in the events of the story and showed an up/side gaze, characteristic to thoughtfulness [11; 12], or (c) the subjects did not establish communicative contact with the robot due to its lack of communicative behavior.



Figure 3: Participants look forward/upward, aside from the robot while telling the story

(2) **Shifting attention to the robot – following the direction of the robot's attention.** During the planning of the experiment, the responsive robot was designed as the more natural version, and the avoiding robot was designed with a kind of communicative deficit. At the same time, 26.0% of respondents ($n = 12$) indicated that they liked the avoiding robot more, and in particular, that they liked *telling the story* to the avoiding robot (15.2%, $n = 7$), because, for example, *the reactions of the avoiding robot were more relevant*. The deficit in the control of the avoiding robot's gaze had been provoking diverse communicative reactions by subjects. When the subjects were showing the card to the robot, and the robot was responding with an avoiding gaze, many subjects moved the card following the robot's gaze, as if trying to adapt to the direction of the robot's attention. This interaction with the supposed direction of the robot's attention was characteristic both for subjects who have noticed and described

the difference between the robots (Fig. 4) and for subjects who did not describe the difference (Fig. 5). This allows us to assume that the interpretation of the direction of the interlocutor's attention may be implicit: it triggers the subject's behavioral response, related to handling of other's attention, but it is not noticed by the subject himself and not represented in the report.

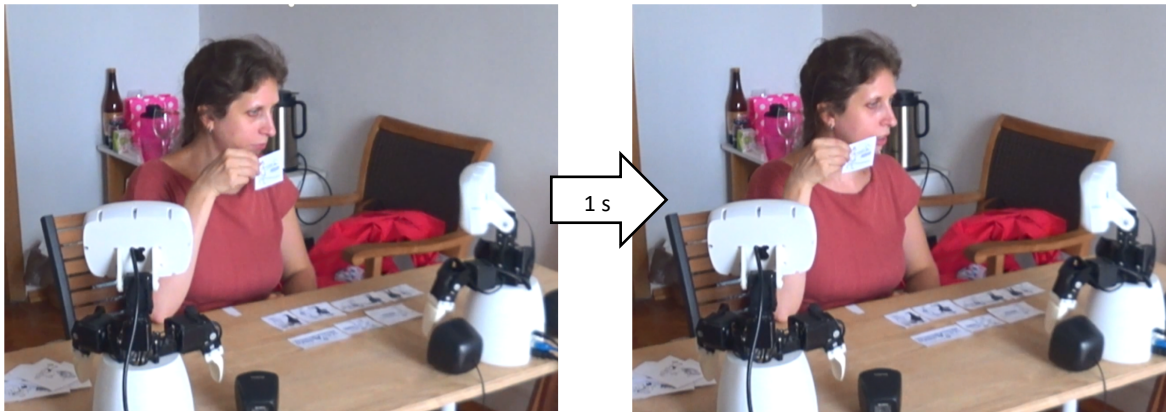


Figure 4: The participant places the card in front of the robot head and, after the robot replies with the avoiding gaze, moves the card in the direction of its gaze. In this case, the participant accurately described the difference between the robots in her report.



Figure 5: The participant places the card in front of the robot head and, after the robot replies with the avoiding gaze, moves the card in the direction of its gaze. In this case, the participant was unable to describe the difference between the robots in her report.

(c) Joint Attention. Many subjects behaved as if trying to attract the attention of the robot through the mechanism of joint attention. As they were showing the card toward the robot, they simultaneously moved their bodies forward and looked at the card, as if trying to gain additional attention from the robot to the picture (Fig. 6).

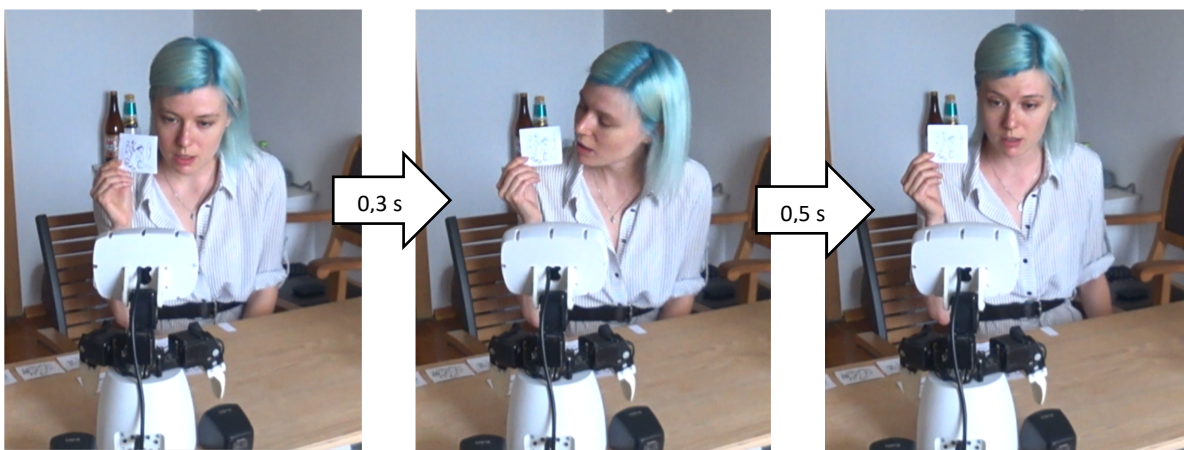


Figure 6: The respondent looks at the card facing the robot as if trying to get the robot's extra attention to the card

This mechanism, however, can be confused with another pattern, when subjects try to control the story and frequently peep into the card to follow the events. In several subjects this behavioral pattern is demonstrated at the first several stories, and is reduced to the end of the experiment, when they adapt to the experimental procedure.

3.2. Analysis of the behavior of the core group

As the core group for the communication behavior analysis, we selected subjects who noticed a difference in robots' behavior (32.6%, n = 15), and who have also been characterized by a higher level of emotional intellect (*Understanding of Emotions of Others* score of EmIn test). These subjects showed different types of communicative responses when interacting with the robots.

(a) No communicative response. These subjects well controlled their communicative behavior, narrated stories while steady holding the card and showed no communicative gestures. In this case no difference between the subjects was observed during the interaction with different robots. This pattern is characteristic for subjects, who have demonstrated numerous auto-manipulations during the preparation of the story (when the cards are on the table but no communication with the robot is established yet), but not on the stage of storytelling.

(b) Positive response to the robot's actions. The subjects in this group accompanied their stories with expressive (mostly, iconic) gestures. The number of such gestures decreased over the course of three stories to the avoiding robot, and increased over the course of three stories for the responsive robot (Fig. 7, 8).

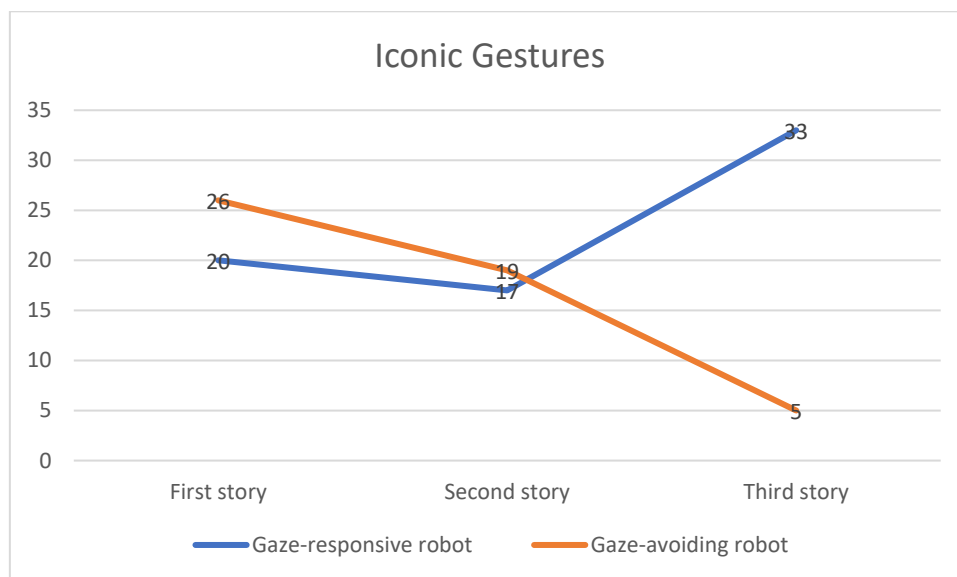


Figure 7: The number of iconic gestures in the stories – Participant №40 (female, 31)

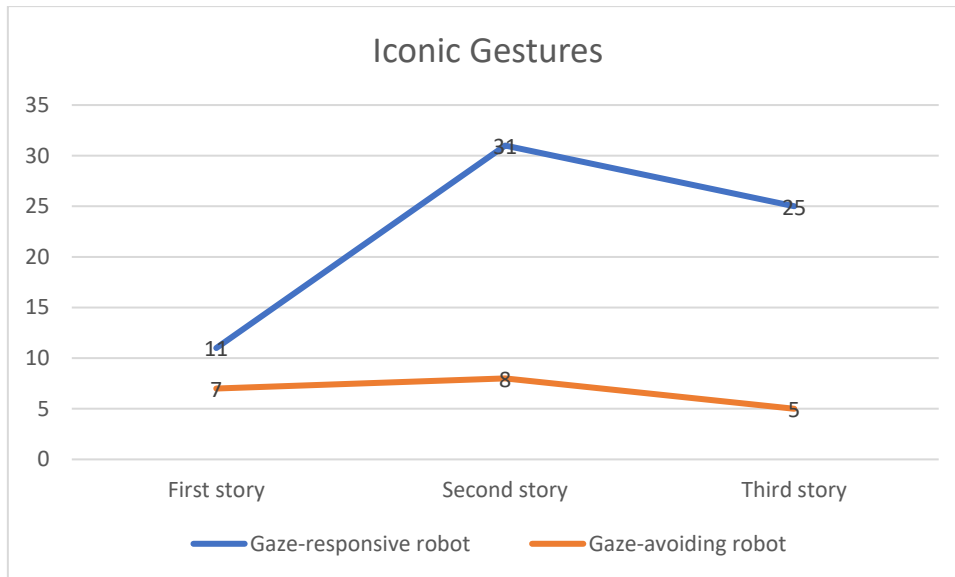


Figure. 8: The number of iconic gestures in the stories – Participant №24 (female, 30)

This behavior allows us to claim that some subjects receive communicative confirmation of attention from the responsive robot and use iconic gestures when telling him a story, while the lack of attention feedback from the avoiding robot causes them to reduce the number of gestures when telling him a story.

