

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной
конференции «Диалог» (2020)

Выпуск 19

Computational Linguistics and Intellectual Technologies

Papers from the Annual International
Conference “Dialogue” (2020)

Issue 19

УДК 80/81; 004
ББК 81.1
К63

Редакционная
коллегия:

*В. П. Селегей (главный редактор),
В. И. Беликов, И. М. Богуславский, Б. В. Добров,
Д. О. Добровольский, Л. М. Захаров, Л. Л. Иомдин,
И. М. Кобозева, Е. Б. Козеренко, М. А. Кронгауз,
Н. И. Лауфер, Н. В. Лукашевич, Д. Маккарти,
П. Наков, Й. Нивре, Г. С. Осипов, А. Ч. Пиперски,
В. Раскин, Э. Хови, С. А. Шаров, Т. Е. Янко*

К63 Компьютерная лингвистика и интеллектуальные технологии:
По материалам ежегодной международной конференции «Диалог» (Москва,
17–20 июня 2020 г.). Вып. 19 (26), основной том. Москва : РГГУ, 2020.

ISBN 978-5-7281-2947-9

Сборник включает 61 доклад международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2020», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

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

ISBN 978-5-7281-2947-9 (осн. том)
ISBN 978-5-7281-2946-2

© Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии» (составитель), 2020

Предисловие

19-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 26-й международной конференции «Диалог». На основании мнений нашего рецензентского корпуса для публикации в ежегоднике редколлекцией были отобраны 61 доклад из примерно ста работ, которые были приняты к представлению на конференции в 2020 году. Работы 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, суммаризация, генерация, анализ тональности и т. д.)
- Глубокое обучение в NLP (новые методы, содержательная интерпретация)
- Язык и компьютерное зрение (новое направление в Диалоге)
- Компьютерный анализ Social Media
- Корпусная лингвистика и корпусометрия (методики создания, использования и оценки корпусов)
- Лингвистический анализ текста (морфология, синтаксис, семантика)
- Представление и извлечение знаний
- Мультимодальная коммуникация
- Модели общения и диалоговые агенты

В соответствии с традициями «Диалога», старейшей и крупнейшей конференции по компьютерной лингвистике в России, отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом. Диалог является де-факто крупнейшим форумом по проблемам создания современных компьютерных ресурсов, моделей и технологий для русского языка.

Поэтому ключевым событием «Диалога» — по уже почти десятилетней традиции является подведение итогов технологических соревнований между разработчиками систем лингвистического анализа русскоязычных текстов, Dialogue Evaluation. В этом году состоялись три соревнования:

- GramEval 2020: соревнование по автоматическому морфосинтаксическому анализу. Такое соревнование проводилось на Диалоге 8 лет назад. И теперь можно увидеть, насколько далеко шагнули методы парсинга за это время.
- RuREBus-2020: соревнование по извлечению отношений из текстов в постановке, максимально приближенной к требованиям бизнеса.
- RUSSE'2020: соревнование по автоматическому позиционированию новой лексики в имеющемся тезаурусе.

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:

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

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что с 2018 года Редаксовет отказался от печати сборника на бумаге, поскольку бумажный вариант пользуется все меньшей популярностью. Сборник, как и в прошлые годы, размещается на сайте конференции и индексируется Scopus.

Программный комитет конференции «Диалог»

*Редакколлегия сборника «Компьютерная лингвистика
и интеллектуальные технологии»*

Организаторы

Ежегодная конференция «Диалог» проводится при организационной поддержке компании АВВУУ.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания АВВУУ
- 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 Жозефа Фурье — Гренобль 1, Франция
Гельбух Александр Феликсович	Национальный политехнический институт, Мексика
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
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	Уппсальский университет, Швеция
Ниренбург Сергей	Университет Мэриленда, Балтимор, США
Осипов Геннадий Семёнович	Институт системного анализа РАН, Россия
Райгородский Андрей Михайлович	МФТИ, Школа прикладной математики и информатики, Россия
Раскин Виктор	Университет Пердью, США
Селегей Владимир Павлович	Компания АВВУУ, МФТИ, Россия
Хови Эдуард	Университет Карнеги — Меллон, США
Шаров Сергей Александрович	Университет Лидса, Великобритания

Организационный комитет

Селегей Владимир Павлович,
председатель

Беликов Владимир Иванович

Браславский Павел Исаакович

Добров Борис Викторович

Захаров Леонид Михайлович

Иомдин Леонид Лейбович

Кобозева Ирина Михайловна

Козеренко Елена Борисовна

Лауфер Наталия Исаевна

Ляшевская Ольга Николаевна

Пиперски Александр Чедович

Толдова Светлана Юрьевна

Федорова Ольга Викторовна

Шаров Сергей Александрович

Компания АBBYУ

Институт русского языка
им. В. В. Виноградова РАН

Уральский федеральный университет

НИВЦ МГУ им. М. В. Ломоносова

МГУ им. М. В. Ломоносова

Институт проблем передачи информации
РАН им. А. А. Харкевича

МГУ им. М. В. Ломоносова

Институт проблем информатики РАН

Компания Yandex

Институт русского языка
им. В. В. Виноградова РАН

РГГУ

НИУ «Высшая школа экономики»

МГУ им. М. В. Ломоносова

Университет Лидса

Секретариат

Родионова Ольга Игоревна,
координатор оргкомитета

Ульянова Анна Вячеславовна,
секретарь оргкомитета

Компания АBBYУ

РГГУ

Рецензенты

Августинова Тая
Антонова Александра Александровна
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Апресян Валентина Юрьевна
Артемова (Черняк) Екатерина Леонидовна
Архангельский Тимофей Александрович
Байтин Алексей Владимирович
Бенко Владимир
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бочаров Виктор Владиславович
Браславский Павел Исаакович
Бурцев Михаил
Васильев Виталий Геннадьевич
Галинская Ирина Евгеньевна
Гершман Анатолий
Гращенков Павел Валерьевич
Диконов Вячеслав
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрушина Нина Роландовна
Жуковский Александр
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Иванов Владимир Владимирович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Ильвовский Дмитрий
Катинская Анисья Юрьевна
Кибрик Андрей Александрович
Клышинский Эдуард
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Кронгауз Максим Анисимович
Кротова Ирина
Кутузов Андрей Борисович
Лапошина Антонина Николаевна
Левонтина Ирина Борисовна
Леонтьев Алексей Петрович
Лобанов Борис Мефодьевич
Логачева Варвара
Лютикова Екатерина Анатольевна
Маккарти Диана
Малафеев Алексей
Марков Александр Юрьевич
Мисюрев Алексей Владимирович
Митрофанова Ольга
Наков Преслав
Недолужко Анна Юрьевна
Никишина Ирина
Новицкий Валерий Игоревич
Пазельская Анна Германовна
Паперно Денис Аронович
Панченко Александр Иванович
Переверзева Светлана Игоревна
Петрова Мария
Пивоварова Лидия
Пиперски Александр Чедович
Подлеская Вера Исааковна
Пономарева Мария
Рыгаев Иван Петрович
Смирнов Иван Валентинович
Смууров Иван Михайлович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Сорокин Алексей Андреевич
Татевосов Сергей
Тестелец Яков Георгиевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Тутубалина Елена
Урысон Елена Владимировна
Усталов Дмитрий Алексеевич
Федорова Ольга Викторовна
Феногенова Алена
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шаров Сергей Александрович
Шелманов Артём Олегович
Юдина Мария Владимировна
Ямщиков Иван

Contents*

Anastasyev D. G. Exploring Pretrained Models for Joint Morpho-Syntactic Parsing of Russian	1
Arefyev N. V., Fedoseev M. V., Kabanov A. V., Zizov V. S. Word2vec not dead: predicting hypernyms of co-hyponyms is better than reading definitions	13
Badryzlova Yu. G. Exploring Semantic Concreteness and Abstractness for Metaphor Identification and Beyond	33
Баранов А. Н., Добровольский Д. О. Динамика стиля русской письменной речи XIX века: корпусный эксперимент	48
Беликов В., Селегей В., Селегей Д. Интернет-корпус как инструмент лингвистических исследований: дифференциальность, авторизация, тематические смещения (или корпуса, которым так хочется верить) ..	62
Blinova O. V., Tarasov N. A., Modina V. V., Blekanov I. S. Modeling lemma frequency bands for lexical complexity assessment of Russian texts	76
Bocharov V. V., de Chalendar G. The Russian Language Pipeline in the LIMA multilingual analyzer	93
Boguslavsky I. M., Dikonov V. G., Frolova T. I., Iomdin L. L., Lazursky A. V., Rygaev I. P., Timoshenko S. P. Full-fledged semantic analysis as a tool for resolving Triangle-COPA social scenarios	106
Bolshina A. S., Loukachevitch N. V. Generating training data for word sense disambiguation in Russian	119
Budennaya E. V., Evdokimova A. A., Nikolaeva Ju. V., Sukhova N. V. Referential phenomena in speakers kinetic channels	133
Чернова Д. А., Алексеева С. В., Слюсарь Н. А. Чему нас учат ошибки: трудности при обработке слов с частотными орфографическими ошибками	147
Чуйкова О. Ю. Об особенностях вторичной имперфективации глаголов с префиксом по- в русском языке	160

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

Dale D. S.	
A simple solution for the Taxonomy enrichment task: Discovering hypernyms using nearest neighbor search	177
Davletov A. A., Gordeev D. I., Rey A. I., Arefyev N. V.	
RENERSANs: Relation Extraction and Named Entity Recognition as Sequence Annotation	187
Derbanosov R., Bakhanova M.	
Stability of Topic Modeling via Modality Regularization	198
Деткова Ю., Новицкий В., Петрова М., Селегей В.	
Дифференциальные семантические скетчи для русскоязычных интернет-корпусов	211
Dyachkov V. V., Khomchenkova I. A., Pleshak P. S., Stoynova N. M.	
Annotating and exploring code-switching in four corpora of minority languages of Russia	228
Эндресен А. А., Жукова В. А., Мордашова Д. Д., Рахилина Е. В., Ляшевская О. Н.	
Русский Конструктик: новый лингвистический ресурс, его устройство и специфика	241
Eremeev M. A., Vorontsov K. V.	
Quantile-based approach to estimating cognitive text complexity	256
Evseev D. A., Arkhipov M. Yu.	
SPARQL query generation for complex question answering with BERT and BiLSTM-based model	270
Feldman D. G., Sadekova T. R., Vorontsov K. V.	
Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining	283
Fenogenova A. S., Tikhonova M. I., Filipetskaya D. V., Mironenko F. D., Tabisheva A. O.	
Event2mind for Russian: understanding emotions and intents in texts. Corpus and model for evaluation	299
Гончаров А. А., Инькова О. Ю.	
Имплицитные логико-семантические отношения и метод их поиска в параллельных текстах	310
Горбова Е. В.	
Видовые тройки русского глагола в диахронии (на материале НКРЯ) ...	321
Gordeev D. I., Davletov A. A., Rey A. I., Akzhigitova G. R., Geymbukh G. A.	
Relation extraction dataset for the Russian	348
Ilvovsky D. A., Galitsky B. A.	
Dialogue management using extended discourse trees	361

Инькова О. Ю.

Количественный метод анализа коннекторов: «портрет» русского союза или в надкорпусной базе данных коннекторов 372

Inshakova E. S., Sizov V. G.

An experimental rule-based parser for Russian employing the NLP resources of the ETAP system 387

Иомдин Б. Л., Иомдин Л. Л.

Валентная структура некоторых речевых предикатных слов: новые находки 400

Ivanin V. A., Artemova E. L., Batura T. V., Ivanov V. V., Sarkisyan V. V., Tutubalina E. V., Smurov I. M.

RuREBus-2020 Shared Task: Russian Relation Extraction for Business 416

Kononenko I. S., Sidorova E. A., Akhmadeeva I. R.

Comparative analysis of rhetorical and argumentative structures in the study of popular science discourse 432

Konovalov V. P., Gulyaev P. A., Sorokin A. A., Kuratov Y. M., Burtsev M. S.

Exploring the BERT Cross-Lingual Transfer for Reading Comprehension 445

Korotaev N. A., Podlesskaya V. I., Smirnova K. V., Fedorova O. V.

Disfluencies in Russian spoken monologues: a distributional analysis 454

Korzun V. A.

R-BERT for relationship extraction on Russian business documents 467

Kunilovskaya M., Kutuzov A., Plum A.

Taxonomy Enrichment for Russian: Synset Classification Outperforms Linear Нуponym-Нypernym Projections 474

Кустова Г. И.

Семантические эффекты времени во вводных конструкциях с ментальными глаголами 485

Kutuzov A., Fomin V., Mikhailov V., Rodina J.

ShiftRy: Web Service for Diachronic Analysis of Russian News 500

Kuvshinova T.

Sentence compression for Russian: dataset and baselines 517

Левонтина И. Б.

«Understatement» и сарказм: лексикализация риторического приема ... 529

Loukachevitch N. V., Rusnachenko N. L.

Sentiment Frames for Attitude Extraction in Russian 541

Lyashevskaya O. N., Shavrina T. O., Trofimov I. V., Vlasova N. A.

GramEval 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing 553

Malykh V., Cherniavskii D., Valukov A. Summary Construction Strategies for Headline Generation in the Russian ...	570
Nikishina I., Logacheva V., Panchenko A., Loukachevitch N. RUSSE2020: Findings of the First Taxonomy Enrichment Task for the Russian Language	579
Оленикова А. В., Федорова О. В. Совместный синтаксис в диалогах с заикающимися	596
Pimonova E., Durandin O., Malafeev A. Doc2Vec or better interpretability? A method study for authorship attribution	606
Пиперски А. Ч. Русский язык и корпусное разнообразие	615
Подлеская В. И. «А тот Перовской не дал власть поспать»: просодия и грамматика анафорического <i>тот</i> в зеркале корпусных данных	628
Shaheen Z., Wohlgenannt G., Zaity B., Mouromtsev D. I., Pak V. G. Russian Natural Language Generation: Creation of a Language modeling Dataset and Evaluation with Modern Neural Architectures	644
Шмелев А. Д. Лингвоспецифичные слова в зеркале перевода: <i>тоска</i>	
Sorokin A. A., Smurov I. M., Kirianov D. P. Tagging and parsing of multidomain collections	670
Stenger I., Avgustinova T. Visual vs. auditory perception of Bulgarian stimuli by Russian native speakers	684
Tarasov D., Matveeva T., Galiullina N. An Empirical Investigation of Language Model based Reverse Turing Test as a Tool for Knowledge and Skills Assessment	696
Татевосов С. Г., Киселева К. Л. Семантика <i>обратно</i>: возвращение в прерванное состояние	708
Tikhomirov M. M., Loukachevitch N. V., Sirotnina A. Yu., Dobrov B. V. Pretraining and Augmentation in Named Entity Recognition Task for Cybersecurity Domain in Russian	724
Tikhomirov M. M., Loukachevitch N. V., Parkhomenko E. A. Combined Approach to Hypernym Detection for Thesaurus Enrichment	736
Toldova S., Davydova T., Kobozeva M., Pisarevskaya D. Discourse Features of Blogs in Subcorpus of Russian Ru-RSTreebank	747

Yadrintsev V. V., Ryzhova A. A., Sochenkov I. V.

**Distributional models and auxiliary methods for determining the
hypernyms of words in Russian** 762

Янко Т. Е.

Наречие *давно* по данным звучащего корпуса 773

Зализняк Анна А.

Русское *как бы*: семантика, прагматика, диахрония 784

Zimmerling A. V.

Zero forms in morphological paradigms: the verb “be” in Russian 795

Zinina A. A., Zaidelman L. Y., Kotov A. A., Arinkin N. A.

**The perception of robots emotional gestures and speech by children
solving a spatial puzzle** 811

Abstracts 827

Авторский указатель 848

Author Index 850

EXPLORING PRETRAINED MODELS FOR JOINT MORPHO-SYNTACTIC PARSING OF RUSSIAN

Anastasyev D. G. (dan-anastasev@yandex-team.ru)

Yandex, Moscow, Russia

In this paper, we build a joint morpho-syntactic parser for Russian. We describe a method to train a joint model which is significantly faster and as accurate as a traditional pipeline of models. We explore various ways to encode the word-level information and how they can affect the parser's performance. To this end, we utilize learned from scratch character-level word embeddings and grammeme embeddings that have shown state-of-the-art results for similar tasks for Russian in the past. We compare them with the pretrained contextualized word embeddings, such as ELMo and BERT, known to lead to the breakthrough in miscellaneous tasks in English. As a result, we prove that their usage can significantly improve parsing quality.

Key words: morphological analysis, dependency parsing, pretrained language models, multitask learning

DOI: 10.28995/2075-7182-2020-19-1-12

АНАЛИЗ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ ДЛЯ МОРФО-СИНТАКСИЧЕСКОГО ПАРСИНГА РУССКОГО ЯЗЫКА

Анастасьев Д. Г. (dan-anastasev@yandex-team.ru)

Яндекс, Москва, Россия

1. Introduction

Morpho-syntactic parsing is an important part of various natural language processing systems. In our work, we consider the following components of a parser: morphological parser, lemmatizer and syntactic parser.

The whole parsing is typically performed in a pipeline manner when each of the mentioned components works separately and relies on another component's predictions. It may lead to error accumulation and requires multiple trained models to be run sequentially. We designed a joint model that learns to perform the whole

morpho-syntactic parsing task in a multi-task mode. This way, all tasks are solved using a single model and there are no dependencies between the components.

In the past few years, pretrained language models have become almost a standard way to improve the model’s quality. However, the research of these models was mostly conducted for English. To fill the gap, we compared two of the most popular language models—ELMo [Peters et al., 2018] and BERT [Devlin et al., 2018]—on our task in application to Russian. Despite the reasonable improvement in terms of accuracy, such models are significantly larger and slower than uncontextualized word embedding models. For a fair comparison, we also implemented a model based on character-level word embeddings which have shown the best results for morphological parsing of Russian on the MorphoRuEval-2017 dataset [Sorokin et al., 2017].

We tested our models on the GramEval-2020 shared task. The proposed BERT-based joint model achieved the state-of-the-art results on it.

2. Background

In this section, we give an overview of the parser’s components we considered and describe the choices we made designing our parser.

2.1. POS and Morphological Features Tagger

In the Universal Dependencies notation, a morphological parser’s prediction consists of a part-of-speech (POS) prediction and a morphological features prediction. Some systems make the prediction separately, some condition the morphological features prediction on the POS prediction [Qi et al., 2019], others predict them simultaneously [Kanerva et al., 2018].

We have chosen the latter approach because it’s known to lead to the state-of-the-art results on Russian [Anastasyev et al., 2018]. Thus, the POS tag and morphological features were simply concatenated (e.g., “NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Plur”) and model had to perform multiclass classification with a significant number of classes (about 1200 in our case). It means that the model cannot predict inconsistent grammatical values. A drawback of such an approach is that the model learns to ignore low frequent morphological features at the tail of the distribution. However, we assumed that errors from this tail cannot possibly affect the model’s accuracy much.

2.2. Lemmatizer

A straightforward approach is a lemmatization based on a dictionary. Frequent non-homonymous words may be lemmatized using a dictionary only: you just need to look up the word in the dictionary and return the corresponding lemma. Frequent, but homonymous words should be lemmatized with the help of a morphological parser which is required to disambiguate the word. The disambiguated word again may be found in a dictionary. The

dictionary might be built from the train data only, yet better results can be achieved with a dictionary based on external resources (such as [Korobov, 2015]). This method fails when an out-of-vocabulary word appears. However, you still can return the word itself as the lemma (this assumption is reasonably good in case of a large enough dictionary).

Another approach utilizes sequence-to-sequence models [Kanerva et al., 2018] to translate the sequence of symbols of the word to the sequence of symbols of its lemma. This way, the problem of the out-of-vocabulary words is solved, but the inference is slower, the model is prone to hallucinations and output of the morphological parser is still required to disambiguate the word.

In our method, we used a method similar to the one in UDPipe [Straka et al., 2016] and in [Sharoff et al., 2011]. It is based on modification rules that have to be applied to a word to obtain its lemma. Only three spelling modification types might be applied: cut N symbols from the prefix of the word, cut N symbols from its suffix and append a specific sequence of symbols to its suffix. Also, three case modifications could be used: a word might be lowercased, uppercased and capitalized. A result modification rule is a combination of such modifications where each modification type appears exactly once. For example: {"cut_prefix": 0, cut_suffix": 1, "append_suffix": "ый", "lower": false, "capitalize": false, "upper": false}.

Therefore, the lemmatizer is a simple classifier that learns to predict a rule to be applied from the rules found in the training set. We considered only rules that appeared more than twice in the train data, which led to less than 1000 classes.

The designed lemmatizer can be trained in the multitask mode with the rest of the parser's components. In this case, the lemmatizer doesn't rely on the morphological parser explicitly anymore, but they learn shared representations. This way, the error propagation issue is solved, and a single model is trained. Also, the model is less likely to hallucinate an invalid lemma than in the sequence-to-sequence approach.

2.3. Dependency Parser

In our work, we considered biaffine dependency parser [Dozat et al., 2017], which applies biaffine classifiers for pairs of words to predict whether a directed arc between these two words should exist and which label this arc should have. We used Edmonds' algorithm for finding minimum spanning trees on directed graphs for decoding.

Typically, this model relies on morphological parser's predictions. We compared such an approach with both unconditional dependency parser (that doesn't receive morphological predictions as an input) and joint model parser (which is trained to predict morphology and syntax relations simultaneously). The latter model uses shared representations between the morphological parser, the lemmatizer and the dependency parser. This approach should help them to learn an additional useful signal from the data and to regularize them.

2.4. Word Embeddings

One of the most important parts of our study is the choice of the word embeddings. We considered the following options.

2.4.1. Character-level Word Embeddings

Character-level word embeddings have led to state-of-the-art models in the past. They were shown to be useful for morphological parsing [Ling et al., 2015] as well as dependency parsing [Qi, 2019]. They are also useful because any word can be encoded and the problem of out-of-vocabulary words doesn't exist. Moreover, the model can learn morphology and semantics from the word's spelling which is likely to be meaningful in Russian.

We used the most common approach to train the character-level embeddings based on BiLSTM. This way, BiLSTM is applied to the sequence of symbol embeddings and word's embedding is constructed as a concatenation of the last states of the forward and backward LSTMs.

2.4.2. Contextualized Word Embeddings

Recently, contextualized word embeddings have led to a breakthrough in many NLP tasks. Contextualized word embeddings are pretrained using a language model task on a large amount of data. The pretraining step helps to learn useful relations from unlabeled data. The contextualization of these embeddings usually leads to better representations of words: it disambiguates the word and encodes some important information about the word's context. We considered the following models: ELMo [Peters et al., 2018] and BERT [Devlin et al., 2018].

The first model uses two LSTMs as forward and backward language models. Those language models encode the context information from the previous and the next words. Yet, they are independent, which leads to a certain lack of context information, because each of them has access only to some part of the sentence. ELMo also utilizes character-level word embeddings to build an uncontextualized word representation. However, unlike the model mentioned in Section 2.4.1, their character-level word embeddings are based on the convolutional neural network instead of a BiLSTM.

BERT is based on a transformer model, and it is trained in a masked language model manner. It should help to produce more contextualized word representation because BERT can access both the previous and the next words. To deal with the out-of-vocabulary words, BERT is designed to work on subword-level, which means that a word can be represented using any number of subword embeddings. To obtain a word's embedding, we used mean-pooling of its subword embeddings.

The main drawback of such embeddings is their speed: they tend to be by a few orders of magnitude slower than a simple lookup in a word embeddings table. They are also much larger than character-level word embeddings from the previous section (e.g., in our case, a character-level model was 10 times smaller than an ELMo-based one and almost 30 times smaller than a BERT-based model).

2.4.3. Grammeme Embeddings

It was shown that additional grammeme embeddings can give a significant improvement of the model quality for morphological parsing [Anastasyev et al., 2018]. They encode ambiguous information about the word's grammatical value and give the model an additional signal about the relations between this word and others. We experimented with grammeme embeddings built using the pymorphy2 dictionary [Korobov, 2015]: we concatenated the word embeddings mentioned in the previous two sections with the grammeme embeddings.

3. Experimental Setup

As already mentioned in [Section 2](#), we explored the following embedders in our experiments: character-level BiLSTM, ELMo¹ and BERT² with possibly concatenated grammeme embedding.

For the character-level BiLSTM we used a 32-dim character embedding with a dropout rate equal to 0.3 and a single-layer 128-dim BiLSTM (which means that the result word embedding has 256 dimensions).

We considered two types of the encoder: a pass-through encoder which simply passes the word embeddings to the decoder and a single-layer 256-dim BiLSTM encoder. The latter encoder helps to contextualize the word embeddings, while the pass-through one utilizes built-in contextualization of ELMo and BERT models. We applied a dropout with 0.4 rate to the encoders' input word embeddings.

We used a simple logistic regression classifier as the morphological parser's and lemmatizer's decoders (we compared it with a multi-layer feedforward classifier, but it didn't lead to any improvements). We utilized biaffine classifiers for the dependency parser's decoder with the 256-dim tag and 512-dim arc representation. We applied a dropout with 0.1 rate to the encoder's output before the classifiers.

In our proposed multitask approach, all decoders were trained jointly in a single model. We compared it with a single-task mode, where the lemmatizer and dependency parser were either conditioned on the morphology predictions (conditional case) or used only word embedding features (unconditional models). In the first situation, the predictions of the morphological parser were embedded and concatenated to the word embeddings.

We trained the model with slanted triangular learning rates (STLR) and discriminative fine-tuning [[Howard et al., 2018](#)]: during the first epoch pretrained embedders were frozen and during the next epochs they were trained with a learning rate by an order of magnitude lower than the learning rate of the trained from scratch parameters. The maximum learning rate of the trained from scratch parameters was 10^{-3} . The warm-up period of the STLR scheduler was 0.1 of the training iterations.

We used the data provided by the organizers of GramEval-2020³ for training and compared models on the official dev⁴ and test⁵ sets. We didn't use the 17th-century subset either for training or for evaluation: it wasn't clean enough and the gains from improvements on the data are not obvious.

We utilized allennlp [[Gardner et al., 2018](#)] and transformers [[Wolf et al., 2019](#)] libraries for our implementation. We made the code used for those experiments publicly available⁶.

¹ ruwikiruscorpora_tokens_elmo_1024_2019 from <http://rusvectors.org> (<http://vectors.nlpl.eu/repository/20/195.zip>)

² RuBERT from DeepPavlov (http://files.deepavlov.ai/deeppavlov_data/bert/rubert_cased_L-12_H-768_A-12_pt.tar.gz)

³ <https://github.com/dialogue-evaluation/GramEval2020>

⁴ <https://github.com/dialogue-evaluation/GramEval2020/tree/master/dataOpenTest>

⁵ <https://competitions.codalab.org/competitions/22902>

⁶ <https://github.com/DanAnastasyev/GramEval2020>

Fig. 1 schematically shows the model and summarizes the mentioned options.

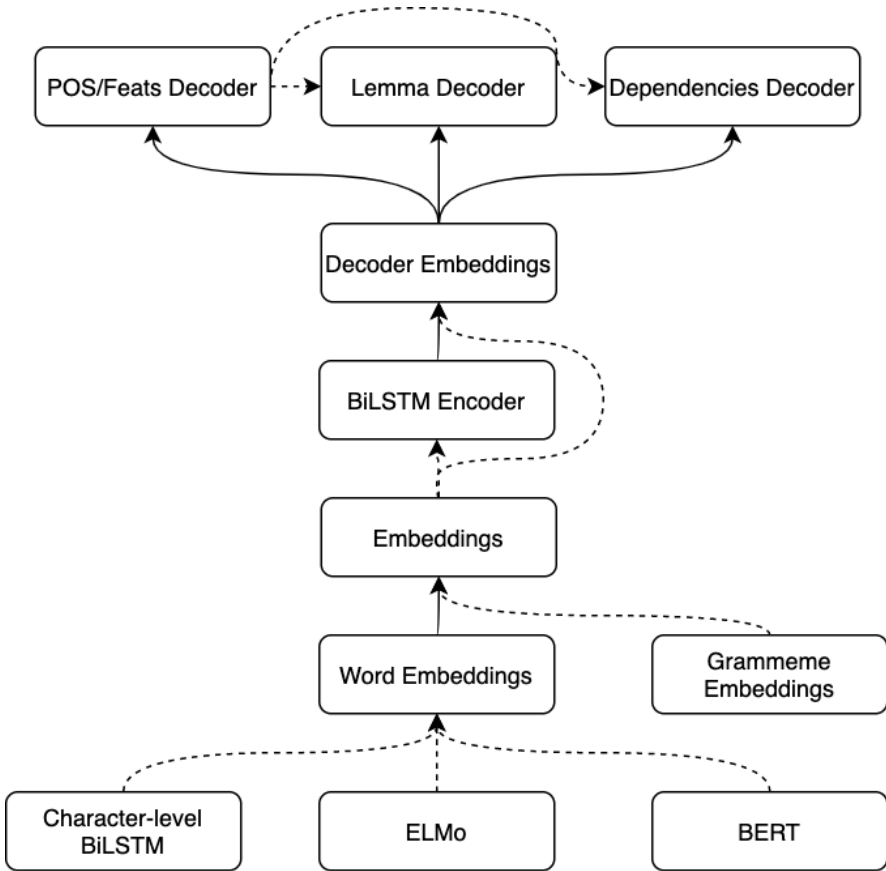


Fig. 1. Schematic representation of the explored options. The dashed lines show optional dependencies between the components

4. Experiments and Results

4.1. Embedders and Encoders Comparison

Table 1 summarizes the results of the different encoders and embedders described in **Section 3**. The prefixes in the model names correspond to different embedders: “chars_” denotes the character-level BiLSTM embedder, “elmo_” represents the ELMo-based embedder and “bert_” stands for the BERT-based embedder. Models with an additional grammeme embedding have “_morph_” substring. Models with the “_lstm” suffix use the BiLSTM encoder while models without it use the pass-through encoder. All models were trained in the joint mode.

Table 1. Comparison of different embedders and encoders on dev/test sets

Model	POS	MorphoFeats	Lemma	LAS	Overall
chars	94.7% / 91.7%	92.2% / 90.9%	95.5% / 93.6%	41.1% / 38.1%	80.9% / 78.6%
chars_lstm	97.2% / 94.1%	96.9% / 94.6%	97.3% / 95.0%	87.2% / 77.8%	94.7% / 90.4%
chars_morph_lstm	97.5% / 94.4%	97.7% / 95.2%	98.1% / 95.6%	89.6% / 77.0%	95.7% / 90.6%
elmo	97.4% / 95.4%	96.2% / 95.8%	93.1% / 92.8%	80.3% / 74.1%	91.8% / 89.5%
elmo_lstm	97.9% / 95.9%	97.5% / 95.9%	97.0% / 95.3%	88.9% / 80.3%	95.3% / 91.9%
elmo_morph_lstm	97.8% / 95.7%	97.7% / 96.1%	97.3% / 95.3%	89.5% / 79.6%	95.6% / 91.7%
bert	98.4% / 96.2%	98.3% / 96.4%	98.6% / 96.5%	93.1% / 84.6%	97.1% / 93.4%
bert_lstm	98.6% / 95.8%	98.4% / 96.3%	98.5% / 96.2%	93.2% / 83.5%	97.2% / 92.9%
bert_morph_lstm	98.4% / 95.9%	98.4% / 96.4%	98.5% / 96.4%	93.3% / 84.1%	97.2% / 93.2%
bert_random	97.0% / 92.3%	96.7% / 93.3%	97.1% / 93.1%	88.0% / 73.1%	94.7% / 88.0%

Clearly, the BERT-based models significantly outperform others. As expected, the character-level models are worse than the models with contextualized embeddings. However, the difference between the ELMo-based and character-level embeddings model with the grammeme embeddings in terms of quality is relatively insignificant.

The first model (*chars*) shows the possible performance of a completely uncontextualized model. It’s quite interesting that the model can solve morphological tasks with reasonably good quality. It demonstrates that the homonymy level in the data is moderate.

The fourth model (*elmo*) is contextualized, but the embeddings are frozen (in accordance with the guidelines from the ELMo authors). The performance of this model shows the amount of morphological and syntactic information the model learned during the pretraining step. It outperforms the uncontextualized *chars* model, especially on the syntax metrics.

BiLSTM encoder is designed to adjust the word representations, to make them more contextualized. In the case of the fine-tunable BERT, it doesn’t improve the model’s quality. For the character-level embeddings and frozen ELMo the BiLSTM encoder is the only way to adjust the representations to the task. It explains the huge enhancement of all metrics when such encoder is applied.

Grammeme embeddings encode an external knowledge about the given word. Given the performance of the models, we can hypothesize that they can improve the performance of a model trained on the provided data only. In contrast to it, BERT or ELMo seem to be able to learn all the required information from the pretraining step.

However, it’s also possible, that larger capacity of those models (the superior number of the learned parameters compared to the *chars* models) leads to their improved quality. To rule out such a hypothesis, we designed an additional experiment. We trained a BERT-based model with randomly initialized weights (*bert_random*). We used a standard random initialization from the transformers library, which is similar to the original BERT implementation. Turns out, this model is on a par with *chars_lstm* model on the dev set and even worse than it on the test set. We consider it as a proof of the importance of the pretraining step.

4.2. Frozen and Tunable Pretrained Embedders Comparison

To compare ELMo and BERT more fairly, both of them should be pretrained and fine-tuned. In **Table 2** we summarize the model qualities in those setups. The “frozen_” prefix refers to frozen embedders, while “trainable_” corresponds to the embedders, which were fine-tuned with the rest of the model.

Table 2. Comparison of frozen and fine-tuned models on dev/test sets

Model	POS	MorphoFeats	Lemma	LAS	Overall
frozen_elmo	97.4% / 95.4%	96.2% / 95.8%	93.1% / 92.8%	80.3% / 74.1%	91.8% / 89.5%
frozen_elmo_lstm	97.9% / 95.9%	97.5% / 95.9%	97.0% / 95.3%	88.9% / 80.3%	95.3% / 91.9%
trainable_elmo	98.2% / 95.5%	97.8% / 95.8%	98.2% / 95.9%	91.5% / 79.7%	96.4% / 91.7%
trainable_elmo_lstm	98.3% / 95.7%	97.9% / 95.8%	98.3% / 95.8%	92.2% / 81.2%	96.7% / 92.1%
frozen_bert	96.0% / 94.0%	95.5% / 94.3%	86.6% / 86.6%	81.7% / 76.7%	89.9% / 87.9%
frozen_bert_lstm	97.1% / 95.3%	96.6% / 95.1%	92.3% / 91.0%	86.8% / 82.0%	93.2% / 90.9%
trainable_bert	98.4% / 96.2%	98.3% / 96.4%	98.6% / 96.5%	93.1% / 84.6%	97.1% / 93.4%
trainable_bert_lstm	98.6% / 95.8%	98.4% / 96.3%	98.5% / 96.2%	93.2% / 83.5%	97.2% / 92.9%

Frozen ELMo performs surprisingly better than frozen BERT. The main cause, most likely, is the lack of alignment between the BERT’s subwords used for pretraining and the complete words required for the parsing tasks. However, BERT’s fine-tuning leads to a greater improvement of the final score, while ELMo gains from the fine-tuning step considerably less.