(c) Negative response to robot' deficit. Subjects tell the story neutrally to the responsive robot, but react negatively to the lack of attention from the avoiding robot. For example, subject №41 (female, 33) tries to get the attention of the avoiding robot before the beginning of each task: (1) she makes circular movements with the card and says *Are you listening to me?*, (2) while showing the card, she asks *Are you ready to listen to me? Are you looking at me?* and (3) while moving her gaze to the avoiding robot: *And you, are you ready to listen to my story?* Subject №35 (male, 35) tries to address the experimenter's assistant: *I don't think he's looking at me!*, and then – *Should I hold the card on the robot's right side then or on the left too? He [avoiding robot] keeps looking to the right.* Participant №38 (female, 22) tries to get feedback from the robot after the story – she holds up the card and says *Look!*, expecting a responsive gaze (or other feedback) from the robot. These reactions indicate expressed frustration during the participants' interaction with the robot with communicative deficit.

4. Conclusion

Interactive gaze management for emotional companion robots constitutes a promising feature. While only a minority of people (32.6%, $n = 15$) recognize the difference in robot's gaze patterns, these people also prefer the gaze responsive robot and are characterized by higher level of emotional intelligence. They may constitute the core user group for future emotional companion robots. These people may adjust their behavior in communication with the two robots: negatively reacting to the deficit of the gaze avoiding robot and/or positively reacting to the gaze responsive robot, in particular, by increasing the number of iconic gestures in the communication. People may try to adjust to the presumed attention of the robot by moving the card (object of the talk) after the robot's gaze – this may apply even to the people, who do not report the difference between the robot gaze patterns. People may also try to manipulate the attention of the robot through the mechanisms of joint attention – while showing a card and simultaneously looking at it. These results show that people (mostly with high level of emotional intelligence) successfully recognize natural gaze behavior patterns of companion robots and may demonstrate natural responsive cues – in gestures and joint attention. Thus, the robot may be perceived as a natural companion partner in the aspect of compound gaze behavior.

Acknowledgements

The research was in part supported by the grant of the Russian Science Foundation No. 19-18-00547, <https://rscf.ru/project/19-18-00547/>

References

- [1] Beyan C., Murino V., Venture G., Wykowska A. Computational Approaches for Human-Human and Human-Robot Social Interactions // *Frontiers in Robotics and AI*. — 2020. — Vol. 7.
- [2] Ekman P., Friesen W. Facial Action Coding System: A Technique for the Measurement of Facial Movement. — Palo Alto: Consulting Psychologists, 1978 — Vol. 3(2), pp. 5.
- [3] Chevalier P., Kompatsiri K., Ciardo F., Wykowska A. Examining joint attention with the use of humanoid robots-A new approach to study fundamental mechanisms of social cognition // *Psychonomic Bulletin & Review*. — 2020. — Vol. 27(2) — P. 217-236.
- [4] Falikman M.V. (2006) General psychology. In seven volumes: a student's course book for higher education institutes — Volume 4. Attention [Obshchaya psihologiya. V 7 t.: uchebnik dlya stud. vyssh. ucheb, zavedenij]. — M.: Academy.
- [5] Fedorova O.V. (2017), Distribution of visual attention of interlocutors in natural communication: 50 years later [Raspredelenie zritel'nogo vnimaniya sobesednikov v estestvennoj kommunikacii: 50 let spustya], *Cognitive Science in Moscow: New Research. Proceedings of the conference*, pp. 370-375.
- [6] Frischen A., Bayliss A. P., Tipper S. P. Gaze cueing of attention // *Psychological Bulletin*. — 2007. — Vol. 133. — P. 694-724.
- [7] Iriskhanova O.K., Cienki A. The semiotics of gestures in cognitive linguistics: Contribution and challenges // *Voprosy Kognitivnoy Lingvistiki*. — 2018. — Vol. 4. — P. 25-36.
- [8] Jording M., Hartz A., Bente G., Schulte-Rüther M., Vogeley K. The “social gaze space”: A taxonomy for gaze-based communication in triadic interactions // *Frontiers in Psychology*. — 2018. — Vol. 9.
- [9] Kaplan F., Hafner V. V. The challenges of joint attention // *Interaction Studies*. — 2006. — Vol. 7(2). — P. 135-169.
- [10] Kendon A. Some functions of gaze direction in social interaction // *Acta Psychologica*. — 1967. — V. 26. — P. 22-63.
- [11] Kotov A. A., Zinina A. A. (2015), Functional analysis of nonverbal communicative behavior [Funkcional'nyj analiz neverbal'nogogo komunikativnogo povedeniya], *Computational Linguistics and Intellectual Technologies [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii]*, M., V. 1(14), pp. 299-310.
- [12] Kotov A., Zinina A., Filatov A. Semantic Parser for Sentiment Analysis and the Emotional Computer Agents // *Proceedings of the AINL-ISMW FRUCT*. — 2015. — P. 167-170.
- [13] Lyusin, D.V. (2006) A New Technique for Measuring Emotional Intelligence: The EmIn Questionnaire [Novaya metodika dlya izmereniya emocional'nogo intellekta: oprosnik EmIn], *Psychol. Diagn.*, Vol. 4, pp. 3-22.
- [14] Lyusin D.V. (2009) EMIN Emotional Intelligence Questionnaire: New psychometric data [Oprosnik na emocional'nyj intellekt EmIn: novye psihometricheskie dannye], *Social and Emotional Intelligence: From Models to Measurements*, M., pp. 264-278.
- [15] Müller C. Gesture and sign: Cataclysmic break or dynamic relations? // *Frontiers in Psychology*. — 2018. — Vol. 9. — P. 1651.
- [16] Pagnotta M., Laland K.N., Coco M.I. Attentional coordination in demonstrator-observer dyads facilitates learning and predicts performance in a novel manual task // *Cognition*. — 2020. — P. 201.
- [17] Pannasch S., Schulz J., Velichkovsky B.M. On the control of visual fixation durations in free viewing of complex images // *Attention, Perception, and Psychophysics*. — 2011. — V. 73(4). — P. 1120-1132.
- [18] Velichkovsky B.M. Communicating attention: Gaze position transfer in cooperative problem solving // *Pragmatics & Cognition*. — 1995. — V. 3(2). — P. 199-223.
- [19] Velichkovsky B. M. (2006) *Cognitive Science: Fundamentals of the Psychology of Cognition [Kognitivnaya nauka: Osnovy psihologii poznaniya]*, M.: Akademia, Vol. 1.
- [20] Velichkovsky B. M., Korosteleva A. N., Pannasch S., Helmert J. R., Orlov V. A., Sharaev M. G., Velichkovsky B.B., Ushakov V. L. Two visual systems and their eye movements: A fixation-based event-related experiment with ultrafast fMRI reconciles competing views // *STM*. — 2019. — V. 11(4) — P. 7-16.
- [21] Velichkovsky B. M., Kotov A., Arinkin N., Zaidelman L., Zinina A., Kivva K. From Social Gaze to Indirect Speech Constructions: How to Induce Impression that Your Companion Robot is a Conscious Creature // *Appl. Sci*. — 2021. — V. 11(21). — P. 10255.

Abstracts

SODNER FOR RUSSIAN NESTED NAMED ENTITY RECOGNITION

Abrosimov K. I., Mosyagina A. G., Seldon, Nizhny Novgorod, Russia

The article describes the solution for Russian nested named entity recognition that we presented in the RuNNE competition. The solution is based on the Sodner model that predicts named entities in a text as a graph. During the competition we improved the training dataset and annotated the additional corpus that contains entities of the few-shot classes. After several experiments with different model parameters high macro F1 and few-shot F1 scores were obtained — 74.08 and 64.41 respectively.

ANALYZING COVID-RELATED STANCE AND ARGUMENTS USING BERT-BASED NATURAL LANGUAGE INFERENCE

Alibaeva K., Loukachevitch N., Lomonosov Moscow State University, Moscow, Russia

In this paper we present our approach for stance detection and premise classification from COVID-related messages developed for the RuArg-2022 evaluation. The methods are based on so-called NLI-setting (natural language inference) of BERT-based text classification (Sun et al., 2019), when the input of a model includes two sentences: a target sentence and a conclusion (for example, positive to masks). We also use translating Russian messages to English, which allows us to leverage COVID-trained BERT model. Besides, we use additional marking techniques of targeted entities. Our approach achieved the best results on both RuArg-2022 tasks. We also studied the contribution of marking techniques across datasets, tasks, models and languages of RuArg evaluation. We found that "</A:ASPECT> keyword </A:ASPECT>" gave the highest average increase over corresponding basic methods.

RUSSIAN ADVERBS OF FREQUENCY: A LEXICOGRAPHIC SKETCH

Apresjan V.^{1,2}, Shmelev A.^{2,3}, ¹HSE University; ²Vinogradov Russian Language Institute of the Russian Academy of Sciences; ³Moscow Pedagogical State University, Moscow, Russia

The paper is a corpus study of Russian frequency adverbs *chasto* 'frequently', *zachastuju* 'often', *redko* 'rarely', *izredka* 'rarely', etc. In Russian lexicographic tradition, frequency adverbs either lack separate entries and are explained via references to their adjectival counterparts or are treated exclusively as denotations of intervals between events. As our study demonstrates, this covers only a small fraction of their actual corpus usage. Many frequency adverbs can quantify over subjects, and thus resemble classical quantifiers such as 'many' or 'few'. Even when frequency adverbs quantify over predicates, they mostly refer not to intervals between events, but merely to their number. In some cases, they quantify over aspects of events, expressed by adjectives. There are also other important properties of Russian frequency adverbs missed by the dictionaries yet revealed by corpus analysis. Most frequency adverbs have a strong preference for topic or focus position, as motivated by their semantics. Some adverbs are preferable in generalized contexts, while others refer to specific events. Certain adverbs describe violations of the norm or undesirable events. Different adverbs quantify over different time periods: while some require a long time period, others may focus on very short stretches of time.

RUNNE-2022 SHARED TASK: RECOGNIZING NESTED NAMED ENTITIES

Artemova E. L.^{1,2}, Zmeev M. V.¹, Loukachevitch N. A.³, Rozhkov I. S.³, Batura T. V.^{3,4,8}, Ivanov V. V.⁵, Tutubalina E. V.^{1,6,7}, ¹HSE University, ²Huawei Noah's Ark lab, ³Lomonosov Moscow State University, ⁴Novosibirsk State University, ⁵Innopolis University, ⁶Kazan Federal University, ⁷Sber AI, ⁸Ershov Institute of Informatics Systems, Russia

The RuNNE Shared Task approaches the problem of nested named entity recognition. The annotation schema is designed in such a way, that an entity may partially overlap or even be nested into another entity. This way, the named entity "The Yermolova Theatre" of type ORGANIZATION houses another entity "Yermolova" of type PERSON. We adopt the Russian NEREL dataset (Loukachevitch et al., 2021) for the RuNNE Shared Task. NEREL comprises news texts written in the Russian language and collected from the Wikinews portal. The annotation schema includes 29 entity types. The nestedness of named entities in NEREL reaches up to six levels. The RuNNE Shared Task explores two setups. (i) In the general setup all entities occur more or less with the same frequency. (ii) In the few-shot setup the majority of entity types occur often in the training set. However, some of the entity types are have lower frequency, being thus challenging to recognize. In the test set the frequency of all entity types is even. This paper reports on the results of the RuNNE Shared Task. Overall the shared task has received 156 submissions from nine teams. Half of the submissions outperform a straightforward BERT-based baseline in both setups.

This paper overviews the shared task setup and discusses the submitted systems, discovering meaning insights for the problem of nested NER. The links to the evaluation platform and the data from the shared task are available in our github repository.

CORPUS EXPERIMENT IN FORENSIC LINGUISTICS

Baranov A. N., Russian Language Institute of RAS, Moscow, Russia

The report deals with current trends in cases concerning protection of honor, dignity and business reputation, which lead to the situations when courts use the category «abuse of law». Linguistic criteria are proposed that make it possible to identify linguistic signs of abuse of law in the text. One of the criteria is a corpus experiment. The proposed corpus experiment allows to compare the frequency of use of evaluation markers, opinions, assumptions, probabilities, etc. with the frequency of use of these forms in the representative corpus of the Russian language. The criterion of the corpus experiment is supplemented by the semantic criterion and the metatext criterion.

PROPER NAMES IN THE SCOPE OF METALINGUISTIC NEGATION

Boguslavsky I. M., A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia; Universidad Politécnica de Madrid, Madrid, Spain, **Vilinbakhova E. L.**, St. Petersburg State University, St. Petersburg, Russia

This paper analyses constructions with proper personal names negated metalinguistically, i.e. the objection is aimed not at the truth of the proposition (ne Vasja, a Kolja), but at the assertability of a linguistic expression (ne Vasja, a Vasilij Pavlovič). Our data from the Russian National Corpus shows, first, that metalinguistic negation constructions with proper names are used primarily to correct the representation of the referent which is associated with the particular form of naming, and the cases of correction of form only, such as linguistic mistakes,

are by far less typical. Next, among the additional elements there is a prevalence of temporal markers, which indicates that metalinguistic negation constructions are commonly used to point to the change of the referent with time. Another observation is that the correction towards a more formal way of naming prevails over the opposite direction, showing that the lack of formality is regarded as a more serious mistake than the other way around. Finally, it turned out that in our sample the examples of self-correction prevail over the correction of others, which demonstrates a wider range of functions of metalinguistic negation than it was acknowledged in the previous studies.

REFINING CRITERIA OF PARONYMY FOR BUILDING COMPUTER DICTIONARIES OF RUSSIAN PARONYMS

Bolshakova E. I., Lomonosov Moscow State University, HSE, Moscow, Russia, **Telegina A. D.**, Lomonosov Moscow State University, Moscow, Russia

Paronyms are words that have some similarity in sounding and spelling, but differ in meaning and usage (e.g., *sensitive—sensible, излишек—излишество*). In morphologically rich languages like Russian, paronymy is rather frequent phenomenon and one of the sources of speech difficulties. However, known dictionaries of Russian paronyms are not complete enough to help language learning or to support automatic correction of paronymy errors, and they do not provide precise definition of paronymy, which is necessary for constructing more extensive computer dictionaries. Aiming to clarify the concept of paronymy and to refine the previously proposed formal affix criterion of paronymy, we have performed a statistical study of paronyms taken from two printed dictionaries of Russian paronyms. Formal and semantic similarity of paronymy pairs were numerically estimated across various dimensions: proximity in affixes, in sounding, and in word meanings (the latter with the aid of neural models of distributive semantics and with an extensive base of Russian word combinations). Based on results of the study, refined criteria of paronymy and thresholds were proposed, which can be useful to automatically construct computer dictionaries of Russian paronyms, as well to replenish them by diagnostic contexts.

CONTRASTIVE FINE-TUNING TO IMPROVE GENERALIZATION IN DEEP NER

Bondarenko I., Novosibirsk State University, Novosibirsk, Russia

A novel algorithm of two-stage fine-tuning of a BERT-based language model for more effective named entity recognition is proposed. The first stage is based on training BERT as a Siamese network using a special contrastive loss function, and the second stage consists of fine-tuning the NER as a "traditional" sequence tagger. Inclusion of the contrastive first stage makes it possible to construct a high-level feature space at the output of BERT with more compact representations of different named entity classes. Experiments have shown that this fine-tuning scheme improves the generalization ability of named entity recognition models fine-tuned from various pre-trained BERT models. The source code is available under an Apache 2.0 license and hosted on GitHub https://github.com/bond005/runne_contrastive_ner

THE DATASET FOR PRESUICIDAL SIGNALS DETECTION IN TEXT AND ITS ANALYSIS

Buyanov I., Sochenkov I., FRC CSC RAS, Moscow, Russia

The paper says about dataset for presuicidal signal detection in Russian posts from social media. To the best of our knowledge, it is a first dataset of a such type for this language. We develop a collection methodology and conduct linguistic analysis of completed dataset. We also build a classification baseline with machine learning models to solve the detection task.

DISCOURSE-AWARE TEXT CLASSIFICATION FOR ARGUMENT MINING

Chistova E., Smirnov I., FRC CSC RAS / Moscow, Russia

We show that using the rhetorical structure automatically generated by the discourse parser is beneficial for paragraph-level argument mining in Russian. First, we improve the structure awareness of the current RST discourse parser for Russian by employing the recent top-down approach for unlabeled tree construction on a paragraph level. Then we demonstrate the utility of this parser in two classification argument mining subtasks of the RuARG-2022 shared task. Our approach leverages a structured LSTM module to compute a text representation that reflects the composition of discourse units in the rhetorical structure. We show that: (i) the inclusion of discourse analysis improves paragraph-level text classification; (ii) a novel TreeLSTM-based approach performs well for the computation of the complex text hidden representation using both a language model and an end-to-end RST parser; (iii) structures predicted by the proposed RST parser reflect the argumentative structures in texts in Russian.

RUSSIAN VERBAL AFFIXATION IN MENTAL LEXICON: PRIMING STUDY AND ITS ONLINE REPLICATION WITH TRUE AND STEM-MODIFIED RELATIVE PRIME VERBS

Chuprina A. O., NRU HSE, Moscow, Russia

While suffixed and prefixed words share common lexical features with their base word in the mental lexicon, the two derivational processes have their own properties. Whether their differences are reflected in the mental storage of the group of relative words or not is one of the topical psycholinguistic questions. My experimental results indicate that memory representations of the derivatives differ: while between the stem and the suffixed relative, the relationship is closer and based on transparency of the derived meaning, the relationship between the stem and the prefixed derivative is rather formal. The results also signal that the decompositional route is not a preferred strategy in complex verb processing for a Russian speaker. I base this conclusion on the results of two in-person experiments and their online replicas. Additionally, the data suggest that lexical organization is modified through the aspectual information of family members. These findings need to be taken into account in future studies, both within psycholinguistic and computational fields, using verbal material of the Russian language.

RUSSE-2022: FINDINGS OF THE FIRST RUSSIAN DETOXIFICATION SHARED TASK BASED ON PARALLEL CORPORA

Dementieva D.^{1,5}, **Nikishina I.**¹, **Logacheva V.**¹, **Fenogenova A.**³, **Dale D.**¹, **Krotova I.**², **Semenov N.**², **Shavrina T.**^{3,4}, **Panchenko A.**¹, ¹Skolkovo Institute of Science and Technology (Skoltech), ²Mobile TeleSystems (MTS), ³SberDevices (Sber), ⁴AI Research Institute (AIRI), ⁵Technical University of Munich (TUM)

Text detoxification is the task of rewriting a toxic text into a neutral text while preserving its original content. It has a wide range of applications, e.g. moderation of output of neural chatbots or suggesting less emotional version of posts on social networks. This paper provides a description of RUSSE-2022 competition of detoxification methods for the Russian language. This is the first com-

petition which features (i) parallel training data and (ii) manual evaluation. We describe the setup of the competition, the solutions of the participating teams and analyse their performance. In addition to that, the large-scale evaluation allows us to analyse the performance of automatic evaluation metrics.