The latter result is somewhat surprising given the common guidelines. Nonetheless, the difference between frozen and fine-tunable ELMo is less significant when an additional encoder is used. It should be noted that the difference in the highest scores achieved by the ELMo- and BERT-based models doesn’t necessarily reflect the BERT’s architecture superiority. They have to be pretrained on the same corpus to allow a proper comparison.

4.3. Specialized and Joint Models Comparison

To ensure that we didn’t lose anything in our joint-training setup, we compared a joint BERT-based model without encoder with similar models but trained in a single-task manner. Table 3 shows the obtained results: *multitask_bert* refers to the baseline model, *conditional_bert* and *unconditional_bert* correspond to the single-task models with and without conditioning on a morphological parser’s predictions.

Table 3. Comparison of single and multitask training on dev/test sets

Model	POS	MorphoFeats	Lemma	LAS	Overall
multitask_bert	98.4% / 96.2%	98.3% / 96.4%	98.6% / 96.5%	93.1% / 84.6%	97.1% / 93.4%
unconditional_bert	98.5% / 95.7%	98.4% / 96.1%	98.6% / 96.4%	92.9% / 83.6%	97.1% / 92.9%
conditional_bert	98.5% / 95.7%	98.4% / 96.1%	98.7% / 96.5%	92.7% / 83.2%	97.1% / 92.9%
conditional_distorted_bert	98.5% / 95.7%	98.4% / 96.1%	94.2% / 89.6%	65.8% / 58.7%	89.2% / 85.0%

These models have shown approximately similar accuracies. The multitask model even outperformed single-task ones on the test set. Moreover, it’s required to perform about three times more computations on both training and inference stages to use the single-task models.

The lack of difference in the performance of the conditional and the unconditional models shows that the BERT embeddings already contain the required morphological information which is passed explicitly in the conditional model.

However, the conditional model is still prone to error propagation issue. To prove it, we passed random grammatical values to the conditional lemmatizer and the dependency parser of the conditional_bert model (*conditional_distorted_bert* row). It led to a drastic drop in quality, especially in case of dependency parsing performance.

We believe, that these findings can prove the superiority of the proposed joint model.

4.4. Robustness to an Unseen Domain

Mostly, the data we used for training came from the news and fiction domains. However, it is likely to encounter sentences from other domains in a real-life application, which may lead to a considerable drop in the model’s performance [Giesbrecht, 2009]. To estimate our model’s ability to work on an unseen domain, we evaluated models trained on the data without the poetry and without the social data. We compared them with the baseline model trained on the whole data.

Table 4 summarizes our findings. The last column shows the number of samples removed from the train set.

Table 4. Comparison of models trained without some specific domain

Model	Fiction	News	Wiki	Poetry	Social	Data D
bert	99.2% / 96.3%	97.3% / 96.0%	95.9% / 90.6%	99.0% / 93.0%	94.0% / 91.2%	0
bert_w/o_poetry	99.1% / 96.2%	97.2% / 95.9%	96.0% / 89.4%	91.1% / 91.9%	93.8% / 91.3%	-915
bert_w/o_social	99.2% / 96.2%	97.3% / 95.9%	96.0% / 88.9%	99.0% / 92.3%	93.4% / 89.7%	-2773

The quality drops on 1–1.5% on the unseen domains, which seems to be reasonably significant, but not drastic. However, the performance of the model on those genres was poorer compared to the fiction or the news subsets. So, we expect our model to work well enough on other unseen domains, yet not without errors.

4.5. Speed and Size Comparison

As we mentioned in [section 2.4.2](#), the main issue of the large pretrained models such as BERT or ELMo is their lower speed and larger size. We summarize the speed and size of the presented models in [Table 5](#).

Table 5. Comparison of speed in samples per second and size of the models

Model	Speed, it/s, CPU	Speed, it/s, GPU	Size, Mb
chars_morph_lstm	39.3	30.3	25
elmo_lstm	11.8	24.3	243
bert	3.0	8.4	702

The speed evaluation was conducted on an Intel Xeon E5-2660 v4 CPU and a Tesla M40 GPU. We used a batch size equal to 64 for ELMo-based model on GPU, a batch size equal to 16 for BERT-based model on GPU and a single sample batch for all other cases. Surprisingly, the smallest model (chars_morph_lstm) has better performance on CPU rather than on GPU. Most likely, the model is too small to benefit from the faster matrix multiplications that a GPU can provide.

As expected, BERT is significantly larger and considerably slower than others, while character-level is the fastest and the smallest model.

4.6. Comparison on GramEval-2020

We also compared our model with the results of the contestants of GramEval-2020. The results are presented in [Table 4](#). The results on the 17th century subset are ignored.

Table 6. Results of GramEval-2020

Contestant	POS	MorphoFeats	Lemma	LAS	Overall
qbic	95.8%	96.6%	96.5%	84.2%	93.3%
ADVance	95.3%	96.3%	95.8%	82.2%	92.4%
lima	95.4%	96.5%	93.6%	76.6%	90.5%
vocative	94.0%	91.9%	94.6%	73.5%	88.5%
baseline	92.0%	91.7%	92.2%	57.9%	83.5%

Clearly, our BERT-based models achieve state-of-the-art performance on this dataset.

Interestingly, the uncontextualized *chars* model from [Table 1](#) shows comparable results to the baseline. Most likely, it shows the importance of the training on the data from a similar distribution and with similar markup guidance as in the test set.

Besides, our model with the character-level embeddings (*chars_morph_lstm* from [Table 1](#)) might have achieved third place despite its simplicity. Moreover, its lemmatization quality is on par with the second-place result which proves that our proposed lemmatization technique is effective in terms of quality as well as in terms of speed.

As a result, it's possible to choose an optimal model by the quality-size tradeoff varying the embeddings and encoder, and even the smallest model will be reasonably accurate.

5. Conclusions

We proposed a joint morpho-syntactic parser which shows state-of-the-art performance on the GramEval-2020 dataset. We proved that joint training can simplify the parsing pipeline and improve its accuracy. Moreover, our designed lemmatization technique is more straightforward than the dictionary-based or the sequence-to-sequence lemmatization, but it shows promising results.

We explored different pretrained contextualized word embeddings in an application for this task and showed the way to improve the parsing quality using them. We hope that our analysis may offer an insight into the ability of the pretrained language models to increase the model's quality. Based on the results, it should be easier to make a choice on the balance between the accuracy, the size and the speed of the parser.

References

1. *Anastasyev D. et al.* (2018), Improving Part-of-Speech Tagging via Multi-task Learning and Character-level Word Representations, Computational linguistics and intellectual technologies: Proceedings of the International Conference “Dialog 2018”, pp. 14–27.
2. *Eugenie Giesbrecht, Stefan Evert* (2009). Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus, Proceedings of the 5th Web as Corpus Workshop (WAC5), San Sebastian, Spain.
3. *Jacob Devlin et al.* (2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, available at <https://arxiv.org/pdf/1810.04805.pdf>.
4. *Jenna Kanerva et al.* (2018), Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task, Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, pp. 133–142.
5. *Jeremy Howard, Sebastian Ruder* (2018), Universal Language Model Fine-tuning for Text Classification, available at <https://arxiv.org/pdf/1801.06146.pdf>.
6. *Korobov M.* (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages, Analysis of Images, Social Networks and Texts, Vol. 542, pp. 320–332.
7. *Ling Wang et al.* (2015), Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 1520–1530.
8. *Matt Gardner et al.* (2018), A Deep Semantic Natural Language Processing Platform, Proceedings of Workshop for NLP Open Source Software (NLP-OSS), Melbourne, Australia, pp. 1–6.
9. *Matthew Peters et al.* (2018), Deep Contextualized Word Representations, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, Vol. 1, pp. 2227–2237.

10. *Milan Straka* (2016), UDPipe: Trainable Pipeline Sharoff for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, pp. 4290–4297.
11. *Peng Qi et al.* (2018), Universal Dependency Parsing from Scratch, Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, pp. 160–170.
12. *Sharoff S., Nivre J.* (2011), The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Computational linguistic and intellectual technologies, pp. 591–604.
13. *Sorokin A., et al.* (2017), MorphoRuEval-2017: An Evaluation Track for the Automatic Morphological Analysis Methods for Russian, Proceedings of the International Conference “Dialog 2017”. Vol. 1, pp. 297–313.
14. *Thomas Wolf et al.* (2019), HuggingFace’s Transformers: State-of-the-art Natural Language Processing, available at <https://arxiv.org/pdf/1910.03771.pdf>.
15. *Timothy Dozat, Christopher Manning* (2016), Deep Biaffine Attention for Neural Dependency Parsing, available at <https://arxiv.org/pdf/1611.01734.pdf>.

WORD2VEC NOT DEAD: PREDICTING HYPERNYMS OF CO-HYPONYMS IS BETTER THAN READING DEFINITIONS

Arefyev N. V. (nick.arefyev@gmail.com)^{1,2,3},
Fedoseev M. V. (maxim.fedoseev13@gmail.com)¹,
Kabanov A. V. (arshehremen@gmail.com)¹,
Zizov V. S. (vzs815@gmail.com)¹

¹Lomonosov Moscow State University

²Samsung R&D Institute Russia

³National Research University Higher School of Economics,
Moscow, Russian Federation

Expert-built lexical resources are known to provide information of good quality for the cost of low coverage. This property limits their applicability in modern NLP applications. Building descriptions of lexical-semantic relations manually in sufficient volume requires a huge amount of qualified human labour. However, given some initial version of a taxonomy is already built, automatic or semi-automatic taxonomy enrichment systems can greatly reduce the required efforts. We propose and experiment with two approaches to taxonomy enrichment, one utilizing information from word definitions and another from word usages, and also a combination of them. The first method retrieves co-hyponyms for the target word from distributional semantic models (word2vec) or language models (XLM-R), then looks for hypernyms of co-hyponyms in the taxonomy. The second method tries to extract hypernyms directly from Wiktionary definitions.

The proposed methods were evaluated on the Dialogue-2020 shared task on taxonomy enrichment. We found that predicting hypernyms of co-hyponyms achieves better results in this task. The combination of both methods improves results further and is among 3 best-performing systems for verbs. An important part of the work is detailed qualitative and error analysis of the proposed methods, which provide interesting observations of their behaviour and ideas for the future work.

Key words: lexical-semantic relations, hypernymy prediction, taxonomy enrichment, distributional similarity, definition extraction, word2vec, neural language models, XLM-R

DOI: 10.28995/2075-7182-2020-19-13-32

WORD2VEC ЖИВ: ПРЕДСКАЗЫВАТЬ ГИПЕРОНИМЫ КОГИПОНИМОВ — ЛУЧШЕ, ЧЕМ ЧИТАТЬ ОПРЕДЕЛЕНИЯ

Арефьев Н. В. (nick.arefyev@gmail.com)^{1,2,3},
Федосеев М. В. (maxim.fedoseev13@gmail.com)¹,
Кабанов А. В. (arshehremen@gmail.com)¹,
Зизов В. С. (vzs815@gmail.com)¹

¹Московский Государственный Университет
им. М. В. Ломоносова

²Московский Исследовательский Центр Самсунг

³Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Известно, что созданные экспертами лексические ресурсы предоставляют информацию высокого качества, но страдают низкой полнотой. Данная проблема влияет на их применимость в современных приложениях обработки текстов на естественных языках. Описание лексико-семантических отношений в достаточном объеме требует серьезных трудозатрат. При наличии уже сформированной начальной версии таксономии системы обогащения могут существенно сократить трудоемкость задачи. Мы предлагаем и сравниваем два подхода к обогащению таксономии, один из которых использует информацию из определений слов, а второй — информацию о контекстах, в которых слова встречаются, а также комбинируем эти подходы. Первый метод ищет когипонимы отсутствующего в таксономии слова либо с помощью дистрибутивных векторных представлений слов (word2vec), либо с помощью языковых моделей (XLM-R), а затем ищет гиперонимы найденных когипонимов в уже имеющейся таксономии. Второй метод старается извлечь гиперонимы непосредственно из определений Wiktionary.

Предложенные методы были оценены в рамках соревнования по обогащению таксономии на конференции Dialogue-2020. Мы показали, что предсказание гиперонимов когипонимов позволяет достичь более высоких результатов. Комбинация обоих методов привела к дополнительному улучшению результатов и вошла в тройку лучших систем для глаголов. Важной частью работы являются детальный качественный анализ и анализ ошибок предложенных методов, которые позволили сделать ряд интересных наблюдений и сформулировать идеи для дальнейших исследований.

Ключевые слова: лексико-семантические отношения, предсказание гиперонимов, обогащение таксономии, дистрибутивная близость слов, извлечение определений, word2vec, нейронные языковые модели, XLM-R

1. Introduction

Structured resources describing lexical-semantic relations between words such as synonymy, hypernymy, meronymy, etc. are known to require a lot of time and expert labour to build. However, modern machine learning methods allow to approximately predict which relations are held between words given unstructured resources like word definitions or word usage examples. Dialogue-2020 shared task on taxonomy enrichment [10] is a competition to compare these methods. The task is to predict possible positions of a word absent in a taxonomy. More technically, for each word in a test set 10 possible synsets of a given taxonomy shall be predicted, which are either direct hypernyms or hypernyms of hypernyms of this word.

We propose two approaches to the task, named Hypernyms-of-Co-Hyponyms (HCH) and Definition Processing (DP). HCH method tries to predict co-hyponyms of the word in question using a distributional semantic model (word2vec) or a language model (XLM-R), then maps them to the taxonomy and returns their hypernyms as the result. DP method extracts hypernyms from word definitions retrieved from Wiktionary. While HCH significantly outperform DP, the best results are obtained using their combination. Our combined system was the 3rd best performing system for verbs and the 5th for nouns.

2. Related work

Since this is the participating system description paper, we will describe only those work, which our methods are based on. Please, refer to [10] for an overview of the field.

To retrieve co-hyponyms of a given word, our HCH method exploits word2vec distributional semantic model [9] or XLM-R cross-lingual language model [3]. Word2vec learns low-dimensional vector representations of words (word embeddings), that are useful to predict their contexts in unlabeled corpora. Empirically, simple algebraic operations with learnt embeddings allow modeling some aspects of corresponding words' meaning. Most importantly for our work, using cosine similarity between word embeddings allows retrieving words, which are frequently co-hyponyms of a given word. In [2] the authors trained word2vec model for the Russian language on 150GB lib.rus.ec corpora and showed that this model outperform several other methods for retrieval of semantically similar words. We employed their word embeddings for the first step of our HCH method. XLM-R is a masked language model (MLM) trained on texts from 100 languages similarly to multilingual BERT [4], but having 3x more parameters, 2x larger vocabulary and trained not only on Wikipedias, but also Common Crawl, which increased training data for low-resourced languages by orders of magnitude resulting in more than 2TB of data in total. As a MLM, it is good at predicting words that can appear in a particular position in a given context. We exploit this ability to find words, that can either replace or appear in coordination with the target word in many different contexts.

A method very similar to our DP method was proposed as one of the baselines for SemEval-2016 Task 14: Semantic Taxonomy Enrichment [5]. In this task systems had to attach a new word into an existing taxonomy. Unlike our task, not only words but also

their definitions were given as input to the systems. First Word First Sense approach was a very strong baseline, which was outperformed by only one participant. This baseline looked for the first word in the definition with the same part of speech as the given word. Then the corresponding synset was assumed to be the correct hypernym. Extending this approach, our DP method first retrieves definitions and then extracts not only words, but also phrases that can represent hypernyms of the word defined. This modification is required since in our task most synsets are represented by phrases.

3. Taxonomy enrichment methods

3.1. Hypernyms-of-Co-Hyponyms (HCH)

The Hypernyms-of-Co-Hyponyms (HCH) approach is based on the assumption, that the words distributionally similar to the target word are often its co-hyponyms, hence they have the same hypernyms we would like to predict (see [Appendix B](#) for detailed analysis regarding this assumption). As the simplest default option, we employ Skip-Gram Negative Sampling word2vec model [9] trained on 150GB lib.rus.ec corpus of books mostly in Russian [2] with cosine similarity metric to find k nearest neighbours for the target. This option works bad for words that are either absent or very rare in lib.rus.ec, thus, do not have good embeddings ([Appendix C](#) explores the correlation between word frequencies and hypernyms prediction quality). Nevertheless, we found it to be the best performing option.

Alternatively, we can find occurrences of the target word in some corpus, retrieve their contexts and ask a language model which words can replace or stand in co-ordinated row with the target in these contexts. Specifically, for each target word we have retrieved examples from the news corpora [10] using Sphinx search engine¹ and passed them to XLM-R [3]. The target can be replaced with the special token <mask>, so the model will receive the same kind of input it was trained on. However, this hides the target from the model and often results in predictions, that are plausible but entirely unrelated to the target. To stimulate predicting co-hyponyms we employ dynamic patterns proposed by [1], i.e. replace the target with a pattern like “<mask> and T” and then replace T back with the target. Thus, instead of *I love <mask>*, the model receives *I love <mask> and cats*. For each example we take 100 tokens that are most probable in the masked position as substitutes for the target in this example (this number is selected intuitively, selecting it as a hyperparameter may improve results). For each target we take k most frequent substitutes across all examples as the nearest neighbours. Since XLM-R has a vocabulary of 250K subwords shared for 100 languages, it contains only a small number of frequent Russian words. We found it beneficial to predict substitutes consisting of two subwords by inserting two masks, taking 100 most probable predictions for the first one and then one most probable continuation for each of them. We leave exploration of other multitoken substitutes generation techniques for the future work.

¹ <http://sphinxsearch.com/>

For each nearest neighbour we find all matching synsets in RuWordNet [7]. Exact matching (i.e. retrieving only synsets having exactly the same word as one of their expressions) resulted in better final performance than inexact matching with limited Levenshtein distance. For XLM-R we additionally performed lemmatization. Then for each matched synset we find its hypernyms. Finally, for each target word we return 10 most frequent hypernyms of its nearest neighbours as the result.