EVIDENTIALITY AND EPISTEMIC MODALITY IN THE SEMANTICS OF THE GERMAN VERBS *SOLLEN* AND *WOLLEN* (BASED ON THE DATA FROM THE GERMAN-RUSSIAN PARALLEL CORPUS)

Dobrovolskij D. O.^{1,2,4}, **Zalizniak Anna A.**^{2,3}, ¹Russian Language Institute of the RAS; ²Institute of Linguistics; ³Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the RAS Moscow, Russia; ⁴Stockholm University, Stockholm, Sweden

Based on the analysis of the semantics of the German modal verbs *sollen* and *wollen* and their translational equivalents identified in the German-Russian parallel subcorpus of the Russian National Corpus (RNC), the article demonstrates that the categories of evidentiality and epistemic modality should be considered as having an independent status. At the same time, these two linguistic meanings can be expressed simultaneously. We consider possible combinations of the types of evidential and epistemic meaning components expressed by these German verbs. We clarify the classification of indirect evidentiality types by introducing the third intermediate type—reportative-inferential evidentiality, i.e. a conclusion made by the speaker based on the interpretation of someone else's utterance. Addressing the parallel corpus made it possible, on the one hand, to distinguish between the types of evidential meanings expressed by the verbs *sollen* and *wollen*, on the other hand, to clarify the semantics and identify the potential polysemy of the Russian translation equivalents.

RUCOCO: A NEW RUSSIAN CORPUS WITH COREFERENCE ANNOTATION

Dobrovolskii V. A., ABBYY, Moscow, Russia, **Michurina M. A.**, MIPT, RSUH, Moscow, Russia

We present a new corpus with coreference annotation, Russian Coreference Corpus (RuCoCo). The goal of RuCoCo is to obtain a large number of annotated texts while maintaining high inter-annotator agreement. RuCoCo contains news texts in Russian, part of which were annotated from scratch, and for the rest the machine-generated annotations were refined by human annotators. The size of our corpus is one million words and around 150,000 mentions. We make the corpus publicly available.

PROSODY AND POLYSEMY IN RUSSIAN DISCOURSE FORMULAE

Duryagin P. V., Ca' Foscari University of Venice / Venice, Italy

The paper presents the first descriptive approach to the prosody of Russian discourse formulae using the methods of experimental phonetics. The prosody of a frequent idiomatic formula *da nu* was studied. The analysis of pitch contours revealed that this unit can be marked by two tonal configurations: the falling one that can be identified as the IK-2 of the traditional holistic approach, and the rising one that is distinct from IK-3 and apparently contains a unique high tonal target followed by an irregularly truncated low boundary tone. The usage of these configurations is distributed unevenly and depends on the additional pragmatic meanings set by the dialogical context. In addition, the data indicates that some pragmatic meanings can be disambiguated by segmental duration. The subjects used significantly longer vowels in *da nu* to express surprise, compared to the rejection of new information. Meanwhile, the expression of doubt takes an intermediate position in this pragmatic continuum, marked by a lengthened pre-stressed and a shortened stressed vowel.

MOTION VERBS IN MULTIMODAL COMMUNICATION

Evdokimova A.¹, **Nikolaeva Y.**², **Budennaya E.**^{3,1}, ¹Institute of Linguistics, Russian Academy of Sciences; ²Moscow State University; ³Higher School of Economics, Moscow, Russia

The article explores correlations between motion verbs and head and hands gestures using the RUPLEX corpus. The verbs are divided into four groups based on their meanings. Monological and dialogical parts of the recordings are compared along with the speaker's role and viewpoint in gestures. The pilot analysis of motion verbs in the multimodal corpus showed that the relationships between verb type, non-verbal behavior and speaker's role depend on a complex set of factors and manifests itself in different ways in different channels. In the verbal channel no direct relationship between the semantic type of the verb and the speaker's role was detected; however, the narrators and commentators who have seen the film used more affectional vocabulary than the reteller while the latter tended to use more vector-prefixed verbs. In manual channel prefixes or their absence do not influence the use of hand gestures. Transitive verbs meaning manipulations of different items are more probable to be illustrated by depictive gestures. Predictably, motion verbs in the strict sense are more prone to be supported by observer viewpoint (O-VPT) gestures, while verbs of manipulation are usually used with C-VPT gestures. In cephalic channel motion verbs in the strict sense (relocation of a character) are usually illustrated by O-VPT depictive gestures, and manipulation verbs are more probably supported by pantomime C-VPT gestures similar to manual channel. In some head gestures the viewpoint is combined. If the verb is repeated by the same or another speaker the gestures differ in both manual and cephalic channels. Cephalic gesture clusters on motion verbs have mostly a depictive function, which may be considered a gestural illustration.

LIGHTWEIGHT AND ACCURATE SYSTEM FOR ENTITY EXTRACTION AND LINKING

Evseev D. A., Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Entity extraction and linking components in dialogue assistants should meet the requirements of low resource consumption and high accuracy. In this paper we present lightweight system which extracts entity mentions from the text and finds corresponding Wikidata ids and Wikipedia pages links. Entity extraction and linking is performed into the following steps: extraction of entity substrings from the text, retrieval of candidate entities from Wikidata knowledge base and entity disambiguation. Entity extraction is based on RoBERTa-tiny model for token classification. Extracted substrings are classified into 42 fine-grained tags for filtering of candidate entities. Candidate entities are ranked by number of connections of candidate entities in the text in Wikidata knowledge graph. The proposed system outperforms on WNED-WIKI other lightweight solutions, such as REL and OpenTapioca. The system supports easy adding new Wikidata entities to the database and using other knowledge bases for entity linking.

CONTROLLABLE MULTI-ATTRIBUTE DIALOG GENERATION WITH PALS AND GROUNDING KNOWLEDGE

Evseev D. A., **Nagovitsin M. S.**, **Kuznetsov D. P.**, Moscow Institute of Physics and Technology Dolgoprudny, Russia

Today, neural language models are commonly employed for generation of natural like responses in dialog systems. The main issue that limits wide adoption of neural generation is related to poor predictability of responses in terms of content, as well as dialog attributes such as dialog acts and sentiment. In this paper we propose a method based on projected attention layers (PALS) for

controllable multi-attribute knowledge grounded dialog generation. We compared a number of methods for training and blending representations produced by PALs combined with DialoGPT base model. Results of our experiments demonstrate that separate pre-training of PAL branches for different attributes followed by transfer and fine-tuning of dense blending layer gives the highest accuracy of control of a generated response for fewer trainable parameters per an attribute. Furthermore, we applied our approach for controllable multi-attribute generation with grounding knowledge to Blenderbot model. Our solution outperforms the baseline Blenderbot and CRAYON model in control accuracy of dialog acts and sentiment on Daily Dialog as well demonstrates a comparable overall quality of dialog generation given grounding knowledge on Wizard of Wikipedia.

“PEARS FILM” LIVE: COGNITIVE PECULIARITIES OF THE REPORTAGE

Fedorova O. V., Lomonosov Moscow State University, Moscow, Russia

This study investigated whether there was a relationship between verbal working memory capacity and speech production in the genre of live reportage. The participants were 16 students of the Lomonosov Moscow State University. Participants' working memory capacity was assessed by means of the speaking span test [12]. Speech production was elicited by means of the “Pears Film” by W. Chafe [2]. Three aspects of speech production were assessed: continuity of the reportage, speech rate, and lexical diversity. Statistical analyses revealed that working memory capacity correlates positively with speech rate and lexical diversity, as predicted, and negatively with continuity of the reportage, contrary to predictions.

ARGUMENTATIVE TEXT GENERATION IN ECONOMIC DOMAIN

Fishcheva I. N.¹, **Osadchiy D.**², **Bochenina K. O.**², **Kotelnikov E. V.**¹, ¹Vyatka State University, Kirov, Russia; ²ITMO University, Saint Petersburg, Russia

The development of large and super-large language models, such as GPT-3, T5, Switch Transformer, ERNIE, etc., has significantly improved the performance of text generation. One of the important research directions in this area is the generation of texts with arguments. The solution of this problem can be used in business meetings, political debates, dialogue systems, for preparation of student essays. One of the main domains for these applications is the economic sphere.

The key problem of the argument text generation for the Russian language is the lack of annotated argumentation corpora. In this paper, we use translated versions of the Argumentative Microtext, Persuasive Essays and UKP Sentential corpora to fine-tune RuBERT model. Further, this model is used to annotate the corpus of economic news by argumentation. Then the annotated corpus is employed to fine-tune the ruGPT-3 model, which generates argument texts. The results show that this approach improves the accuracy of the argument generation by more than 20 percentage points (63.2% vs. 42.5%) compared to the original ruGPT-3 model.

AUTOMATIC SUMMARIZATION OF RUSSIAN TEXTS: COMPARISON OF EXTRACTIVE AND ABSTRACTIVE METHODS

Goloviznina V. S., **Kotelnikov E. V.**, Vyatka State University, Kirov, Russia

This paper investigates the problem of creating summaries of Russian-language texts based on extractive (TextRank and LexRank) and abstractive (mBART, ruGPT3Small, ruGPT3Large, ruT5-base and ruT5-large) methods. For our experiments, we used the Russian-language corpus of news articles Gazeta and the Russian-language parts of the MLSUM and XL-Sum corpora. We computed ROUGE-N, ROUGE-L, BLEU, METEOR and BERTScore metrics to evaluate the quality of summarization. According to the experimental results, the methods are ranked (from best to worst) as follows: ruT5-large, mBART, ruT5-base, LexRank, ruGPT3Large, TextRank, ruGPT3Small. The study also highlights the salient features of summaries obtained by various methods. In particular, mBART summaries are less abstractive than ruGPT3Large and ruT5-large, and ruGPT3Large summaries are often incomplete and contain errors.