For a few target words, **table 1** shows their nearest neighbours, that were matched to some synsets. Nearest neighbours having direct or second level hypernyms that are also correct hypernyms of the target word, or nearest neighbours that are correct hypernyms themselves, are in bold. These nearest neighbours contribute towards correct predictions. See **section 4.1** and **appendix A** for additional analysis and discussion.

Table 1: Target words with their top-15 nearest neighbours, that were found in the taxonomy. Neighbours that resulted in correct hypernyms predicted are in **bold**

Target word	word2vec	XLM-R (<mask><mask> or T)
переиздавать	переиздать, публиковать, издаваться, рецензировать, перепечатывать, печататься, печатать, переиздание, опубликовать, издание, напечатать, двухтомник, переписать, допечатывать, перечитывать, ...	исполнять, переписывать, выдавали, повторили, дополнять, переводить, издавать, повторять, издали, издавали, переписали, ставил, повторял, показывал, издавая, ...
пылать	полыхать, гореть, пламенеть, сгорать, запылать, тлеть, сиять, кипеть, разгораться, трепетать, загораться, дымиться, вспыхивать, сверкать, загореться, ...	горят, падает, горит, спала, горела, трус, горящего, горящих, пламя, тает, гаснет, шатается, погибает, тонет, жрет, ...
прожевывать	пережевывать, прожевать, жевать, откусывать, разжевать, проглатывать, заглатывать, пережевать, глотать, съесть, наедаться, дожевать, съесть, проглотить, запивать, ...	солить, чистить, кушать, жуть, жрать, пробовать, пропускать, варить, дробить, растворять, резать, мять, жевать, смывать, удалять, ...
первомай	первомайский, субботник, праздник, новогодний, юбилей, праздничный, годовщина, праздновать, отмечать, предпраздничный, пятидесятилетие, митинг, парад, именины, летие, ...	первого мая, Праздник, рабочий день, Первомай, первое мая, Мая, Независимость, апрель, выходной, Великий, Новый год, ВОВ, рабочего дня, праздника, День труда, ...
атлетизм	спортивность, атлетичность, спорт, бодибилдинг, культуризм, выносливость, атлетический, олимпизм, мужественность, культуристский, мотоспорт, артистизм, техничность, мускулистость, физкультура, ...	гольф, ловкость, кросс, скоростью, сильным, умом, умением, активностью, фитнеса, бокса, гимнастики, секции, ретро, конь, чутье, ...

POS mapping. While the correct hypernyms always have the same part of speech (POS) as the target, nearest neighbours often contain words with different POS, resulting in incorrect predictions. Predictions with incorrect POS can be simply filtered. However, we noticed that derivationally related words with different POS usually correspond to synsets with identifiers differing only by the last letter. For instance, the synset ‘144051-N’ consists of words: *издательское дело* (publishing), *печатание* (printing), *напечатание* (typing), *издание* (publication), and ‘144051-V’: *выходить из печати* (see the light), *печатать* (type), *напечатать* (print), *издать* (to issue), *издаваться* (to be published). Changing POS in the synset identifier, i.e. replacing the letter with the desired, often results in additional correct predictions. Therefore, after finding the nearest neighbours and their synsets, for synsets with the wrong POS we replace their POS with the POS of the target. Only if the resulting synset is not found in the taxonomy, we remove it from predictions.

Hypernyms ordering and merging. We use not only direct hypernyms of nearest neighbours (denoted as *degree 1*), but also hypernyms of hypernyms (*degree 2*), they are compared in [section 4.1](#). Given a list of nearest neighbours of a particular target, we build two ordered dictionaries, for their direct and indirect hypernyms separately, with hypernyms as keys and their counts as values. Hypernyms appear in the decreasing order of their counts, and if the counts are equal, in the order of corresponding nearest neighbours. Thus, if there are several nearest neighbours with the same hypernym, this hypernym will be among the first hypernyms in the dictionary and will more likely be returned as the result. Also, if there are several hypernyms with the same counts, we will return those corresponding to more similar neighbours first. The best results are achieved by merging direct and indirect hypernyms (*degree 1 + 2*). Counts of each synset occurred as a direct and indirect hypernym of nearest neighbours are added. For hypernyms with the same counts the order of corresponding nearest neighbours are preserved. Thus, we return 10 most frequent hypernyms of nearest neighbours, preferring hypernyms of neighbours that are more similar to the target word if the counts are equal.

3.2. Definition processing (DP)

The definition processing method extracts hypernyms from Wiktionary definitions. Wiktionary is a lexical semantic resource, introduced in [\[11\]](#). It contains word definitions, examples of word usage, and some meta-information. Definitions can be classified into two large groups, *intensional* definitions try to give the sense of a term and *extensional* definitions try to impose the objects that a term describes [\[8\]](#). We are expecting to see the *intensional* ones. The simplified version of the DP method assumes that some phrase in the definition is the hypernym or at least reflects some connection with the hypernym of the word defined, and this hypernyms is already described in the taxonomy. Thus, for each N it looks at all the N-grams of the definition D and finds out which have corresponding synsets $S(N, D)$ in the taxonomy. We fix the maximum N for which $S(N, D)$ is not empty. Such N-grams and their corresponding synsets are considered to be hypernyms. For example, the phrase ‘A B C’ will be processed from $N = 3$ to $N = 1$. First, if the taxonomy has synset ‘A B C’, it will be the only answer. Next, the method will search ‘A B’ and ‘B C’ in the taxonomy and return those

that were found. If none of them was found, the same procedure will be performed for 'A', 'B', 'C' separately.

After analysing errors of this simplified method, we found situations when the extracted hypernym is not found in the original taxonomy. This problem was solved by recursively enriching of the taxonomy. In the first iteration we extract from each definition D all the N -grams with $N = 0.8 * \text{len}(D)$ and map them to the taxonomy. For definitions with non-empty $S(N, D)$ we add the defined terms to the taxonomy and specify $S(N, D)$ as their hypernyms. Then we exclude these definitions from the iterative process. We decrease N for each definition by one on each iteration and repeat hypernyms search and taxonomy enrichment procedure. Thus, terms with short definitions are added to the taxonomy first. For short definitions, the precision of each iteration (proportions of true positives in $S(N, D)$) is larger than for long definitions. The intuition is that in initial iterations we introduce less noise to the taxonomy, and that is why the noise grows more slowly in the later iterations. Thanks to such step-by-step filling in, we can detect, for example, the relation “айпад — планшетный компьютер” (“ipad” IsA “tablet computer”). This happens in the following way. There are two different definitions: *Планшет — то же, что планшетный компьютер...* (a tablet is the same as a tablet computer...), and *айпад — планшет марки iPad (ipad is an iPad brand tablet)*. First, a relation: “планшет — планшетный компьютер” (“tablet” IsA “tablet computer”) will be built, and then “айпад — планшет” (“ipad” IsA “tablet”).

4. Experiments

To develop our models and select their hyperparameters during the evaluation period, instead of using the development set provided by the organizers, we have decided to build another development set that is more similar to the test set. To achieve this, the words from the public test set were sorted by their frequencies in the news corpus, and then divided into 10 bins containing equal number of words. All the words from the given (train) part of the taxonomy were mapped to the same bins. We have sampled 50 train words from each bin. This resulted in the development set, that consists of 500 words having the same frequency distribution as the test words. However, other characteristics of the obtained development set may still be far from the test set. Hence, in during post-evaluation period we have decided to perform hyperparameter and error analysis on the public test set, leaving the private test set for the final evaluation of our methods performance.

4.1. Hypernyms-of-Co-Hyponyms

Figure 1 shows the dependence of the results of HCH method using word2vec from hyperparameters on the public test set. MAP is low if we use only the synsets of nearest neighbours (*degree 0*). This is intuitive because only few nearest neighbours are hypernyms of the target word. It is better to use direct hypernyms (*degree 1*), than indirect ones (*degree 2*). The combinations of hypernyms perform significantly better. Also, we see that larger top-k perform better. We pre-calculated only 300 nearest neighbours, however, increasing this number will likely increase the results further.

Table 2 compares HCH method using word2vec (default) and XLM-R to retrieve nearest neighbours. Evidently, word2vec results in better nearest neighbours for the task. Regarding the XLM-R model, from **figure 2** and **table 2** we see that two-subword substitutes perform much better, and “<mask><mask> or T” template is consistently better than “<mask><mask> and T”. Additionally, POS mapping gives small but consistent improvements for word2vec.

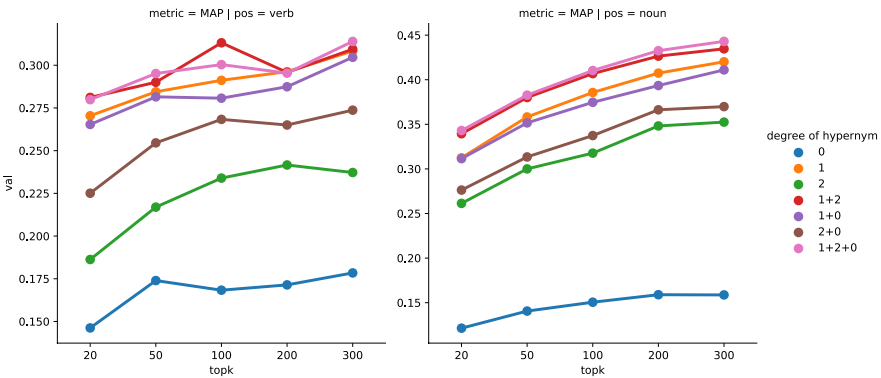


Figure 1: Evaluation of HCH on the public test set

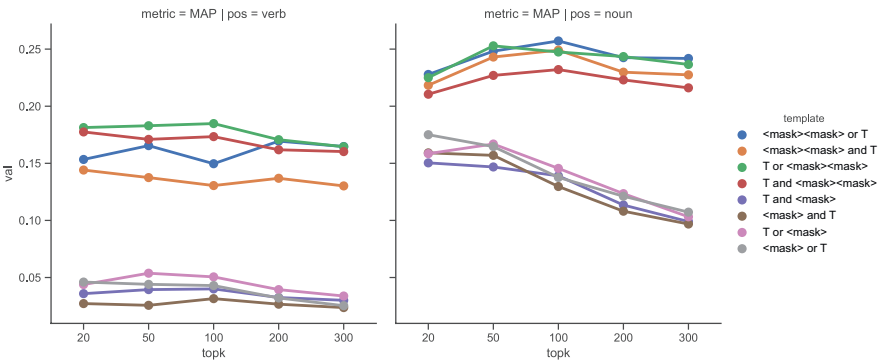


Figure 2: Evaluation of HCH with XLM-R on the public test set with degree of hypernyms 1+2

The detailed error analysis of HCH method can be found in **appendix D**. From examples in **table 1** and **appendix A**, it seems that word2vec produces less frequent and more diverse words, which are closely-related to the target, however, they are not exclusively co-hyponyms, but also topically related words. The error analysis revealed that such nearest neighbours is the largest source of errors. XLM-R generates more distantly related (and sometimes seemingly unrelated) words. One possible reason is generation of substitutes consisting of one or two subwords only, which greatly limits possible substitutes. Another possible reason is that XLM-R is trained with cross-entropy

loss resulting in prediction of frequent words first, while word2vec is trained with negative sampling loss, which promotes words having high PPMI with the target [6]. The potential advantage of XLM-R is its ability to generate not only words, but also phrases (notice *New Year* among other holidays in table 1). See appendix B for additional study of relations between targets and their nearest neighbours.

Rare words may have bad word2vec embeddings or no embeddings at all, resulting in inadequate nearest neighbours and bad performance. Appendix C investigates this problem. It shows that the performance of our method drops for the target words having less than 100–200 occurrences in lib.rus.ec. Luckily, the majority of words have more occurrences, so this is not the major problem.

4.2. Definition Processing

Table 2 shows that pure DP method is much worse than HCH method using word2vec. It is comparable with the results of HCH using XLM-R. Detailed error analysis can be found in the appendix E, while this section summarizes problems discovered.

Most errors (about 43%) are words in the test set, that do not have definitions in Wiktionary. Another problem is using exact matching to find synsets corresponding to the extracted N-grams. For example, phrases “ambulance station” and “ambulance substation” are considered different. These mistakes constitute about 10% of errors.

Table 2: Model comparison on the public test set.

DP^k means DP with k iterations. HCH^{topk, degrees} means HCH with top-k neighbours and specified degrees of hypernyms

Method	Nouns MAP	Nouns MRR	Verbs MAP	Verbs MRR
HCH ^{300, 1 + 2 + 0}	0.4430	0.4728	0.3130	0.3531
HCH ^{300, 1 + 2 + 0} w/o POSmap	0.4341	0.4615	0.3061	0.3472
XLM-R (<mask><mask> or T)	0.2480	0.2766	0.1844	0.2121
XLM-R (<mask><mask> and T)	0.2338	0.2659	0.1472	0.1723
XLM-R (<mask> or T)	0.1673	0.1926	0.0543	0.0641
XLM-R (T and <mask>)	0.1539	0.1786	0.0377	0.0410
DP ³⁰	0.1894	0.2175	0.1904	0.2519
HCH ⁵⁰ + DP ³	0.4165	0.4538	0.3122	0.3613
HCH ⁵⁰ + DP ¹²	0.4045	0.4365	0.3586	0.4072
HCH ¹⁰⁰ + DP ¹²	0.4196	0.4540	0.3548	0.4019

Some predictions also have a very low precision, i.e. small proportion of true positives in the predicted answers. Meanwhile, we can identify which of the answers are potentially correct. Particularly, this can be obtained from the fact that sentences are usually built in a sufficiently simple way. The corresponding definition was constructed on average as a single sentence with no more than two objects. It implies that one, two, or three synsets can be found in the definition. As an example, if we have target word *пустельга* (*kestrel*) and DP method returned *хищная птица, легкомысленный человек* (*predatory bird, frivolous person*) (N = 2), we consider them to be

right answers. But if we have *книга, науки, прекратить, исключительно, знакомить, потом* (*book, sciences, stop, exclusively, meet, after*) ($N = 1$) for target word *сонник* (*dream book*), it is the signal that our answer is possibly wrong. Therefore, it was able to see that if we found more than three hypernym synsets for the target word, the output is more likely to be polluted. Such results were rejected, and did not contaminate taxonomy (8%). Moreover, if N was reduced to 1, some individual irrelevant words (auxiliary verbs, as a common example) could appear in the output. At $N > 1$ most of the extracted senses were true hypernyms. The disadvantage of definitions was in a fact that the position in the taxonomy could differ by one or several relation steps from the answer. The problem with precision was solved by intersecting predictions with the HCH predictions. The completeness problem still remains.

5. Final combined model

The combined method used voting. Hypernyms predicted by both HCH and DP methods were returned first, their order in HCH answers was preserved. If the HCH returned less than 10 hypernyms, then the output was supplemented by the answers of the DP method. For instance, if HCH predicted ‘A B C D’ and DP predicted ‘C B E’, then the combined model returned ‘B C A D E’.

The results of the combined model are shown in [table 2](#). Interestingly, iteration DP method improves quality for verbs, but worsens it for nouns. In general, DP method makes HCH results more accurate, but has significant limitations if used alone.

6. Conclusion

In this paper we have proposed and evaluated two methods of taxonomy enrichment, exploiting information both from word definitions and examples of word usage. We have shown that their combination improve both of them, resulting in 3rd best result for verbs and 5th for nouns ([table 3](#)) on Dialogue-2020 taxonomy enrichment shared task.

Table 3: Final results on the private test set

Competitor	Nouns MAP	Nouns MRR	Verbs MAP	Verbs MRR
Our submissions: team <i>MorphoBabushka</i>				
HCH ^{300, 1+2} + DP ³⁰	0.4497 (5)	0.4835 (5)	0.3890 (3)	0.4419 (3)
DP ³⁰	0.1729 (13)	0.1931 (13)	0.1399 (11)	0.1690 (11)
Top participants’ submissions				
<i>cointegrated</i>	0.4178 (6)	0.4503 (6)	0.4483 (1)	0.5049 (2)
<i>Yuriy</i>	0.5522 (1)	0.5940 (1)	0.4355 (2)	0.5130 (1)
<i>xeno</i>	0.5054 (2)	0.5433 (2)	0.3075 (4)	0.3547 (4)
<i>KuKuPl team</i>	0.4976 (3)	0.5332 (3)	0.2470 (6)	0.2897 (6)
<i>RefalMachine</i>	0.4930 (4)	0.5314 (4)	0.2542 (5)	0.2969 (5)

Though it proved to be hard to apply modern language models like XLM-R to the task of taxonomy enrichment compared to the simplicity of word2vec, we believe

that further work on substitute generation using such models can result in new SOTA in this task. At the same time, using dictionaries or similar resources to solve the task seems to give bad results unless combined with other strong methods. These resources contain only a small fraction of the target words and the definitions are usually misleading. Still it worth trying to extract good definitions from the whole Web using search engines and definition extraction techniques.

7. Acknowledgements

We thank the organisers of the competition for such an inspiring task. We are grateful to our reviewers who motivated us to do detailed qualitative and error analysis of our methods, which hopefully make our paper more interesting to the readers. The contribution of Nikolay Arefyev to the paper was partially done within the framework of the HSE University Basic Research Program funded by the Russian Academic Excellence Project ‘5-100’.

Appendix A. Examples of nearest neighbours found in the taxonomy

For a few target nouns and verbs, [table 5](#) shows nearest neighbours (15 at most) from top-300 word2vec and top-100 XLM-R nearest neighbours that were matched to some synsets. Nearest neighbours that resulted in correct predictions (i.e. nearest neighbours having direct or second level hypernyms, that are also correct hypernyms of the target word, or nearest neighbours that are correct hypernyms themselves), are in bold. Words “обгорать” and “пропихнуться” don’t have nearest neighbours in XLM-R because in the news corpus there are no examples with these words, there are only with the words “обгораться” and “пропихнуть”.

Table 4: Distribution of counts of the nearest neighbours found in the taxonomy among top-100 nearest neighbours

quantile	verbs				nouns			
	word2vec	word2vec + lemm	XLM-R	XLM-R + lemm	word2vec	word2vec + lemm	XLM-R	XLM-R + lemm
0.05	1	9	0	57	0	5	4	31
0.25	7	27	3	75	4	16	11	58
0.50	16	46	32	84	9	29	25	75
0.75	26	62	67	90	15	42	39	85
0.95	43	78	85	96	39	63	61	94

[Table 4](#) shows distribution of counts of the matched nearest neighbours among top-100 nearest neighbours. Obviously, lemmatization increases the number of matched neighbours for both models. However, our experiments have shown that for word2vec it does not improve final results, so we did lemmatization only for XLM-R. This is due to the fact that given a word in its base form, word2vec often returns similar words

in their base forms also. Unlike word2vec, XLM-R is a language model and returns those forms that are appropriate given particular contexts. Also, there more matched neighbours for XLM-R than those for word2vec, probably because XLM-R generates only 2 subword units, which are mostly frequent words and phrases, hence, they are more likely described in the taxonomy.

Table 5: Nearest neighbours that were found in the taxonomy.
Neighbours that resulted in correct hypernyms predicted are in **bold**

target word	word2vec	XLM-R (<mask><mask> or T)
варево	похлебка, пойло, суп, рагу, помешивать, зелье, котелок, супчик , кушанье , котел, снадобье, половник, хлебать, размешивать, бульон , ...	квас, варка, мясо, сало, смесь, картофель, соус, зелень, ведро, супа , соль, творог, специи, еда, медь, ...
доктор	профессор , врач , психиатр , медик , хирург , коллега, ассистент, невропатолог, пациент	профессора , мастера , докторант , врачей , психолога , хирурга , специалиста, ученого, бывшего, ученый, академии, адвоката, юриста, студента , журналиста, ...
радиожурналист	тележурналист , журналист, радиоведущий , телеведущий , телеобозреватель , эссеист, колумнист , прозаик, обозреватель , корреспондент , международник, фотожурналист, литератор, публицист , телепродюсер, ...	авторов, журналистов, экспертов, редакторов, репортеров , ведущих, продюсеров, программистов, операторов, исполнителей, агентов, выпускников, писателей, фотографов, режиссеров, ...
праправнучка	правнучка , внучка, родственница, племянница, прабабка , прабабушка , правнук, ровесница, внук, дочь, кузина, праправнук, сестра , потомок , наследница , ...	внука , бабушка, дочка, наследница , вдова, наследника , няня, сынок, девица , супруга , невеста, подруга, сестра , щенка, родная, ...
задушевность	проникновенность, сердечность , музыкальность, душевность , непринужденность , доверительность, ласковость, мелодичность, искренность , непосредственность , трогательность, трепетность, выразительность, приветливость, естественность , ...	душевность , легкость, живость, искренность , откровенность , непосредственность , страсть, романтика, торжественность, чувственность, тонкость, эмоциональность, ярость, холодность, мягкость, ...
переплести	переплестать , сплести, заплести, вплести, расплести, переплетенный, переписать , перепечатать , отпечатать , скрепить, сплестись, вклеить, сцепить, обмотать, напечатать , ...	связать , разделить, развязать, совместить , связывать , синхронизировать, увязать , перевязать , сочетать , привязать, завязать, смешать , вязать , свести , объединять , ...
отвертеть	открутить , отвинтить , отвинтиться, развертеть, откручиваться, накутывать	<i>no nearest neighbours</i>

target word	word2vec	XLM-R (<mask><mask> or T)
обгорать	обгореть , облезать, шелушиться, обугливаться, выгорать, смуглеть , отслаиваться, обуглиться, чернеть , дымиться, поджариваться, загореть , плавиться, трескаться, посмуглеть , ...	падает, распадается, разводится, горит , выпадает, красится, белеет , попадает, страдает, тонет, шатается, стареет, цветет, теряется, портится, ...
пропихнуться	протолкнуться, протиснуться, протолкаться , втиснуться, продохнуть, пролезть, протискаться , давка, протискиваться, пропихиваться , впихнуть, пропихнуть, затолкать, проталкиваться , запихнуть, ...	<i>no nearest neighbours</i>
утаскивать	утащить , затаскивать, оттаскивать, увозить, уводить , забирать, выволакивать, заманивать, притаскивать, похищать , подтаскивать, вытаскивать, тащить, таскать, волочь , ...	брали, берут , везут, несут, переносили, ведут, водили , рвет, несли, выносили , носили, забрали, убирали , поднимали, бросали , ...

Appendix B. Relations between target words and their nearest neighbours

Table 6 shows the nearest neighbours, that were found in the taxonomy (10 at most), and their relation to the target. Only synsets that have a common hypernym (direct or transitive) are considered related. CH^{X+Y-} stands for co-hyponyms, which are connected to the target by a path of X upward and Y downward edges, *hypo* X and *hyper* X stand for hyponyms and hypernyms of level X and *syn* are synonyms of the target. *none* means that the target and its nearest neighbour do not have any common hypernym at any level.

Table 6: Examples of relations between target words and nearest neighbours

target word	word2vec	XLM-R
балкарка	карачаевка(CH^{1+1-}), кизилюрт(CH^{5+5-}), адыгейка(CH^{1+1-})	татарка(CH^{1+1-}), украинка(CH^{1+2-}), чешка(CH^{1+2-}), грузинка(CH^{2+2-}), чувашка(CH^{1+1-}), арабка(CH^{2+1-}), русалка(CH^{5+3-}), киргизка(CH^{1+1-}), монголка(CH^{2+2-}), северянка(CH^{5+3-})

target word	word2vec	XLM-R
узость	ограниченность(суп), односторонность(гипо1), поверхностность(CH ²⁺¹⁻), неразвитость(CH ¹⁺¹⁻), замкнутость(CH ³⁺⁴⁻), однобокость(гипо1), узкость(суп), примитивность(CH ²⁺³⁻), убогость(CH ¹⁺¹⁻), мелочность(CH ³⁺³⁻)	слабость(CH ¹⁺¹⁻), недостаточность(гипер1), ограниченность(суп), грубость(поне), незнание(CH ⁴⁺²⁻), закрытость(CH ³⁺²⁻), плотность(CH ²⁺²⁻), неопределенность(CH ³⁺³⁻), низость(CH ³⁺⁴⁻), особенность(CH ³⁺¹⁻)
интеллект	интеллектуальный(поне), разум(суп), мышление(CH ²⁺¹⁻), потенциал(CH ⁴⁺⁴⁻), интеллектуальность(суп), ум(суп), мозг(CH ⁶⁺⁵⁻), умственный(поне), сообразительность(CH ³⁺¹⁻)	нейрон(CH ⁶⁺⁴⁻), НИИ(CH ⁶⁺⁵⁻), интуиция(поне), совесть(CH ⁵⁺³⁻), смех(поне), воспитание(поне), мусор(CH ⁶⁺⁴⁻), ученый(поне), геном(CH ⁶⁺⁴⁻), искусственный(поне)
восставить	восстать(суп), бунтовать(суп), покоряться(CH ⁵⁺⁶⁻), отречься(CH ³⁺³⁻), ополчаться(CH ²⁺²⁻), роптать(CH ¹⁺¹⁻), ниспровергать(CH ⁴⁺⁵⁻), негодовать(CH ⁵⁺³⁻), ополчиться(CH ²⁺²⁻), возмущать(CH ⁴⁺⁴⁻)	становиться(поне)
застег- нуться	застегнуть(суп), расстегнуть(CH ³⁺³⁻), застегивать(суп), зашнуровать(CH ¹⁺³⁻), запахнуть(CH ²⁺²⁻), одеться(CH ⁵⁺³⁻), застегиваться(суп), надеть(CH ³⁺⁵⁻), расстегнуться(CH ³⁺³⁻), раздеться(CH ³⁺⁵⁻)	завязать(CH ³⁺⁵⁻), обуться(CH ³⁺⁶⁻), остановиться(поне), запнуться(CH ³⁺⁵⁻), встать(поне), одеться(CH ⁵⁺³⁻), убежать(CH ³⁺⁵⁻), связать(CH ¹⁺¹⁻), задеть(CH ³⁺²⁻), нажать(CH ³⁺²⁻)

Figure 3 shows the proportion of nearest neighbours related by different relations to the target words among topk nearest neighbours. These proportions are estimated on the nouns from the public test set. Only nearest neighbours, that are *synonyms*, direct or second level hypernyms (*hypernym12*), direct hyponyms (*hyponym1*), or cohyponyms connected to the target by 1 or 2 upward followed by 1 or 2 downward edges (*CoHypo12*) result in correct answers predicted. The proportion of the synonyms and suitable hypernyms shrink rapidly, when topk grows, because nearest neighbours are sorted by cosine similarity metric, and we take the less close vectors for the target word, increasing topk. Proportion of the *CoHypo12* practically does not shrink. XLM-R has much less synonym, hypernyms and *CoHypo12*, much more distantly related words, that's why it works worse than word2vec. The target words with wrong predictions have more distantly related (only by topic) or unrelated neighbours.

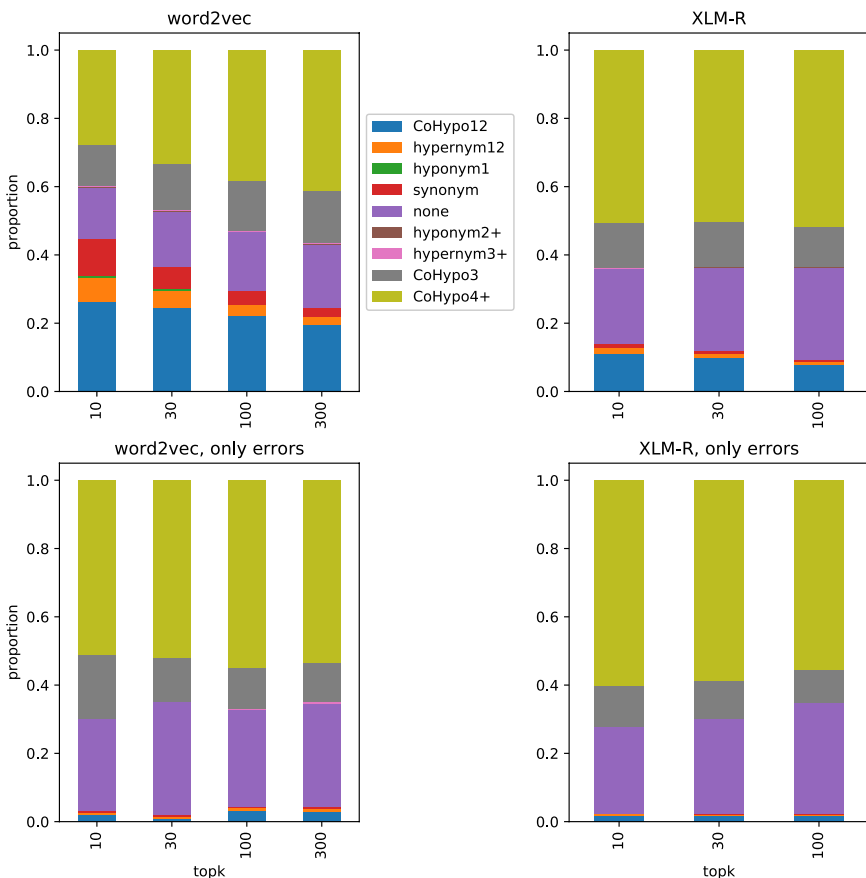


Figure 3: Proportions of nearest neighbours related by different relations to the target

Appendix C. Performance depending on the target word frequency

In this section we explore how the results of our HCH method vary depending on the frequency of the target word in the corpus that word2vec was pretrained on. **Figure 4** shows MAP and MRR for the words from the public test set depending on their frequencies in lib.rus.ec. Approximately 7% of the words have no word2vec embeddings (because they occurred less than $\text{mincnt}=5$ times in lib.rus.ec). For them, we backoff to predicting 10 most frequent hypernyms as estimated on the development set. All target words that have word2vec embeddings were divided by their frequencies into 10 bins, containing equal number of words each. Hence, performance on each bin contribute equally to the final system performance. From the figure we conclude that for the words that occurred at least about 200 times or more there is no large dependence between their frequency and performance. For the words with

less than 100–200 occurrences the performance drops significantly. Finally, most frequent hypernyms backoff doesn't work at all.

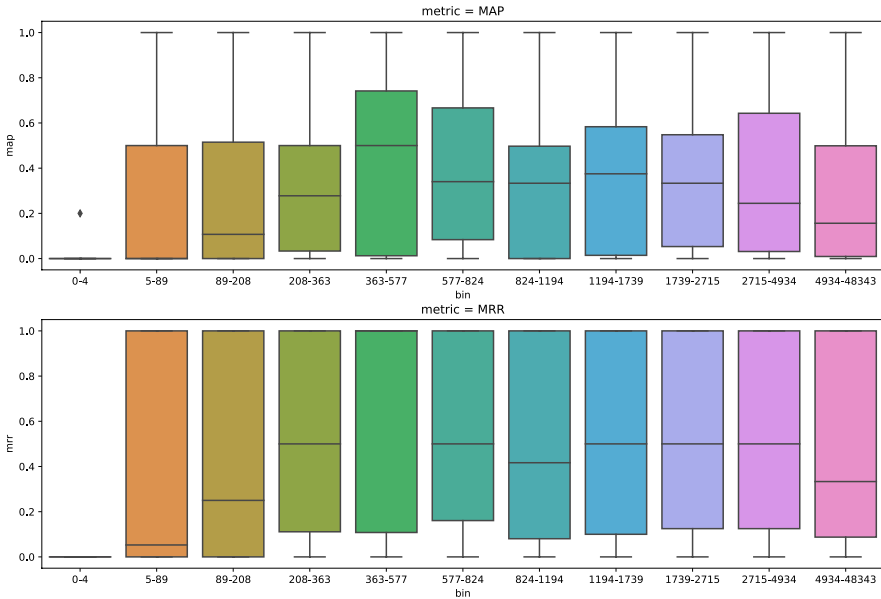


Figure 4: Dependence of the result of the word frequency

Appendix D. Error analysis of HCH method

Table 7 shows the results of error analysis of HCH method made for the public test set. In total, 100 randomly selected errors were examined, including 50 for XLM-R and 50 for word2vec. We divided all errors into two large groups: those examples which had correct answers among all HCH predictions, but not among top-10 final results (not-in-10), and those with no correct predictions at all (not-in-preds).

Related words. The most common type of errors are due to the nearest neighbours that are somehow related to the target word, but are not its co-hyponyms. For example, the word *обух* (butt/head) has correct hypernyms *сторона*, *бок*, *боковая часть* (side). Its substitutes are *топор* (axe), *тесак* (cleaver), *заточка* (sharpening), *рукоять* (handle), *клинок* (blade), which are not hyponyms of *боковая часть* (side). Potentially, this may be fixed by intersecting word2vec and XLM-R outputs. In the case of XLM-R itself, it is worth trying to experiment with other templates.

High-level hypernym. The second type of errors is related to the hierarchy of hypernyms. This means that the obtained answers of the system turned out to be either too abstract, or even the substitutes themselves are already very high-level hypernyms. An example is the word *саркома* (sarcoma) that has true answer *опухоль*, *злокачественное новообразование* (tumor) and predicted answers *болезнь* (illness),

воспаление (inflammation), *инфекционная болезнь* (infectious disease), *онкологическое заболевание* (cancer), *физическое самочувствие* (physical well-being). As another example, the word *космы* (tousle) has the true answer *прядь волос, пук, пучок* (lock of hair, topknot), while substitutes already contain a more abstract hypernym *волосы* (hair). At the same time, we have found no cases when true hypernyms are more high-level than the predicted ones. In the process of analyzing errors, it was noticed that for errors of this type the right answers nevertheless have a large number of occurrences, but not as large as those of higher-level hypernyms. Therefore, it makes sense to give higher priority to the predicted hypernyms with lower level of abstraction, rather than just the most frequent ones.

Table 7: Error types and counts

Error type	XLM-R		word2vec	
	not-in-10	not-in-preds	not-in-10	not-in-preds
Related words, not co-hyponyms:	12	6	6	17
High-level hypernym	13	2	3	6
incl. prevalence of higher-level hypernyms in the answers	11	1	3	5
incl. substitutes are already higher level hyponyms	2	1	0	1
Incompleteness of taxonomy	4	2	4	5
incl. synonymy	1	2	2	4
incl. minor meaning	3	0	2	1
Complex taxonomy	3	4	2	4
Incorrect interpretation of the word	1	3	2	1
incl. truncated words	1	0	2	1
incl. names	0	3	—	—
Total	33	17	17	33

Incompleteness of taxonomy. Some errors are not exactly errors, but cases where the system found correct hypernym that was not among gold answers. It can be a synonym of some gold answer. For example, the word *кейтеринг* (catering) has the gold answer *есть, принимать пищу* (eat, digest food), while the system predicts *питаться, кормиться* (feed). Sometimes the system predicts hypernyms of some secondary meanings of the target word. For example, *конторка* (check/bureau) has only gold hypernyms for one of its senses *письменный стол* (desk), but not the other sense *маленькая контора* (small office). Another good example is the word *аквабайк* (jetski). In the context of *Помимо дайвинга, здесь можно заняться виндсерфингом, аквабайком, кайтингом, полетать на парашюте и вертолете* (In addition to diving, you can do windsurfing, jetskiing, kiteboarding, parachuting and helicopters), the word *аквабайк* (jetski) is clearly used as a sporting event, while in the correct answers it appears only as a vehicle.