RUSSIAN NOUN *PRICHINA CAUSE, REASON* REVISITED: CONSTRUCTIONS WITH *CHTO THAT*-CLAUSE

Goncharov A. A.¹, **Kobozeva I. M.**^{2,3}, ¹Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (FRC CSC RAS); ²Lomonosov Moscow State University; ³Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

The paper deals with Russian syntactic constructions *prichina*, *chto P* (lit. ‘cause that P’) and *prichina togo*, *chto P* with expletive pronoun *to* in genitive case, that are in most cases interchangeable. The purpose is to describe the syntactic and semantic properties of these constructions in Russian using corpus data. The semantics of the noun *prichina* ‘cause, reason’ and its argument structure were critically examined, and a representative set of examples where this noun is a part of one of these constructions was analyzed. The following results were obtained: 1) the description of the argument structure of the word *prichina* was supplemented; 2) the constraint on the use of *chto*-complements with deverbative nouns observed in [10] was shown to cover non-deverbative noun-predicate *prichina* ‘cause’; 3) it was shown that the semantic role of the *chto*-clause may be either Cause or Effect depending on the syntactic function of the construction; 4) it was determined that both meanings of the noun *prichina* (objective vs subjective cause) may be expressed in both constructions — with and without *to* — irrespective of syntactic function of the construction.

SUFFIXAL IMPERFECTIVATION OF PREFIXED VERBS: RECORD BREAKERS AND OUTSIDERS (BASED ON THE DICTIONARY, CORPUS AND RUNET)

Gorbova E. V., **Chuikova O. Iu.**, Herzen State Pedagogical University of Russia

The paper sums up the results of the study on the imperfectivability of the Russian prefixed verb using such sources of linguistic data as the Small Academic Dictionary, the corpus (RNC — ruscorpora.ru) and Runet. The focus is on those subsets of the set of prefixed perfectives that show specificity in relation to suffixal imperfectivation, differing in increased (record breakers) and decreased (outsiders) imperfectivability compared to the average level. The former are represented by denominatives and prefixed perfectives derived from perfective simplex stems, while the latter are represented by most Aktionsarten and verbs with the *-i(zi)rova-* formant. Systemic and morphological explanations for the specifics of these subsets are proposed.

RUSSIAN TEXTS DETOXIFICATION WITH LEVENSHTEIN EDITING

Gusev I., Moscow Institute of Physics and Technology, Moscow, Russia

Text detoxification is a style transfer task of creating neutral versions of toxic texts. In this paper, we use the concept of text editing to build a two-step tagging-based detoxification model using a parallel corpus of Russian texts. With this model, we achieved the best style transfer accuracy among all models in the RUSSE Detox shared task, surpassing larger sequence-to-sequence models.

THE ROLE OF PARAGRAPH IN THE CORPORA OF ANNOTATED TEXTS

Inkova O. Y.^{1,2}, **Nuriev V. A.**¹, **Popkova N. A.**¹, ¹Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; ²University of Geneva, Geneva, Switzerland

The paper focuses on the function of paragraph both in text organization and in text annotation from the point of view of coherence. Taking as examples three major types of corpora (the RST, ANNODIS, and PDTB corpora), it shows whether and to what extent the existing approaches account for the paragraph when a discourse relation gets annotated. Then it presents the theoretical principles underlying text annotation in two databases: the Supracorpora database of connectives and the Supracorpora database of hierarchical logical-semantic relations (a new linguistic resource). Text coherence is shown to result from the interaction of various discourse phenomena, acting at the level of local and global structures. In this approach, the paragraph is assigned to the meso-level, positioned between local and global levels. The researcher may analyze the internal organization of the paragraph, limiting oneself to the inter-sentential level. Yet, to analyze and describe how paragraphs follow one another in the text, it is necessary to operate at the supra-sentential level, adopting a conceptual apparatus fundamentally different from the one for the description of local text structure.

“WORD-BY-WORD” MELODIC CONTOUR IN RUSSIAN DIALECTS: QUANTITATIVE APPROACH

Knyazev S. V., Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia, **Evsstigneeva M. Y.**, Banks Soft Systems, Moscow, Russia

The paper presents results of quantitative analysis of phrasal tonal structure in two Northern Russian dialects with different types of “word-by-word” melodic contour. These dialects differ from Modern Standard Russian by the quantity of pitch accents since their 60% of words bear pitch accent, thus the prosodic unit in them is not a (phonological) word, but an accent group. In addition, the dialects differ from Standard Russian by regular presence of even tone on the accented vowel (in Arkhangelsk dialect 86% of all accents have it; in Vologda dialect it is less frequent: 33%) and higher frequency of pitch accents with increased interval. The main differences between Arkhangelsk and Vologda dialects are 1) the ratio of rising and falling pitch accents: 2.6% falling in Arkhangelsk dialect and 56% in Vologda dialect, it brings the latter closer to Standard Russian (53%) and 2) the level of the base tone on which the main tonal changes occur (high and medium, respectively). Thus “word-by-word” melodic contour exists at least in two varieties: with rising tonal movement and with a falling tone in the function of an ornamental accent. In general, the intonation system of Vologda dialect, though there are a lot of significant differences, is much closer to Modern Standard Russian than to Arkhangelsk dialect.

KNOWLEDGE DISTILLATION OF RUSSIAN LANGUAGE MODELS WITH REDUCTION OF VOCABULARY

Kolesnikova A.¹, **Kurатов Y.**^{1,2}, **Konovalov V.**¹, **Burtsev M.**^{1,2}, ¹Neural Networks and Deep Learning Lab, MIPT, Dolgoprudny, Russia; ²AIRI, Moscow, Russia

Today, transformer language models serve as a core component for majority of natural language processing tasks. Industrial application of such models requires minimization of computation time and memory footprint. Knowledge distillation is one of approaches to address this goal. Existing methods in this field are mainly focused on reducing the number of layers or dimension of embeddings/hidden representations. Alternative option is to reduce the number of tokens in vocabulary and therefore the embeddings matrix of the student model. The main problem with vocabulary minimization is mismatch between input sequences and output class distributions of a teacher and a student models. As a result, it is impossible to directly apply KL-based knowledge distillation. We propose two simple yet effective alignment techniques to make knowledge distillation to the students with reduced vocabulary. Evaluation of distilled models on a number of common benchmarks for Russian such as Russian SuperGLUE, SberQuAD, RuSentiment, ParaPhaser, Collection-3 demonstrated that our techniques allow to achieve compression from 17× to 49×, while maintaining quality of 1.7× compressed student with the full-sized vocabulary, but reduced number of Transformer layers only. We make our code and distilled models available.

EMOTIONAL ANALYSIS OF POSTS IN VKONTAKTE: CLASSIFICATION OR REGRESSION?

Kolmogorova A. V., **Kalinin A. A.**, Siberian Federal University, Krasnoyarsk, Russia

The article summarizes the results of two tasks in machine learning paradigm: the task of classification according to the criterion of dominating emotion on the data of social networks posts in Russian and the regression task using the same data. The experiments are conducted on the data set collected from VKontakte social network and consisted of 3879 posts assessed by 2000 informants on Toloka crowd sourcing platform. The annotation procedure was based on the original interface for non-discrete emotion assessment elaborated by researchers.

SPEAKER-AGNOSTIC MOUTH BLENDSHAPE PREDICTION FROM SPEECH

Korzun V.^{1,3}, **Gadecky D.**^{1,3}, **Berzin V.**^{1,3}, **Ilin A.**^{2,3}, ¹Moscow Institute of Physics and Technology; ²Skolkovo Institute of Science and Technology; ³Tinkoff, Moscow, Russia

This paper describes a simple end-to-end deep learning approach for automated 3D lip animation from audio. Our solution is speaker-independent, which means that once trained on one voice, the model can be applied to any voice without need for retraining. This solution only requires a small amount of data, which can be easily obtained with a modern iPhone. Along with that we also propose a new combined approach for evaluating blendshape prediction models.

RUARG-2022: ARGUMENT MINING EVALUATION

Kotelnikov E.¹, **Loukachevitch N.**², **Nikishina I.**³, **Panchenko A.**³, ¹Vyatka State University, Kirov, Russia; ²Lomonosov Moscow State University, Moscow, Russia; ³Skolkovo Institute of Science and Technology, Moscow, Russia

Argumentation analysis is a field of computational linguistics that studies methods for extracting arguments from texts and the relationships between them, as well as building argumentation structure of texts. This paper is a report of the organizers on the first competition of argumentation analysis systems dealing with Russian language texts within the framework of the Dialogue conference. During the competition, the participants were offered two tasks: stance detection and argument classification. A corpus containing 9,550 sentences (comments on social media posts) on three topics related to the COVID-19 pandemic (vaccination, quarantine, and wearing masks) was prepared, annotated, and used for training and testing. The system that won the first place in both tasks used the NLI (Natural Language Inference) variant of the BERT architecture, automatic translation into English to apply a specialized BERT model, retrained on Twitter posts discussing COVID-19, as well as additional masking of target entities. This system showed the following results: for the stance detection task an F1-score of 0.6968, for the argument classification task an F1-score of 0.7404. We hope that the prepared dataset and baselines will help to foster further research on argument mining for the Russian language.

GRAMMATICAL VARIATION: METHODS FOR QUANTITATIVE CORPUS ANALYSIS (THE GERMAN PREPOSITION WEGEN)

Krotova E. B.¹, **Tsvetaeva E. N.**², **Sharandin A. V.**³, **Dobrovolskij D. O.**^{4,1,5}, ¹Institute of Linguistics, RAS; ²Moscow State Linguistic University; ³Russian State University for the Humanities; ⁴Russian Language Institute, RAS, Moscow, Russia; ⁵Stockholm University, Stockholm, Sweden

The corpus study of grammatical variation is not considered a complex task anymore. However, the corpus annotations used to solve the task are based on specific theoretical assumptions which do not necessarily cover all possible variants found in empirical material. On the example of the German preposition *wegen* and its variation in case government this article shows why available tools in corpus managers can be insufficient. An experiment is presented in which data obtained from annotated corpora are automatically labeled using a developed algorithm. An analysis of the corpus material showed that the preposition *wegen* primarily governs the genitive (about 60% of all uses). The dative, which is recognized by modern linguistics as a conventional variant, is relatively rare (5%). Additionally, the analysis also revealed that for 35% of nouns following *wegen* case cannot be determined, a phenomenon which is not mentioned in the corresponding dictionary entries of all lexicographic sources known to us. Thus it is not possible to identify the proportion of such underdetermined instances of case government using only the existing annotation of the analyzed corpora, since it did not include tags for nouns with no clear case assignment.