Complex taxonomy. This type of errors means that the system's answers are suitable, but very different from the correct answers. Examples are the words:

- *каторжник* (convict)—the correct answer is *лицо, отбывающее наказание* (person, sending a punishment), answers of the system is *преступник* (criminal), *осужденное лицо* (convicted person)
- *фианит* (cubic zirconia)—the correct answer is *кристалл, твердое тело, химическая продукция, синтетические материалы* (crystal, solid, chemical products, synthetic materials), answers of the system is *драгоценный камень* (gemstone), *минерал* (mineral), *корунд* (corundum), *природное минеральное образование* (natural mineral formation).

Incorrect interpretation of the word. There are errors associated with incorrect interpretation of the target word. For example, *стрип* (strip) has the correct answer *денежная ценная бумага, ценная бумага* (monetary securities, securities), while word2vec returns *стриптиз* (striptease) as its nearest neighbour. In the case of XLM-R, there are contexts where target words are used as person names or as a part of some titles. This can be fixed using named entity recognition to filter such contexts.

The performed error analysis revealed that the most common type of errors is due to the fact that the nearest neighbours are not exclusively co-hyponyms, but often topically related words. Hence, it may be worth working on better co-hyponyms prediction.

Appendix E. Error analysis of DP method

Table 8 shows the approximate percentage of errors, that were detected in a random sample (49 target words) from the public test.

In reality, some sentences have such a construction, that it is impossible to extract a hypernym from them. In addition to that, they can seek hypernym that does not explicitly correspond to any synset. In fact, we are saying that we are looking for a synset hypernym, for example, for $N = 5$. Let's suggest that we haven't found it. We go down to $N = 4$, $N = 3$, $N = 2$, $N = 1$. Eventually, we can come to the point, where our output is littered with a few words (unigrams), which are equally "suitable" for the role of hypernym. In the worst case, it leads to a drop in precision. However, if the correct hypernym can be extracted from the definition, it would be found among the results. This problem was solved because the used method was complementing the others in such a way, that the method's accuracy itself was not critical. That's why it became possible to avoid its disadvantages.

We distinguish several types of errors.

- Words, that were not presented in the dictionary.
- Too general concept as a result. It happened because of errors, appearing during matching synset with taxonomy. If definition contains some right synset 'A B', but has it in form '<target> is A with ... B', it wouldn't match with the whole synset. More evidently that only A or B would take place in the answer. It can be right in some cases, especially if B is property of A, and 'A B' is 'A', but it wouldn't be right in the full sense.
- Skip-grams. The phrase corresponding to the correct hypernym is not contiguous. I.e. N-gram was broken by additional words.

- Rejected examples. If the result has more than 4 alternatives or more than 50% words from definitions, it is more likely to be wrong. Such results were considered inconsistent and have been ignored.
- Deformed correct synset. Method hadn't recognized right synset because of it's form. This type of error emerged because of system errors. The method retrieved some hypothesis, but no one of them was right. It was occurred due to the sliding window, that stopped before the right short answer and returned wrong, but long synset.
- Extensional definitions. Such definitions formulates its meaning by specifying term extension. We found some ostensive definitions, that had only one example or some quotes. For example, we have the definition *Полевод — тот, кто занимается полеводством, возделывает зерновые, технические, кормовые и бахчевые культуры* (*Field crop is the one who practices agriculture*). The DP method can not retrieve any appropriate synset from this definition.

Table 8: Type of errors for DP method

Type of error	%	word	definition	result before matching
Words were not presented in the dictionary	43%			
Too general concept	14%	крыль	«зоол. промысловое название планктонных морских рачков»	'НАЗВАНИЕ'
Skip-gram	10%	мастиф	«древняя английская порода догообразных сторожевых собак»	'ПОРОДА'
Rejected examples	8%	сонник	«устар. книга, содержащая толкование снов Пример употребления: То вдруг велит науки прекратить, а молодых людей исключительно с одними сонниками знакомить, а потом, смотришь, сонники в печку полетели, а науки опять в чести сделались. М. Е. Салтыков-Щедрин, „В среде умеренности и аккуратности“»	'КНИГА', 'ТО', 'НАУКИ', 'ПРЕКРАТИТЬ', 'ИСКЛЮЧИТЕЛЬНО', 'ЗНАКОМИТЬ', 'ПОТОМ'
Deformed correct synset	16%	книгочей	«устар. любитель книг, чтения, знаний; книжник Пример употребления: Буфетная прислуга стала смотреть на меня исподлобья, мне говорили: — Эй ты, книгочей! Ты за что деньги получаешь? Максим Горький, „В людях“»	'МАКСИМ', 'ГОРЬКИЙ'
Extensional definitions	6%	догматик	«Значение ... 2. тот, чьё мышление отличается догматизмом»	'ТОТ'
Ostensive definition		гирька	«Пример употребления: Раз только Соломин рассердился не на шутку и так ударил своим могучим кулаком по столу, что всё на нём подпрыгнуло, не исключая пудовой гирьки, приютившейся возле чернильницы. И. С. Тургенев, „Новь“»	'РАЗ', 'ТОЛЬКО', 'НЕ', 'НА', 'И', 'ПО', 'ЧТО', 'ВСЁ', 'НА', 'НЕ', 'ИСКЛЮЧАЯ'

References

1. *Amrami, A., Goldberg, Y.*: Towards better substitution-based word sense induction. CoRR. abs/1905.12598, (2019).
2. *Arefyev, N. et al.*: Evaluating three corpus-based semantic similarity systems for russian. In: Proceedings of the 21st international conference on computational linguistics and intellectual technologies (dialogue'2015). (2015).
3. *Conneau, A. et al.*: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116. (2019).
4. *Devlin, J. et al.*: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers). pp. 4171–4186 Association for Computational Linguistics, Minneapolis, Minnesota (2019).
5. *Jurgens, D., Pilehvar, M. T.*: SemEval-2016 task 14: Semantic taxonomy enrichment. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 1092–1102 Association for Computational Linguistics, San Diego, California (2016).
6. *Levy, O., Goldberg, Y.*: Neural word embedding as implicit matrix factorization. In: Ghahramani, Z. et al. (eds.) NIPS. pp. 2177–2185 (2014).
7. *Loukachevitch, N. V. et al.*: Creating russian wordnet by conversion, (2016).
8. *Lyons, J.*: Semantics. Cambridge University Press (1977).
9. *Mikolov, T. et al.*: Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR. 2013, (2013).
10. *Nikishina, I. et al.*: RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian Language. In: Computational linguistics and intellectual technologies: Papers from the annual conference “Dialogue”. (2020).
11. *Zesch, T. et al.*: Using wiktionary for computing semantic relatedness. Proceedings of the National Conference on Artificial Intelligence. 2, 861–866 (2008).

EXPLORING SEMANTIC CONCRETENESS AND ABSTRACTNESS FOR METAPHOR IDENTIFICATION AND BEYOND

Badryzlova Yu. G. (yuliya.badryzlova@gmail.com)

National Research University Higher School of Economics,
Moscow, Russia

The paper presents a method for computing indexes of semantic concreteness and abstractness in two languages (Russian and English). These indexes are used in metaphor identification experiments in both languages; the results are either comparable to or surpass previous work and the baselines. We analyze the obtained indexes of concreteness and abstractness to see how they align with the linguistic intuitions about the corresponding semantic categories. The results of the analysis may have broader implications for computational studies of the semantics of concreteness and abstractness.

Keywords: Metaphor Identification, Semantics, Abstractness, Concreteness, Machine Learning, Distributional Semantics

DOI: 10.28995/2075-7182-2020-19-33-47

1. Introduction

1.1. The task of computational metaphor identification

Metaphor occupies a prominent place in contemporary linguistic theory: it is recognized to be one of the most powerful cognitive tools with which humans conceptualize [Lakoff & Johnson, 1980]. Metaphor is truly ubiquitous in everyday discourse, forming a fundamental part of the language system; not surprisingly, metaphor identification and interpretation pose a serious challenge to a wide range of real-world NLP applications.

Different features have been proposed to train machine learning algorithms to identify metaphor, such as lexical, morphological and syntactic, distributional semantic, and topic modelling features; features extracted from lexical thesauri and ontologies (WordNet, FrameNet, VerbNet, ConceptNet, etc.), and psycholinguistic features including concreteness and abstractness, imageability, affect, and force; for a comprehensive overview of systems for metaphor identification, see [Veale et al., 2016].

Much of the computational metaphor identification work for Russian follows the top-down design, i.e. is aimed at identifying conceptual metaphors [Dodge et al., 2015]; [Dunn et al., 2014]; [Mohler et al., 2014]; [Strzalkowski et al., 2013]. The experiments for identification of linguistic metaphor in Russian include [Badryzlova, 2019]; [Badryzlova & Panicheva, 2018]; [Panicheva, 2019]; [Panicheva & Badryzlova,

2017]; [Tsvetkov et al., 2014, 2013]. The first four of these projects explore different sets of features for metaphor classification, including lexical and morphosyntactic co-occurrence, distributional semantic embeddings, and concreteness-abstractness indexes (the present paper is an extension of the preliminary concreteness-abstractness studies from [Badryzlova, 2019]). The two latter of the aforementioned studies are based on cross-lingual model transfer, when the model is trained on English data using English lexical resources, and then the classification features are translated into Russian and other languages by means of an electronic dictionary (for more detail, see [Section 1.2](#)).

1.2. Concreteness-Abstractness feature in metaphor identification: previous research

Implementation of concreteness and abstractness in metaphor identification experiments builds on the groundwork of the theory of embodied and grounded cognition, and primary and conceptual metaphor [Barsalou, 2008, 2010]; [Lakoff & Johnson, 1999]. These theories claim that human thinking is intrinsically metaphoric, since the conceptual representations underlying knowledge are grounded in sensory and motor systems, and conceptual metaphor is the primary mechanism for transferring conventional mental imagery from sensorimotor domains to the domains of subjective experience.

Since concreteness and abstractness represent two dialectically related facets of meaning, they will be discussed under the dual term ‘Concreteness-Abstractness’.

The established method to compute indexes of concreteness and abstractness of a word is to collect two sets of lexemes (‘seed lists’) consisting of abstract and concrete words (which together constitute the ‘paradigm’ of concreteness-abstractness)—and to measure the lexical similarity between each word in the lexicon and each of the paradigm words [Tsvetkov et al., 2014, 2013]; [Turney et al., 2011]. In cases when researchers are unwilling to resort to the translational method, computation of concreteness and abstractness indexes becomes a nontrivial language-specific task requiring seed lists that are compiled purposefully for a given language.

[Turney et al. 2011] compute the Concreteness-Abstractness indexes of a word by comparing its distributional semantic embedding to the vector representations of twenty abstract and twenty concrete words. The paradigm words are automatically selected from the MRC Psycholinguistic Database Machine Usable Dictionary [Coltheart, 1981], a collection of 4,295 English words rated with degrees of abstractness by human subjects in psycholinguistic experiments. We are replicating the Turney et al. experiments in this work, therefore more detail on their experimental dataset and the obtained results will be reported in [Section 3](#).

[Tsvetkov et al. 2013] also compute the Concreteness-Abstractness indexes of English words by using a distributional semantic model and the MRC database: they train a logistic regression classifier on 1,225 most abstract and 1,225 most concrete words from MRC; the degree of Concreteness-Abstractness of a word is the posterior probability produced by the classifier.

This paper presents two Russian paradigms which are used to compute indexes of Concreteness and Abstractness for a large Russian vocabulary. We also introduce a similar English paradigm and compute Concreteness and Abstractness indexes for a large English vocabulary. Both the Russian and the English resources are made available to the community. We use the obtained indexes in machine learning experiments for linguistic metaphor identification on representative datasets in the two languages and compare the performance with previous research and the baselines. Besides, we look into the Concreteness-Abstractness rankings to see how they align with the intuitions about the semantic domains of abstractness and concreteness. To the best of our knowledge, this is the first research of this kind on Russian data.

2. Concreteness-Abstractness indexes

2.1. Russian Concreteness-Abstractness paradigms

We use two paradigms to compute the indexes of Russian words; each paradigm consists of a Concrete (Con) and an Abstract (Abs) seed list.

The first paradigm is a subset of the psycholinguistic database collected at the Kazan Federal University [Solovyev et al., 2019], therefore it will be referred to as the Kazan Paradigm in this paper. The psycholinguistic database consists of 500 most frequent Russian nouns which have been ranked by human judges to indicate the perceived degree of concreteness of each noun. We took nouns with the highest and the lowest rankings to populate the Con and the Abs seed lists of the Kazan paradigm, respectively; both animate and inanimate nouns were included in the Con list.

The second paradigm was selected from the Open Semantics of the Russian Language, the semantically annotated dataset of the KartaSlov database [Kulagin, 2019]. KartaSlov contains about 71,000 words with human annotations assigning them to one of the semantic categories of the ontology. The Abs seed list was selected from the category ABSTRACT; the nouns for the Con seed list were drawn from the class PHYSICAL ENTITY, namely, the subclasses INORGANIC, THING, and ORGANIC (with the exception of HUMAN and ANIMAL). The decision to discard the categories HUMAN and ANIMAL was prompted by the consideration that HUMAN lexemes mostly denote abstract social roles rather than concrete human beings (e.g. *адвокат* ‘lawyer’, *доброволец* ‘volunteer’, *мачеха* ‘stepmother’, etc.), and much of the ANIMAL lexicon is intrinsically metaphoric (e.g. *ворона* ‘crow’, *орел* ‘eagle’, *лиса* ‘fox’, *осел* ‘donkey’, *корова* ‘cow’, etc.). This second paradigm will be addressed to as the Moscow Paradigm throughout this paper.

As for the size of the paradigms, in the metaphor identification experiments described in Section 3, we experimented with Con-Abs indexes computed with seed lists ranging from 360 to 40 words. No substantial difference in the quality of classification was observed in relation to the size of the paradigm. Thus, we decided to follow [Turney et al. 2011] who recommend 40 as the optimal size of a seed list, since they indicated a problematic amount of overfitting when increasing the number of words. Examples from the Kazan and the Moscow Paradigms are presented in Table 1.

Table 1: Excerpts from the Moscow and the Kazan Paradigms

	Abs seed list (excerpt)		Con seed list (excerpt)	
Moscow Paradigm	домашность	'domesticity'	арфа	'harp'
	мечтательность	'reverie'	теплоход	'motorship'
	автоматизм	'automatism'	зеркальце	'compact mirror'
	распутство	'promiscuity'	пельмени	'dumplings'
	чрезвычайность	'extraordinariness'	электрогитара	'electric guitar'
	партийность	'party affiliation'	гранатомет	'grenade launcher'
	активация	'activation'	линейка	'ruler'
	буддизм	'Buddhism'	фуганок	'jointer plane'
	правильность	'correctness'	узелок	'knot'
	минимализм	'minimalism'	полупальто	'short coat'
Kazan Paradigm	жизнь	'life'	стол	'table'
	процесс	'process'	самолет	'aircraft'
	представление	'representation / show'	мальчик	'boy'
	способность	'ability'	квартира	'apartment'
	радость	'joy'	врач	'doctor'
	соответствие	'correspondence'	девочка	'girl'
	суть	'essence'	телефон	'phone'
	положение	'situation / location'	палец	'finger'
	желание	'desire'	рубли	'rouble'
	ответственность	'responsibility'	стекло	'glass'

2.2. English Concreteness-Abstractness paradigm

To compile the English Paradigm, we used the MRC Psycholinguistic Database (similarly to the previous research presented in [Section 1.2](#)). Nouns from the top and from the end of the MRC concreteness rating were drawn to populate the paradigm. We use three English Concreteness seed lists: one consisting of animate (Anim), and another of inanimate (Inan) nouns, while the third is the composite concreteness (Con) list composed of animate and inanimate nouns in equal proportions. The English seed lists are presented in [Table 2](#). It can be seen that the Anim list is predominantly comprised of words indicating animals rather than humans: this reflects the distribution of animate nouns in the top of the concreteness ranking in the MRC database, where animal nouns rank higher than those denoting humans. This may be due to the consideration discussed above in [Section 2.1](#), namely, that lexemes denoting humans mainly indicate abstract social roles rather than physical human beings.

Table 2: The English Paradigm

Anim	ape, adder, albatross, beetle, carp, cat, catfish, chicken, clown, cow, crab, deer, eagle, frog, goat, gorilla, grasshopper, hare, hen, horse, lion, mackerel, mussel, nightingale, otter, owl, ox, pig, puppy, rabbit, rat, sheep, shrimp, skunk, skylark, sparrow, stoat, stork, turtle, walrus
Inan	balloon, banana, barn, bed, bench, bluebell, bra, bridge, camera, car, carnation, casket, cauliflower, clarinet, collar, corkscrew, cucumber, daisy, egg, garlic, harpsichord, jacket, lamp, lantern, mattress, nightgown, olive, pants, pea, peach, pencil, piano, plum, potato, quilt, saxophone, ship, skyscraper, sofa, tulip
Abs	affirmation, animosity, demeanour, derivation, determination, detestation, devotion, enunciation, etiquette, fallacy, forethought, gratitude, harm, hatred, illiteracy, impatience, independence, indolence, inefficiency, insufficiency, integrity, intellect, interposition, justification, malice, mediocrity, obedience, oblivion, optimism, prestige, pretence, reputation, resentment, tendency, unanimity, uneasiness, unhappiness, unreality, value

2.3. Computing the Concreteness-Abstractness indexes

We computed Con-Abs indexes for a Russian vocabulary with a total of about 18,000 words, and for an English vocabulary of about 17,000 words.

The Russian Con-Abs indexes were computed using a pre-trained Continuous Skip-Gram distributional semantic model [Kutuzov & Kuzmenko, 2016] based on the Araneum corpus [Benko & Zakharov, 2016]. The English indexes were computed with a Continuous Skip-Gram model [Kutuzov et al., 2017] which had been pre-trained on the Gigaword 5th Edition corpus [Parker et al., 2011].