SENTENTIAL ARGUMENTS OF MENTAL PREDICATES WITH CONJUNCTION *KOGDA WHEN* (ACCORDING TO THE RUSSIAN NATIONAL CORPUS)

Kustova G. I., Vinogradov Russian Language Institute of the Russian Academy of Sciences; Moscow Pedagogical State University, Moscow, Russia

It is known that the matrix predicates of emotion and evaluation subordinate sentential arguments not only with conjunction *chto* 'that': *Obidno / plokho, chto komanda proigrala* 'It's a shame / Too bad the team lost', but also with conjunctions *kogda / esli* 'when' / 'if': *Plokho, kogda / esli komanda proigryvaet* 'It's bad if the team loses'. The NRC data show that predicates of the other semantic classes that are not mentioned in grammars can also subordinate clauses with conjunctions *kogda* 'when' and *esli* 'if'.

The article discusses examples of *kogda*-sentences with mental predicates (*znat* 'to know', *pomnit* 'to remember', *ponimat* / *poniatno* 'to understand'): *Ya pomnyu, kogda po Borodinskomu mostu khodili tramvai* 'I remember when trams ran along the Borodino bridge'; *Ponyatno, kogda kletki formiruyutsya v khode razvitiya zarodyscha, no vo vzrosлом organizme?* 'It is clear when cells are formed during the development of the embryo, but in an adult organism?'

SWEET IDEAL

Levontina I. B., Russian Language Institute RAS

The work is devoted to the semantics, the structure of polysemy, the syntax of the word *ideal*. The word *ideal* at first glance does not seem complicated either from the point of view of the semantics and structure of polysemy, or from the point of view of the actant structure. *X is the ideal of Y from the point of view of Z* — means that the object X, either real or imaginary, belongs to class Y and fully corresponds to Z's idea of what an object of class Y should be. Moreover, this is such a complete correspondence that in the life of such practically does not happen. The paper considers some non-trivial features of this word. In particular, it turns out that the genitive form of the word *ideal* can occupy three positions: not only the ideal of Petya (=Petia's ideal) and the ideal of wife (the idea of what a wife should be), but also the ideal of service (*ideal sluzhenija*) — in the sense that service is the content of the ideal. In addition, it turns out that the semantics and structure of the polysemy of the word *ideal* have changed during its existence in the Russian language. In Pushkin's expression *Tatyany milyj ideal* the sweet ideal does not imply that Tatyana has some kind of ideal, but that the ideal is Tatyana herself. More precisely, the *ideal of Tatiana* here is the same as the *image of Tatiana*. Different words with this root have various structure of polysemy.

ARTIFICIAL TEXT DETECTION WITH MULTIPLE TRAINING STRATEGIES

Li B.¹, **Weng Y.**², **Song Q.**¹, **Deng H.**³, ¹College of Electrical and Information Engineering, Hunan University, Changsha, China; ²National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy Sciences, Beijing, China; ³Experimental High School Affiliated to Beijing Normal University, Beijing, China

As the deep learning rapidly promote, the artificial texts created by generative models are commonly used in news and social media. However, such models can be abused to generate product reviews, fake news, and even fake political content. The paper proposes a solution for the Russian Artificial Text Detection in the Dialogue shared task 2022 (RuATD 2022) to distinguish which model within the list is used to generate this text. We introduce the DeBERTa pre-trained language model with multiple training strategies for this shared task. Extensive experiments conducted on the RuATD dataset validate the effectiveness of our proposed method. Moreover, our submission ranked second place in the evaluation phase for RuATD 2022 (Multi-Class).

METHOD AND SOFTWARE MODEL FOR EVALUATING THE STATISTICAL CHARACTERISTICS OF A SPEECH MELODY

Lobanov B. M., **Zhitko V. A.**, United Institute of Informatics Problems NAS Belarus, Minsk, Belarus

A method for estimating the statistical characteristics of speech melody is proposed. The procedure of constructing histograms of the frequency distribution of the pitch frequency over sufficiently long intervals of speech is described. A distinctive feature of the method is that the discrete values of the pitch of speech are measured only at intervals of vowels. Two options of the pitch scales used for analysis of the melody characteristics are selected, namely: linear for speech and logarithmic for singing. A method for estimating three parameters of the histogram is proposed: register—R, range—D, asymmetry—A. Numerous examples are given showing the effectiveness of the proposed method in assessing the individuality of the melody of the speaker's speech, as well as his emotional state. A description of the prototype of the Speech Melody Meter (SMM) system is given (see also: <https://intontrainer.by>). SMM is a software implementation of the proposed method for assessing the statistical characteristics of speech melody.

DIALOG-22 RUATD GENERATED TEXT DETECTION

Maloyan N., **Nutfullin B.**, **Ilyshin E.**, Lomonosov MSU

Text Generation Models (TGMs) succeed in creating text that matches human language style reasonably well. Detectors that can distinguish between TGM-generated text and human-written ones play an important role in preventing abuse of TGM. In this paper, we describe our pipeline for the two DIALOG-22 RuATD tasks: detecting generated text (binary task) and classification of which model

was used to generate text (multiclass task) (Shamardina et al., 2022). We achieved 1st place on the binary classification task with an accuracy score of 0.82995 on the private test set and 4th place on the multiclass classification task with an accuracy score of 0.62856 on the private test set. We proposed an ensemble method of different pre-trained models based on the attention mechanism.

RUSSIAN NEURAL MORPHOLOGICAL TAGGING: DO NOT MERGE TAGSETS

Movsesyan A. A., Institute for Information Transmission Problems (Kharkevich Institute) Russian Academy of Sciences, Moscow, Russia

There are multiple morphologically annotated corpora of Russian available. They have different tagsets and annotation guidelines, which makes them difficult to use together. We proposed a neural morphological tagger for Russian based on multitask learning technique which is able to predict morphological tags of words for different tagsets. We evaluated our model on various corpora and showed that utilising multiple corpora without merging them not only improves tagging performance but allows for scalable indirect conversion between multiple tagsets in all directions. Furthermore, we also showed that treating each corpus separately is more efficient than merging the corpora even if they share the same tagset.

DETECTING AUTO-GENERATED TEXTS WITH LANGUAGE MODEL AND ATTACKING THE DETECTOR

Orzhenovskii M. V., Saint Petersburg, Russia

We propose a simple approach to the detection of automatically generated texts. A pre-trained language model, fine-tuned on the shared task's dataset, achieved 3rd place on the binary task leaderboard with 82.6% accuracy. In the multi-task leaderboard, the language model achieved an F1 score of 64.5% after being fine-tuned with the same procedure. In order to investigate the weaknesses of this approach, we explore two possible attacks on the detector: selecting from language model outputs and directed beam search. These attacks reduce the likelihood of detecting the generated texts without significant loss in quality. Both attacks do not require retraining the generative model and are applied at inference time.

RUSSIAN TEMPORAL CLAUSES IN TERMS OF SEMANTIC-SYNTACTIC INTEGRATION (A CASE STUDY OF A COORDINATING CONJUNCTION *KOGDA WHEN*)

Pekelis O. E., Russian State University for the Humanities/Moscow, Russia

The paper argues that the Russian temporal conjunction *kogda* 'when', which is traditionally treated as subordinating, may also be coordinating. Due to a very subtle semantic difference between the coordinating and subordinating instances of *kogda*, it is difficult to distinguish between them using the well-known criteria of coordination and subordination. This difficulty is solved as follows: first, contexts are introduced in which the two *kogda* clearly differ both semantically and syntactically; then the coordinating *kogda* is analyzed in these contexts based on the criteria. In the syntactic literature, temporal clauses are generally assumed to be more closely integrated with the host clause than, for example, causal or concessive clauses. Given the conclusion about the existence of coordinate temporal clauses, this point of view needs reconsideration.

THE PILOT CORPUS OF THE ENGLISH SEMANTIC SKETCHES

Petrova M.¹, **Ponomareva M.**^{2,1}, **Ivoylova A.**^{3,4,1}, ¹ABBY; ²HSE; ³RSUH; ⁴MIPT, Moscow, Russia

The paper is devoted to the creation of the semantic sketches for English verbs. The pilot corpus consists of the English-Russian sketch pairs and is aimed to show what kind of contrastive studies the sketches help to conduct. Special attention is paid to the cross-language differences between the sketches with similar semantics. Moreover, we discuss the process of building a semantic sketch, and analyse the mistakes that could give insight to the linguistic nature of sketches.

BETWEEN DENOISING AND TRANSLATION: EXPERIMENTS IN TEXT DETOXIFICATION

Pletenev S., National Research University Higher School of Economics, Moscow, Russia

This paper describes a solution for the RUSSE Detoxification competition held as part of the Dialogue 2022 conference. The paper presents experiments based on autoregressive and non-autoregressive models. The following approaches are described in this paper: 1) Detoxification as a special case of the text style-transfer problem and the use of modern approaches to solve this task in Russian. 2) Using the Automatic Post-Editing algorithm as a task of translation from toxic to normative Russian text. The article provides an analysis of the listed models, their results in detoxification of sentences, as well an analysis of errors and reasons why the models gave such a diverse result.

"BECAUSE NO ONE READS PROSE ANY MORE": GRAMMAR AND PROSODY OF INSUBORDINATED REASON CLAUSES

Podlesskaya V. I., Russian State University for the Humanities, Moscow, Russia

Based on data from the multimedia subcorpus of the Russian National Corpus, the paper addresses syntactic, pragmatic and prosodic features of insubordinated adverbial clauses introduced by the adverbial subordinator *potomu čto* 'because'. The quantitative analysis showed that more than 30% of reason clauses in spoken discourse appear to be insubordinated. Qualitatively, we observed symptoms of insubordination at various levels. (1) Prosodically, insubordinated clauses are placed after discourse fragments that are articulated with falling pitch projecting no continuation and are separated from them by the prosodic break. (2) Pragmatically, they can have independent illocutionary force and can form separate turns in dialogues. (3) Grammatically, they allow right dislocation of the adverbial subordinator — otherwise blocked in adverbial clauses.

ARTIFICIAL TEXT DETECTION IN RUSSIAN LANGUAGE: A BERT-BASED APPROACH

Posokhov P. A., **Skrylnikov S. S.**, **Makhnytkina O. V.**, ITMO University, Saint Petersburg, Russia

This paper describes our solution for the RuATD (Russian Artificial Text Detection) competition held within the Dialog 2022 conference. Our approach is based on the idea of transfer learning, using pre-trained RuRoBERTa, RuBERT, RuGPT3, RuGPT2 models. The final solution included Byte-level Byte-Pair Encoding tokenization, and a fine-tuned model RuRoBERTa model. The system got Accuracy metric value of 0.65 and took first place in the multi-class classification task.

SPOKEN CORPORA OF SPONTANEOUS SPEECH AS A SOURCE TO STUDY POLAR QUESTION INTONATION IN RUSSIAN DIALECTS

Post M., Department of Foreign Languages, University of Bergen, Norway

The emergence of several online spoken corpora of Russian regional speech opens new possibilities for the study of regional Russian intonation. The Russian dialect corpora of the Linguistic Convergence Laboratory [32; 1–10] were used to study the intonation of polar (yes/no) questions in regional rural speech. Although using spontaneous speech to study intonation is a challenge, the corpora are large enough to show general tendencies. The typical rising-falling pitch accent of most polar questions in Central Standard Russian is predominant in the regional corpora as well, but with possible variation in phonetic implementation and in the association of the fall. This accent is the most common even in the majority of question utterances with lowered questionhood, and dominates even in the regions known for rising accents in questions. The corpora show that tag questions are frequent in these interview data, unlike the question particles *li*, *ti* and *či*. Not only the dialectal particles *ti* and *či*, but also the Standard Russian question particle *li* shows a varying regional distribution.

MACHINE READING COMPREHENSION MODEL IN RUNNE COMPETITION

Rozhkov I. S., Loukachevitch N. V., Lomonosov Moscow State University, Moscow, Russia

The paper studies machine reading comprehension model (MRC) (Li et al., 2020) in its application to extracting nested named entities (nested NER) in the RuNNE-2022 evaluation (Artemova et al., 2022). The model transforms named entity recognition tasks to a question-answering task. In this paper we compare several approaches to formulating "questions" for the MRC model such as entity type names (keywords), entity type definitions, most frequent examples for the train set, combinations of definitions and examples. We found that using two most frequent examples from the training set is comparable in quality of nested NER with gathering qualitative definitions from different dictionaries, which is much more complicated. In the RuNNE evaluation, the MRC model obtained the best results among models without any manual work (rules or additional manual annotation of texts).

FINDINGS OF THE THE RUATD SHARED TASK 2022 ON ARTIFICIAL TEXT DETECTION IN RUSSIAN

Shamardina T.¹, Mikhailov V.², Chernianskii D.^{3,4}, Fenogenova A.², Saidov M.⁶, Valeeva A.⁷, Shavrina T.^{2,3}, Smurov I.^{1,7}, Tutubalina E.^{5,6,8}, Artemova E.⁶, ¹ABBY; ²SberDevices; ³AIRI; ⁴Skolkovo Institute of Science and Technology; ⁵Sber AI; ⁶HSE University; ⁷Moscow Institute of Physics and Technology, Moscow, Russia; ⁸Kazan Federal University, Kazan, Russia

We present the shared task on artificial text detection in Russian, which is organized as a part of the Dialogue Evaluation initiative, held in 2022. The shared task dataset includes texts from 14 text generators, i.e., one human writer and 13 text generative models fine-tuned for one or more of the following generation tasks: machine translation, paraphrase generation, text summarization, text simplification. We also consider back-translation and zero-shot generation approaches. The human-written texts are collected from publicly available resources across multiple domains.

The shared task consists of two sub-tasks: (i) to determine if a given text is automatically generated or written by a human; (ii) to identify the author of a given text. The first task is framed as a binary classification problem. The second task is a multi-class classification problem. We provide count-based and BERT-based baselines, along with the human evaluation on the first sub-task. A total of 30 and 8 systems have been submitted to the binary and multi-class sub-tasks, correspondingly. Most teams outperform the baselines by a wide margin. We publicly release our codebase, human evaluation results, and other materials in our GitHub repository.

TOPIC MODELING OF THE RUSSIAN SHORT STORIES OF 1900–1930S: THE MOST FREQUENT TOPICS AND THEIR DYNAMICS

Sherstinova T. Y., Moskvina A. D., Kirina M. A., Karysheva A. S., Kolphashchikova E. O., National Research University Higher School of Economics, Saint Petersburg, Russia

The article describes the results of an experiment on topic modeling of Russian short stories for three successive historical periods of the early 20th century: 1) the beginning of the 20th century until 1913, 2) the war-revolutionary period (1914–1922), and 3) the early Soviet period (1923–1930). Using the Latent Dirichlet Allocation (LDA) algorithm, 9 models were built — 3 samples of different sizes (100, 500, and 1000 stories) for each of the periods. It turned out that in every model there are very frequent "themes" (topics) that characterize with a high probability a fairly significant share of texts in each sample. Moreover, one can also observe a meaningful dynamics of these frequent topics over different time periods, which allows us to consider them as thematic and stylistic markers of the analyzed text collections along with the more traditional quantitative measures of text analysis. The variety of frequent topics turned out to be higher in the second and third periods, which can be explained by the greater lexical and stylistic diversity of the prose of the "era of change".

THE DELIMITATIVE IN RUSSIAN: LINEAR ORDERING OR A PATH TOWARDS THE CULMINATION?

Tatevosov S. G., Lomonosov Moscow State University, Interdisciplinary Scientific and Educational 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

In this paper, we explore the distribution of the delimitative in Russian in combination with eventuality descriptions that denote culminating processes. With H.R. Mehlig, we assume that homogenization of a process is a *sine qua non* condition on the delimitative. We argue for a modal treatment of homogeneity whereby a process is (locally) homogeneous at an interval if the cardinality of the set of metaphysically accessible non-culminating worlds does not change significantly at that interval.

DISTILLED MODEL FOR RUSSIAN NEWS CLUSTERING: MUCH LIGHTER AND FASTER, STILL ACCURATE

Trofimchuk D., National University of Science and Technology "MISIS", Moscow, Russia

This paper explores abilities of knowledge distillation for the purposes of News clustering which also can be generalized as an event detection task. We used a BERT-based clustering model as a teacher and tested various student networks based on different architectures (RNN, FFN, convolutional and Transformer-based networks) in order to get a faster lightweight analogue that is more likely to be deployed in real products. We tried two distillation strategies: the first one combined an original loss function from the initial model with a distillation objective, for the second one we used only a specific distillation loss. This approach turned out to be more successful. It let us extend training and validation datasets and gave significantly better results. One of our distilled models scored about 1% lower than the teacher network, but is more than 20 times smaller and 5 times faster by inference.

IS NEURAL LANGUAGE ACQUISITION SIMILAR TO NATURAL? A CHRONOLOGICAL PROBING STUDY

Voloshina E.^{1,2}, **Serikov O.**^{1,2,3}, **Shavrina T.**^{1,4}, ¹AIRI; ²HSE University; ³DeepPavlov; ⁴SberDevices, Moscow, Russia

The probing methodology allows one to obtain a partial representation of linguistic phenomena stored in the inner layers of the neural network, using external classifiers and statistical analysis.

Pre-trained transformer-based language models are widely used both for natural language understanding (NLU) and natural language generation (NLG) tasks making them most commonly used for downstream applications. However, little analysis was carried out, whether the models were pre-trained enough or contained knowledge correlated with linguistic theory.

We are presenting the chronological probing study of transformer English models such as MultiBERT and T5. We sequentially compare the information about the language learned by the models in the process of training on corpora. The results show that 1) linguistic information is acquired in the early stages of training 2) both language models demonstrate capabilities to capture various features from various levels of language, including morphology, syntax, and even discourse, while they also can inconsistently fail on tasks that are perceived as easy.

We also introduce the open-source framework for chronological probing research, compatible with other transformer-based models. https://github.com/EkaterinaVoloshina/chronological_probing

COLLOCATION2TEXT: CONTROLLABLE TEXT GENERATION FROM GUIDE PHRASES IN RUSSIAN

Vychegzhanin S. V., **Kotelnikov E. V.**, Vyatka State University, Kirov, Russia

Large pre-trained language models are capable of generating varied and fluent texts. Starting from the prompt, these models generate a narrative that can develop unpredictably. The existing methods of controllable text generation, which guide the narrative in the text in the user-specified direction, require creating a training corpus and an additional time-consuming training procedure. The paper proposes and investigates *Collocation2Text*, a plug-and-play method for automatic controllable text generation in Russian, which does not require fine-tuning. The method is based on two interacting models: the autoregressive language ruGPT-3 model and the autoencoding language ruRoBERTa model. The idea of the method is to shift the output distribution of the autoregressive model according to the output distribution of the autoencoding model in order to ensure a coherent transition of the narrative in the text towards the guide phrase, which can contain single words or collocations. The autoencoding model, which is able to take into account the left and right contexts of the token, “tells” the autoregressive model which tokens are the most and least logical at the current generation step, increasing or decreasing the probabilities of the corresponding tokens. The experiments on generating news articles using the proposed method showed its effectiveness for automatically generated fluent texts which contain coherent transitions between user-specified phrases.