As shown in Equation 1, we measured semantic similarity (cosine distance, Sim) between the vectors of each word in the vocabulary and each word in a seed list, and took the mean of the ten nearest semantic neighbors (NN , Equations 2–3) in order to obtain the indexes.

$$\forall v_i, \forall s_j \exists D_i = \{Sim(v_i, s_1), Sim(v_i, s_2), \dots, Sim(v_i, s_j), \dots, Sim(v_i, s_k)\}, \quad (1)$$

where V is the set of words in the vocabulary, S is the set of words in the seed list, k is the number of elements in S

$$NN = \{d'_{i_1}, d'_{i_2}, \dots, d'_{i_{10}}\}, \quad (2)$$

where D'_i is a linearly ordered set of D_i (in ascending order)

$$index = Mean\{NN\} \quad (3)$$

Since we had two Con and two Abs seed lists in the Russian part of the experiment, and three Con and one Abs seed list in the English part, the following rankings were generated: Moscow Con, Moscow Abs, Kazan Con, and Kazan Abs for Russian; and English Anim, English Inan, English Con, and English Abs, for English¹.

¹ The full English and Russian rankings computed with the described method, as well as the RusMet corpus are available at <https://github.com/yubadryzlova/Concreteness-Abstractness-in-Metaphor-Identification>.

3. Metaphor identification experiments

We conduct two experiments with English data and two with Russian data.

We begin by replicating the experiments with the TroFi (Trope Finder) Example Base [Birke & Sarkar, 2006] presented in [Turney et al. 2011]. TroFi is a collection of verbal metaphor: it is built around 50 polysemous target verbs (e.g. *absorb*, *assault*, *die*, *drag*, *drown*, *smooth*, *step*, *stick*, *strike*, *touch*, etc.). Each target verb has from 1 to 115 sentences which are annotated as either literal or non-literal. For example, see the sentences with the target verb *besiege*:

- (Literal) *In 1347, Mongols < besieging > the Black Sea port of Caffa began to sicken and die from the plague.*
- (Non-literal) *... Powelson began to < besiege > me with letters asking for an invitation.*

In the first of the English experiments following Turney et al. we use a subset of 25 target verbs. Each group of sentences for a given target verb is treated as a separate learning problem, by learning and testing a separate model with ten-fold cross-validation. The performance is measured as macro-averaged Accuracy and F1-score (this experiment will be referred to as *TroFi-1* below).

In our second English experiment after Turney et al., one model is trained on the entire subset of the 25 target verbs from *TroFi-1* (1,965 sentences), which is then tested on the new unseen sentences (numbering 1,772) with the other 25 verbs (this experiment will be called *TroFi-2* in the rest of this paper).

The two Russian experiments are performed on RusMet, the Russian corpus of metaphor-annotated sentences presented in [Badryzlova & Panicheva, 2018]. The corpus consists of 7,020 sentences; each of them contains one of the 20 polysemous target verbs (e.g. *бомбардировать* ‘to bombard’, *нападать* ‘to attack’, *утюжить* ‘to iron (about clothes)’, *взрывать* ‘to explode (smth)’, *взвешивать* ‘to weigh’, etc.) which are used either metaphorically or non-metaphorically. The number of sentences per target verb ranges from 225 to 693; each of these subsets is balanced by class. The following sentences demonstrate examples of the metaphoric and non-metaphoric classes with the target verb *укалывать* ‘to prick’:

- (Metaphorical) *Это поражение серьезно < укололо > меня.* This defeat seriously piqued (lit. < ‘pricked’ >) me.
- (Metaphorical) *Самолюбие—это наполненный ветром воздушный шар, из которого вырывается буря, лишь < уколешь > его.* ‘Vanity is a balloon filled with the wind; once you < prick > it, you release a storm’.
- (Non-metaphorical) *А маникюрша, помнится, как-то до крови < уколола > мне палец.* ‘I recall that a manicurist has once < pricked > my finger so that it bled’.

Here, we start by training and testing a separate model with ten-fold cross-validation for each of the 20 subsets with individual target verbs; the overall performance is measured in terms of macro-averaged Accuracy and F1-score. Then, we split the dataset into two approximately equal parts: one containing the first ten target verbs (3,617 sentences in total), and the other containing the remaining ten target verbs (3,504 sentences). We train the classifier on the first subset and test it on the

second. These two experiments will be respectively referred to as *RusMet-1* and *RusMet-2* in the discussion below.

In all the four experiments, the metaphor identification task was formulated as sentence-level binary classification. The Support Vector Machine (SVM) classifier with linear kernel was used to learn and test the models.

For each sentence, we computed the mean of the scores of its constituent content words. The decision to use entire sentences rather than syntactic dependencies of the target verbs was motivated, firstly, by the design of the Turney et al. experiments replicated in *TroFi-1* and *TroFi-2* and, secondly, following the observations by [Mu et al. 2019] and [Badryzlova & Panicheva 2018] that accuracy of metaphor identification tends to grow with the increase of the window size, suggesting that discourse-level features consistently enhance performance of conventional, non-neural classifiers making them competitive with sophisticated neural models.

In each of the experiments we compare the performance of the Con-Abs model with the baseline lexical model in which the frequencies of lemmas are represented by the ΔP indexes of co-occurrence [Ellis, 2006]. The choice of the baseline was prompted by the findings of [Klebanov, Leong, Heilman, and Flor 2014] and [Badryzlova 2019] who reported that simple lexical unigram baselines achieve surprisingly good results for some of the datasets

Since in both of the pre-trained distributional semantic models which were used to compute the Con-Abs indexes (see Section 2.3) the tokens are lemmatized, the *TroFi* and the *RusMet* datasets were also POS-tagged. *TroFi* was preprocessed with Stanford Log-linear Part-Of-Speech Tagger [Toutanova & Manning, 2000], and *RusMet* with the Ru-syntax parser [Droganova & Medyankin, 2016].

4. Results

The results of the classification experiments are presented in Tables 3 and 4 (along with the lemma baselines).

In the *TroFi* experiments, our Con-Abs models either outperform the Turney et al. models, both in Accuracy and F1-score (sometimes by a sizeable margin, as in the case of *TroFi-1*), or they fall behind, yet by a slim margin.

When comparing the Con-Abs results with the lemma baseline, we see that in all the cases the Con-Abs models either outstrip the baseline both in Accuracy and F1-score (sometimes by a substantial margin, e.g. the Accuracy in both of the *TroFi* experiments), or they fall behind, but by a very narrow margin, as in *RusMet-1*.

The overall lower results of the *TroFi* experiments, as compared to *RusMet*, are presumably due to the unbalanced setup of the former dataset.

When comparing the performance of the models learned and tested separately for each target verb (*TroFi-1* and *RusMet-1*) with the models that were trained on one set of target verbs and tested on another set of new verbs (*TroFi-2* and *RusMet-2*), the former models, quite expectedly, produce higher results. This is not surprising considering that the second task is more challenging and ambitious, since the model has to capture semantic regularities that are common for all the verbs irrespective of the idiosyncratic patterns of their combinability.

Table 3: Results of TroFi-1 and TroFi-2 (English)

	lemma		Con-Abs	
	Acc	F1	Acc	F1
TroFi-1, Turney et al.	NA	NA	0.73	0.64
TroFi-1, this work	0.7	0.76	0.75	0.77
TroFi-2, Turney et al.	NA	NA	0.68	0.68
TroFi-2, this work	0.63	0.67	0.7	0.67

Table 4: Results of RusMet-1 and RusMet-2 (Russian)

	lemma		Con-Abs	
	Acc	F1	Acc	F1
RusMet-1	0.83	0.83	0.82	0.82
RusMet-2	0.73	0.76	0.75	0.77

Table 5 compares models based on different types of Con-Abs indexes and their combinations in the *TroFi-2* and the *RusMet-2* experiments. The uni-feature Con models outperform the Abs models twice (Mos-Con and Eng-Con vs. Mos-Abs and Eng-Abs, respectively); besides, the combination of two Con features (Mos-Con and Kaz-Con) outstrips the combination of the two Abs features (Mos-Abs and Kaz-Abs) by a safe margin—thus suggesting that indexes of Concreteness may serve as more reliable predictors of metaphoricity.

Table 5: Comparison of Con-Abs rankings and their combinations

Russian rankings	Acc	F1	English rankings	Acc	F1
Mos-Con	0.72	0.75	Anim	0.66	0.66
Mos-Abs	0.67	0.68	Inan	0.67	0.67
Kaz-Con	0.62	0.63	Con	0.68	0.68
Kaz-Abs	0.68	0.68	Abs	0.59	0.57
Mos-Con, Mos-Abs	0.74	0.77	An, Inan	0.67	0.67
Kaz-Con, Kaz-Abs	0.74	0.74	Con, Abs	0.7	0.67
Mos-Con, Kaz-Con	0.72	0.75	An, Inan, Abs	0.69	0.67
Mos-Abs, Kaz-Abs	0.69	0.69	An, Inan, Con, Abs	0.69	0.66
Mos-Con, Mos-Abs, Kaz-Con, Kaz-Abs	0.75	0.77			

The Mos-Con model substantially outperforms Kaz-Con, while both of the Russian Abs models perform on the par with each other, which may indicate that the two Abstractness rankings resemble each other. There is not much difference between Eng-Anim and Eng-Inan; besides, the combination of Eng-Anim and Eng-Inan fares

relatively similarly to the composite Eng-Con feature; this may suggest that both animate and inanimate nouns may be equally representative of the semantic category of concreteness. In the English experiments, the best result is achieved by combining only two features, Eng-Con and En-Abs, while with the Russian models the top performance is produced by combination of all the four types of indexes.

5. Concreteness-Abstractness indexes: a reliability test

In order to put our Con-Abs indexes to test, we took 451 known abstract nouns and 571 known concrete nouns from the KartaSlov thesaurus; we computed their Moscow Con-Abs indexes as described in Section 2 and clustered them using k-means algorithm with N clusters = 2.

The resulting clustering Accuracy was 0.99: Figure 1 shows that the Con-Abs indexes of known concrete and known abstract words form two distinct clusters with little overlap. The misclassified concrete nouns are *афиша* ‘poster’, *бандероль* ‘parcel’, *бланк* ‘form (document)’, and *записка* ‘note’; the misclassified abstract nouns are *номинал* ‘nominal value’ and *присловье* ‘proverb’.

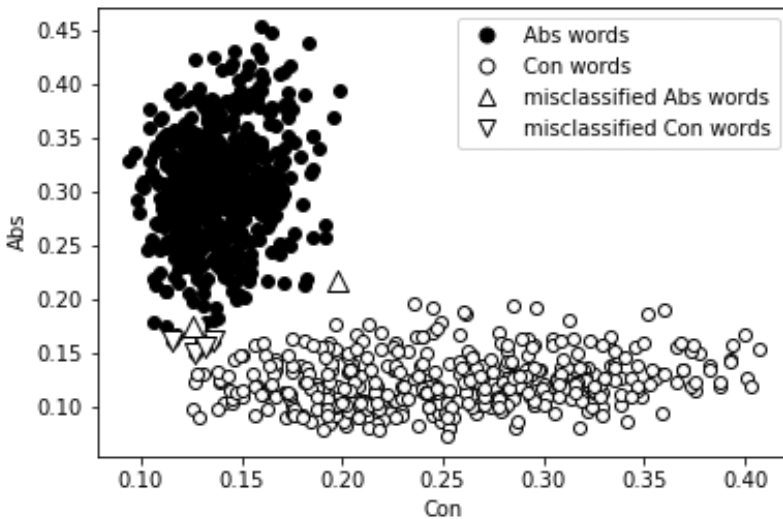


Figure 1: Reliability test: clustering known Con and Abs words (Russian)

6. A closer look at Concreteness-Abstractness indexes

The overall encouraging results of the metaphor identification experiments and the reliability test motivate us to take a more in-depth look at the computationally generated indexes of Concreteness and Abstractness to see whether their behavior aligns with the intuitive expectations about the corresponding semantic categories.

Idealistically, one could presume that Con and Abs indexes should stand in negative correlation to each other: the higher the Concreteness of a word, the lower its Abstractness should be, and vice versa. Yet, no statistical correlation between our Con and Abs rankings has been found. Obviously, the linguistic reality is not so straightforward, especially when taking into account the issue of polysemy, when different meanings of one word may vary in Concreteness and Abstractness. A glance at visualizations of the pairs of index rankings (Con vs. Abs) reveals interesting insights.

Figure 2 shows the correspondences between the Con and the Abs indexes of the Russian and the English vocabulary. The lexemes are ordered by the Con indexes (in descending order) which thus form a smooth sigmoid curve; the small dots are the Abs indexes which correspond to each of the Con indexes in the sigmoid.

In plots A and C (depicting the Moscow and the English indexes), in the beginning of the sigmoid curve, where the Con indexes are the highest, all the Abs indexes are located below the curve, with the dots forming areas of high density. This distribution indicates that there is a well-defined group of concrete words characterized by very high concreteness and very low abstractness. At the opposite end of the curve, where the Con indexes are the lowest, all the Abs indexes are above the curve, although there is no such pronounced gap between them as in the beginning: we see that, although abstract vocabulary does form a semantically uniform group, it is not as numerous as the concrete vocabulary, and the Concreteness and Abstractness are not so distinctly contrasted in it. The long steep part of the curve shows that the Abs indexes are scattered both below and above it—yet, as the curve goes down, the number of the Abs dots above the curve increases. Thus, there is a broadly defined trend for Abstractness to grow with the decrease of the Concreteness, which conforms to the linguistic intuitions; this trend is profiled in **Figure 3** which shows the number of the Abs indexes located above and below the Con line in the English vocabulary.

Table 5 demonstrates English and Russian examples of Con and Abs indexes from the beginning, the middle, and the end of the ranking. These examples are also intuitively feasible: the nouns from the top are concrete, the words in the middle are a semantic mix of the concrete and the abstract, and the words from the end of the ranking are highly abstract.

Plot B in **Figure 2**, which depicts the Con-Abs indexes computed using the Kazan paradigm, shows that the Abs indexes have much lower variance than in the Moscow and the English rankings. This may be a result of the fact that the lexemes in the Kazan paradigm are high-frequency words: thus, the mean frequency (ipm, according to [Lyashevskaya & Sharoff, 2009]) of the Moscow paradigm is 17.8, while the mean frequency of the Kazan paradigm is 244.23. Words with high frequency tend to co-occur with a wide range of vocabulary, so their distributional semantic vectors may be less indicative of the concreteness and abstractness expected of a paradigm word. Still, comparison of the Mos-Abs and the Kaz-Abs rankings provides an interesting observation: there is a strong positive Pearson correlation (0.73) between them—which is not the case between Mos-Con and Kaz-Con ($\text{corr} = 0.5$). The presence of correlation between the rankings computed with two different Abs paradigms seems to confirm the intuition that the category of Abstractness is semantically more homogeneous than the category of Concreteness. The Anim and Inan indexes of the English ranking

also show strong positive correlation (0.8), which may indicate that animate and inanimate nouns are equally representative of the semantic category of concreteness.

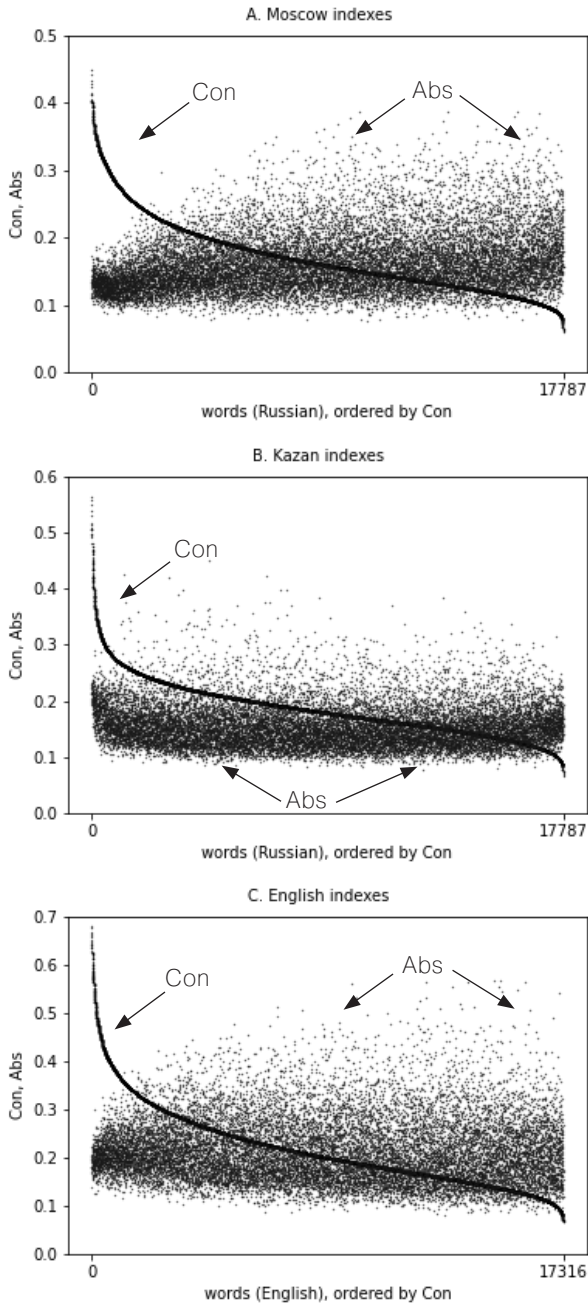


Figure 2: Correspondences between Con and Abs indexes

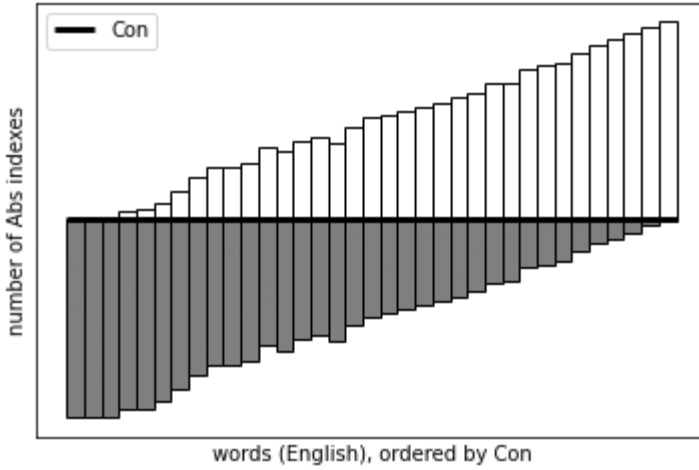


Figure 3: Number of Abs indexes above and below the Con line (English)