KEYWORDS IN RETRIEVAL OF PROSODIC DATA

Yanko T. E., Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

Data retrieval can be based on the keywords found in the orthographic records of sound speech. This paper is aimed at analyzing the communicative meanings, compositions of meanings expressed by prosody, and the structure of segmental material aligned with prosodic variations. The meanings and their compositions expressed by prosody can have segmental counterparts, e.g., lexemes, or some other linguistic units. These counterparts, in turn, are used as keywords for retrieval. The results of searching are used as a material for analyzing the syntactic structure of the communicative components of sentences, such as themes and rhemes of statements, and the components of some specific illocutions (dreams, recollections, explanations), and combinations of illocutions with discourse incompleteness. The data source for this analysis is the Russian National Corpus (Multimodal sub-corpus Murko), Prosodically Annotated Corpus of Spoken Russian (SpokenCorpora.ru), and video-hosting Youtube. As the instrument, the computer system for sound speech analyzing Praat is used. The paper is illustrated throughout with pitch contours of sound records.

HISTORICAL TEXT CORPORA AND THE CONCLUSIVENESS OF LINGUISTIC ANALYSIS

Zimmerling A. V., Pushkin State Russian Language Institute, Moscow, Russia; Institute of Linguistics, Russian Academy of Science, Moscow, Russia

I discuss the methodology and conclusiveness of the corpus-based historical linguistics and analyze two formal models predicting the language-internal variation in Early Old Russian syntax. Linguistic models claiming a rigid distribution of grammatical features like \pm overt realization of agreement markers activate hidden corpus characteristics such as profiles of text genres, chronology, vector of change, \pm impact of L2, \pm presence of supra-dialect features. In this case they can be valued and checked on text samples, where genre features are stable, while location and time vary.

HUMAN COMMUNICATIVE RESPONSES TO DIFFERENT MODES OF GAZE MANAGEMENT BY THE ROBOT

Zinina A. A.^{1,2,3}, **Kotov A. A.**^{1,2,3}, **Zaidelman L. Y.**^{1,2}, **Arinkin N. A.**^{1,2}, ¹Kurchatov Institute; ²Russian State University for the Humanities; ³Moscow State Linguistic University, Moscow, Russia

We investigated communicative reactions of people (N = 46), while telling stories to two companion robots, who reacted differently to the human gaze (head turning). In response to a human gaze the “aversive” robot averted its gaze away from the user, while the “responsive” robot, lifted its head and showed a responsive gaze. We found that users with high level of emotional intelligence prefer the gaze responsive robot and better recognize the difference between the robots. Thus, these users constitute the core group for the technology. In this paper, we further examine behavioral patterns of people in the experiment situation: (a) shift of attention to the story; (b) shift of attention to the robot; (c) joint attention. We also distinguish the communicative reactions of people, mainly from the core group, to the aversive and responsive gazes of the robots: positive responses to gaze contact and negative responses to gaze aversion. We show that for some users the responsive gaze behavior of the robot may serve as positive feedback, increasing the number of human iconic gestures, while telling a story to the responsive robot, and decreasing the number of iconic gestures in a story to the aversive robot.

Авторский указатель

Абросимов К. И.	1	Ивойлова А.	436	Панченко А.	114, 333
Алибаева К.	8	Ивойлова А.М.	141	Пекелис О. Е.	420
Апресян В.	18	Ильин А.	323	Петрова М.	436
Артемова Е.	497	Инькова О. Ю.	273	Плетенев С.	447
Артемова Е. Л.	33	Калинин А. А.	311	Подлеская В. И.	456
Баранов А. Н.	42	Карышева А. С.	512	Пономарева М.	436
Батура Т. В.	33	Кирина М. А.	512	Попкова Н. А.	273
Берзин В.	323	Киселева К. Л.	527	Посохов П. А.	470
Богуславский И. М.	50	Князев С. В.	284	Пост М.	477
Большакова Е. И.	61	Кобозева И. М.	236	Рожков И. С.	33, 488
Бондаренко И.	70	Колесникова А.	295	Саидов М.	497
Боченина К. О.	211	Колмогорова А. В.	311	Семенов Н.	114
Буденная Е.	159	Колпащикова Е. О.	512	Сериков О.	550
Бурцев М.	295	Коновалов В.	295	Скрыльников С. С.	470
Буянов И.	81	Корзун В.	323	Смирнов И.	93
Валеева А.	497	Котельников Е. В.	211, 223, 333, 564	Смууров И.	497
Вен С. Й.	375	Кротова Е. Б.	349	Сон К.	375
Вилинбахова Е. Л.	50	Кузнецов Д. П.	185	Соченков И.	81
Волошина Е.	550	Куратов Ю.	295	Татевосов С. Г.	527
Вычегжанин С. В.	564	Кустова Г. И.	358	Телегина А. Д.	61
Гадецкий Д.	323	Левонтина И. Б.	367	Трофимчук Д.	538
Головизнина В. С.	223	Ли Б.	375	Тутубалина Е. В.	33, 497
Гончаров А. А.	236	Лобанов Б. М.	382	Федорова О. В.	203
Горбова Е. В.	247	Логачева В.	114	Феногенова А.	114, 497
Гусев И.	264	Лукашевич Н. В.	8, 33, 333, 488	Фищева И. Н.	211
Дементьева Д.	114	Махныткина О. В.	470	Цветаева Е. Н.	349
Добровольский В.А.	141	Михайлов В.	497	Циммерлинг А. В.	586
Добровольский Д. О.	132, 349	Мичурина М.А.	141	Чернявский Д.	497
Дурагин П. В.	150	Мовсесян А. А.	402	Чистова Е.	93
Дэн Х. Дж.	375	Москвина А. Д.	512	Чуйкова О. Ю.	247
Евдокимова А.	159	Мосягина А. Г.	1	Шаврина Т.	114, 497, 550
Евсеев Д. А.	176, 185	Наговицин М. С.	185	Шамардина Т.	497
Евстигнеева М. Ю.	284	Никишина И.	114, 333	Шарандин А. В.	349
Житко В. А.	382	Николаева Ю.	159	Шерстинова Т. Ю.	512
Зализняк Анна А.	132	Нуриев В. А.	273	Шмелев А.	18
Змеев М. В.	33	Орженовский М. В.	412	Янко Т. Е.	577
Иванов В. В.	33	Осадчий Д.	211		

Author Index

Abrosimov K. I.	1	Ivoylova A. M.	141, 436	Petrova M.	436
Alibaeva K.	8	Kalinin A. A.	311	Pletenev S.	447
Apresjan V.	18	Karysheva A. S.	512	Podlesskaya V. I.	456
Arinkin N. A.	594	Kirina M. A.	512	Ponomareva M.	436
Artemova E. L.	33, 497	Kisseleva X. L.	527	Popkova N. A.	273
Baranov A. N.	42	Knyazev S. V.	284	Posokhov P. A.	470
Batura T. V.	33	Kobozeva I. M.	236	Post M.	477
Berzin V.	323	Kolesnikova A.	295	Rozhkov I. S.	33, 488
Bochenina K. O.	211	Kolmogorova A. V.	311	Saidov M.	497
Boguslavsky I. M.	50	Kolpashchikova E. O.	512	Semenov N.	114
Bolshakova E. I.	61	Konovalov V.	295	Serikov O.	550
Bondarenko I.	70	Korzun V.	323	Shamardina T.	497
Budennaya E.	159	Kotelnikov E. V. ...	211, 223, 333, 564	Sharandin A. V.	349
Burtsev M.	295	Kotov A. A.	594	Shavrina T.	114, 497, 550
Buyanov I.	81	Krotova E. B.	349	Sherstinova T. Y.	512
Chernianskii D.	497	Krotova I.	114	Shmelev A.	18
Chistova E.	93	Kuratov Y.	295	Skrylnikov S. S.	470
Chuiikova O. Iu.	247	Kustova G. I.	358	Smirnov I.	93
Chuprina A. O.	106	Kuznetsov D. P.	185	Smurov I.	497
Dale D.	114	Levontina I. B.	367	Sochenkov I.	81
Dementieva D.	114	Li B.	375	Song Q.	375
Deng H.	375	Lobanov B. M.	382	Tatevosov S. G.	527
Dobrovolskii V. A.	141	Logacheva V.	114	Telegina A. D.	61
Dobrovol'skij D. O.	132, 349	Loukachevitch N. V. ...	8, 33, 333, 488	Trofimchuk D.	538
Duryagin P. V.	150	Makhnytkina O. V.	470	Tsvetaeva E. N.	349
Evdokimova A.	159	Maloyan N.	396	Tutubalina E. V.	33, 497
Evseev D. A.	176, 185	Michurina M. A.	141	Valeeva A.	497
Evstigneeva M. Y.	284	Mikhailov V.	497	Vilinbakhova E. L.	50
Fedorova O. V.	203	Moskvina A. D.	512	Voloshina E.	550
Fenogenova A.	114, 497	Mosyagina A. G.	1	Vychegzhanin S. V.	564
Fishcheva I. N.	211	Movsesyan A. A.	402	Weng Y.	375
Gadecky D.	323	Nagovitsin M. S.	185	Yanko T. E.	577
Goloviznina V. S.	223	Nikishina I.	114, 333	Zaidelman L. Y.	594
Goncharov A. A.	236	Nikolaeva Y.	159	Zalizniak Anna A.	132
Gorbova E. V.	247	Nuriev V. A.	273	Zhitko V. A.	382
Gusev I.	264	Nutfullin B.	396	Zimmerling A. V.	586
Ilin A.	323	Orzhenovskii M. V.	412	Zinina A. A.	594
Ilyshin E.	396	Osadchiy D.	211	Zmeev M. V.	33
Inkova O. Y.	273	Panchenko A.	114, 333		
Ivanov V. V.	33	Pekelis O. E.	420		

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 21, 2022

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**

Издательский центр
Российского государственного
гуманитарного университета
125047, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06