Table 6: Examples of indexes (Moscow and English Con-Abs) from the top, middle, and end of the ranking

	Russian word	Translation	POS	Con	Abs	English word	POS	Con	Abs
Top	наволочка	'pillow-case'	noun	0.45	0.13	shrimp	noun	0.71	0.17
	салфетка	'napkin'	noun	0.44	0.11	crab	noun	0.68	0.18
	плед	'bed throw'	noun	0.43	0.13	tomato	noun	0.68	0.2
	омлет	'omelette'	noun	0.43	0.14	scallop	noun	0.67	0.18
	одеяло	'blanket'	noun	0.42	0.13	chicken	noun	0.66	0.14
	сумочка	'purse'	noun	0.41	0.13	mussel	noun	0.66	0.18
	скатерть	'tablecloth'	noun	0.41	0.13	lobster	noun	0.66	0.17
	винтовка	'rifle'	noun	0.41	0.13	oyster	noun	0.65	0.18
	шарф	'scarf'	noun	0.4	0.13	onion	noun	0.65	0.19
соус	'sauce'	noun	0.4	0.15	potato	noun	0.64	0.2	
Middle	торрент	'torrent'	noun	0.16	0.14	healey	noun	0.2	0.17
	нагло	'brazenly'	adv	0.16	0.19	swing	verb	0.2	0.16
	бодро	'cheerfully'	adv	0.16	0.15	carrier	noun	0.2	0.15
	раж	'zeal'	noun	0.16	0.21	smash	verb	0.2	0.11
	вальдорфский	'waldorf'	adj	0.16	0.14	unload	verb	0.2	0.14
	впереди	'in front of'	adv	0.16	0.14	stanford	noun	0.2	0.16
	продавать	'sell'	verb	0.16	0.09	north	adv	0.2	0.15
	потренироваться	'practice'	verb	0.16	0.12	perilous	adj	0.2	0.21
	вековой	'centennial'	adj	0.16	0.18	enumeration	noun	0.2	0.29

	Russian word	Translation	POS	Con	Abs	English word	POS	Con	Abs
End	<i>согласоваться</i>	'correspond'	verb	0.07	0.19	<i>vitro</i>	propn	0.2	0.12
	<i>дифференциация</i>	'differentiation'	noun	0.07	0.27	<i>open-ended</i>	adj	0.11	0.24
	<i>вовлечение</i>	'involvement'	noun	0.07	0.26	<i>inter-departmental</i>	adj	0.11	0.18
	<i>распространение</i>	'proliferation'	noun	0.07	0.2	<i>rejection</i>	noun	0.11	0.36
	<i>разделение</i>	'separation'	noun	0.07	0.27	<i>discredit</i>	verb	0.11	0.31
	<i>тотальный</i>	'total'	adj	0.07	0.25	<i>hold</i>	verb	0.11	0.13
	<i>пониматься</i>	'be regarded as'	verb	0.07	0.22	<i>legitimate</i>	adj	0.11	0.28
	<i>всесторонне</i>	'comprehensively'	adv	0.07	0.16	<i>unsuccessfully</i>	adv	0.11	0.15
	<i>выявлять</i>	'detect'	verb	0.07	0.14	<i>long-awaited</i>	adj	0.11	0.19
<i>переподчинение</i>	're-subordination'	noun	0.07	0.23	<i>abolition</i>	noun	0.11	0.25	

7. Conclusions

We have presented a method for computing indexes of semantic Abstractness and Concreteness of words in two languages—Russian and English, and applied them to the task of linguistic metaphor identification in these two languages.

The results of the classification are either comparable with previous research or surpass it, the same holds for the baseline model. The efficiency of metaphor identification with Concreteness-Abstractness indexes suggests that metaphoric and non-metaphoric contexts are semantically different in terms of Concreteness and Abstractness, which conforms to the Conceptual Metaphor Theory. Models based on Concreteness indexes are more efficient than the models based on Abstractness, suggesting that Concreteness may be a more reliable predictor of metaphoricity.

Analysis of the Concreteness-Abstractness indexes reveals that there is a general trend for Abstractness indexes to increase as the corresponding Concreteness indexes decrease, which is in accordance with linguistic intuitions. There is a distinct group of highly concrete words in the lexicon which have very high Concreteness and very low Abstractness indexes; similarly, there is a group of distinctly abstract vocabulary, with low Concreteness and high Abstractness scores.

Presence of statistical correlation between two Russian Abstractness ratings (and absence of correlation between the respective Concreteness ratings) may indicate that the category of Abstractness is more semantically homogeneous than the category of Concreteness. Statistical correlation between the English concrete Animate and Inanimate rankings suggests that both animate and non-animate nouns may be equally representative of the semantic category of Concreteness.

References

1. *Badryzlova, Y.* (2019). Automated metaphor identification in Russian texts. National Research University Higher School of Economics.
2. *Badryzlova, Y., & Panicheva, P.* (2018). A Multi-feature Classifier for Verbal Metaphor Identification in Russian Texts. Conference on Artificial Intelligence and Natural Language, 23–34.
3. *Barsalou, L. W.* (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59, 617–645.
4. *Barsalou, L. W.* (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, 2(4), 716–724.
5. *Benko, V., & Zakharov, V.* (2016). Very large Russian corpora: New opportunities and new challenges. In *Computational linguistics and intellectual technologies*. Russian State University for the Humanities.
6. *Birke, J., & Sarkar, A.* (2006). A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language. *EACL*.
7. *Coltheart, M.* (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4), 497–505.
8. *Dodge, E., Hong, J., & Stickles, E.* (2015). MetaNet: Deep semantic automatic metaphor analysis. *NAACL HLT 2015*, 40.
9. *Droganova, K., & Medyantkin, N.* (2016). NLP pipeline for Russian: An easy-to-use web application for morphological and syntactic annotation. Proceedings of the Annual International Conference “Dialogue”. Annual International Conference “Dialogue”, Moscow.
10. *Dunn, J., de Heredia, J. B., Burke, M., Gandy, L., Kanareykin, S., Kapah, O., Taylor, M., Hines, D., Frieder, O., & Grossman, D.* (2014). Language-Independent Ensemble Approaches to Metaphor Identification. Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence.
11. *Ellis, N. C.* (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1), 1–24.
12. *Klebanov, B. B., Leong, B., Heilman, M., & Flor, M.* (2014). Different texts, same metaphors: Unigrams and beyond. Proceedings of the Second Workshop on Metaphor in NLP, 11–17.
13. *Kulagin, D.* (2019). Opyt sozdaniya mashinno-proveryaemoy semanticheskoy razmetki russkix sushhestvitel'nyx [Developing computationally verifiable semantic annotation of Russian nouns]. Annual International Conference “Dialogue”, Moscow.
14. *Kutuzov, A., Fares, M., Oepen, S., & Velldal, E.* (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. Proceedings of the 58th Conference on Simulation and Modelling, 271–276.
15. *Kutuzov, A., & Kuzmenko, E.* (2016). WebVectors: A toolkit for building web interfaces for vector semantic models. International Conference on Analysis of Images, Social Networks and Texts, 155–161.
16. *Lakoff, G., & Johnson, M.* (1980). *Metaphors We Live By* (2nd ed.). The University of Chicago Press.
17. *Lakoff, G., & Johnson, M.* (1999). *Philosophy in the Flesh* (Vol. 4). New York: Basic books.

18. *Lyashevskaya, O., & Sharoff, S.* (2009). Chastotny`j slovar` sovremennogo russkogo yazy`ka na materialax Nacional`nogo korpusa russkogo yazy`ka [A frequency dictionary of modern Russian language on the basis of the Russian National Corpus]. Azbukovnik.
19. *Mohler, M., Rink, B., Bracewell, D. B., & Tomlinson, M. T.* (2014). A Novel Distributional Approach to Multilingual Conceptual Metaphor Recognition. COLING, 1752–1763.
20. *Panicheva, P.* (2019). Analiz parametrov semantičeskoj svjaznosti s pomošč`ju distributivny xsemantičeskix modelej (na materiale russkogo jazyka) [Analysis of parameters of semantic coherence by means of distributional semantic models (on Russian data)]. St. Petersburg State University.
21. *Panicheva, P., & Badryzlova, Y.* (2017). Distributional semantic features in Russian verbal metaphor identification. Computational Linguistics and Intellectual Technologies, 1, 179–190.
22. *Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K.* (2011). English Gigaword Fifth Edition LDC2011T07 (Tech. Rep.). Technical Report. Linguistic Data Consortium, Philadelphia.
23. *Solovyev, V., Andreeva, M., Solnyshkina, M., Zamaletdinov, R., Danilov, A., & Gaynutdinova, D.* (2019). Computing Concreteness Ratings of Russian and English Most Frequent Words: Contrastive Approach. 2019 12th International Conference on Developments in ESystems Engineering (DeSE), 403–408.
24. *Strzalkowski, T., Broadwell, G. A., Taylor, S., Feldman, L., Yamrom, B., Shaikh, S., Liu, T., Cho, K., Boz, U., Cases, I., & others.* (2013). Robust extraction of metaphors from novel data. Proceedings of the First Workshop on Metaphor in NLP, 67–76.
25. *Toutanova, K., & Manning, C. D.* (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Proceedings of the 2000 Joint SIG-DAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13, 63–70.
26. *Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C.* (2014). Metaphor detection with cross-lingual model transfer. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 248–258.
27. *Tsvetkov, Y., Mukomel, E., & Gershman, A.* (2013). Cross-lingual metaphor detection using common semantic features. Proceedings of the First Workshop on Metaphor in NLP, 45–51.
28. *Turney, P. D., Neuman, Y., Assaf, D., & Cohen, Y.* (2011). Literal and metaphorical sense identification through concrete and abstract context. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 680–690.
29. *Veale, T., Shutova, E., & Klebanov, B. B.* (2016). Metaphor: A Computational Perspective. Synthesis Lectures on Human Language Technologies, 9(1), 1–160.

ДИНАМИКА СТИЛЯ РУССКОЙ ПИСЬМЕННОЙ РЕЧИ XIX ВЕКА: КОРПУСНЫЙ ЭКСПЕРИМЕНТ¹

Баранов А. Н. (baranov_anatoly@hotmail.com),
Добровольский Д. О. (dobrovolskij@gmail.com)

Институт русского языка РАН, Институт
языкознания РАН, Москва, Россия

Рассматривается гипотеза о том, что распределение дискурсивных слов характеризует тенденции развития стиля письменной речи XIX века. Излагаются и обсуждаются результаты эксперимента на материале Национального корпуса русского языка по исследованию частоты использования дискурсивных слов с семантикой эпистемической модальности: *конечно, разумеется, по-видимому, как кажется, казалось бы, наверно, вероятно, пожалуй, действительно* и др. Показывается, что частоты этой группы выражений возрастают во второй половине XIX века. Аналогичная тенденция прослеживается также для некоторых синтаксических конструкций с той же семантикой: *(я) думаю, что...; (я) считаю, что...; (мне) кажется, что...* Выявленная закономерность рассматривается как дискурсивная практика в изменении стиля художественной литературы, которая заключалась в расширении модусной части высказывания по сравнению с более ранним периодом.

Дискурсивная практика расширения модуса была присуща лишь группе писателей-новаторов (прежде всего, Ф. М. Достоевскому, М. Е. Салтыкову-Щедрину, Л. Н. Толстому, И. А. Гончарову, А. Ф. Писемскому, П. И. Мельникову-Печерскому, Н. С. Лескову и И. С. Тургеневу), которые, однако, в силу таланта, общественной значимости и количества опубликованных текстов оказали существенное влияние на язык художественной прозы.

Задача исследования динамики художественного стиля заключается в выявлении и описании комплекса дискурсивных практик, формирующих письменный дискурс как таковой.

Ключевые слова: дискурсивные слова, авторский стиль, корпусный подход, динамика стиля

DOI: 10.28995/2075-7182-2020-19-48-61

¹ Работа выполнена в рамках проекта № 18-012-90025, поддержанного РФФИ.

STYLE DYNAMICS OF THE RUSSIAN WRITTEN SPEECH OF THE 19TH CENTURY: A CORPUS STUDY

Baranov A. N. (baranov_anatoly@hotmail.com),

Dobrovol'skij D. O. (dobrovolskij@gmail.com)

Russian Language Institute of the RAS, Institute of Linguistics of the RAS, Moscow, Russia

The starting point of the present paper is the hypothesis that the distribution of discursive words characterizes the trends in the development of the writing style of the 19th century. The paper presents and discusses the results of an experiment based on the data of the Russian National Corpus on the frequency of using discursive words with the semantics of epistemic modality, such as *konechno*, *razumeetsya* (both roughly meaning 'of course'), *po-vidimomu* 'apparently', *kak kazhetsya*, *kazalos' by* (both ≈ 'it would seem'), *naverno* ≈ 'as it were', *veroyatno* 'probably', *pozhaluy* ≈ 'maybe', *deystvitel'no* 'really', etc. We show that the frequency of this group of expressions increases in the second half of the 19th century. A similar trend is also observed for some syntactic constructions with the same semantics: (*ya dumayu*, *chto...* '(I) think that...'; (*ya schitayu*, *chto...* '(I) believe that...'; (*mne*) *kazhetsya*, *chto...* 'it seems to me that'. The revealed regularity is considered as a discursive practice in changing the style of fiction, which consisted in expanding the modus part of the utterance as compared to the earlier period.

The discursive practice of expanding the modus was inherent only to a group of innovative writers (first of all, F. M. Dostoevsky, M. E. Saltykov-Shchedrin, L. N. Tolstoy, I. A. Goncharov, A. F. Pisemsky, P. I. Melnikov-Pechersky, N. S. Leskov, and I. S. Turgenev), who, however, due to their talent, social significance, and the number of published texts, had a significant impact on the language of fiction.

The task of studying the dynamics of artistic style is to identify and describe a set of discursive practices that establish written discourse as such.

Key words: discursive words, individual style, corpus approach, style dynamics

1. Введение в проблему

Исследования стиля письменной речи XIX в. в основном представлены работами по особенностям текстов художественной литературы (см., например, известные работы В. В. Виноградова — [Виноградов 1961; 1980], а также [Левин 1958; 1964], [Эйхенбаум 1969]). При этом в основе метода исследования лежал индивидуальный подход ученого к пониманию текста, учитывающий в большей или меньшей степени литературные особенности текста или его отдельные лингвистические характеристики. Достоверность и полнота получаемых

результатов всецело определялась кругозором исследователя и его индивидуальным восприятием формальной стороны текста, а также его семантики и прагматики. Отсутствие репрезентативных корпусов текстов, а также эффективных инструментов их обработки не давало возможности изучать возникновение и развитие языковых феноменов в широком диахроническом плане, хотя отдельные наблюдения исследователей имели определенную ценность — в том числе и для лингвистической теории [Лотман 1995]; [Пеньковский 2005].

С созданием представительных корпусов в отечественной традиции появились лингво-статистические исследования языковых феноменов дискурсов различных типов, в том числе и художественных текстов второй половины XIX в. [Шайкевич, Андрущенко, Ребецкая 2003; 2013; 2016]. В центре этих работ стояли распределения частоты употребления слов, входящих в словник корпусов, сформированных для целей каждого из этих исследований.

Корпусный подход к изучению стиля предполагает иной взгляд на текстовый материал. Анализ частоты употребления слов и синтаксических конструкций основан на выявлении тенденций в использовании лексических единиц определенных групп. Лингво-статистическая модель стиля предполагает анализ частотного распределения групп лексем, важных с точки зрения общих тенденций стилевых изменений. В данном случае мы исходим из весьма вероятной гипотезы о том, что развитие стиля художественной литературы в XIX веке и письменной речи в целом происходило за счет усложнения оформления пропозиционального содержания модальным компонентом высказывания. Иными словами, в динамике изменения стиля появлялись новые дискурсивные практики. Общепризнанная простота стиля прозы А. С. Пушкина связана с тем, что пропозиция отражала не только описываемое положение дел (ситуацию), но и ментальные и эмоциональные состояния персонажей, отношение к ним автора и другие смыслы, которые во второй половине XIX века стали оформляться как часть модуса.

Мы предполагаем, что к середине XIX века форма выражения смысла в художественном тексте начинает усложняться. Некоторая часть смыслов, «внешних» по отношению к описываемому положению дел, концентрируются в модальной рамке высказывания. Соответственно, многие лексические единицы и синтаксические конструкции, не имевшие ранее дискурсивных функций, их приобретают (см., например [Баранов, Добровольский 2018; 2019]). Этот процесс хорошо виден на примере ряда крупных писателей второй половины XIX в. Разумеется, различные авторы следуют этой тенденции в разной степени. Ф. М. Достоевский, А. Ф. Писемский и М. Е. Салтыков-Щедрин, скорее, принадлежат к новаторам, формировавшим новую стилистическую традицию, что проявлялось в двух отношениях. Во-первых, авторы-новаторы использовали больше собственно дискурсивных слов, а также недискурсивных слов в дискурсивных употреблениях. Во-вторых, спектр значения слов, допускающих дискурсивное употребление, у них был существенно шире, чем у авторов условно «старого» стиля. Таким образом, указанная тенденция проявляется в письменном тексте в том, что увеличивается количество дискурсивных слов и недискурсивных слов в дискурсивных употреблениях.

Рассмотрим указанную тенденцию к усложнению оформления пропозиции на материале Национального корпуса русского языка (НКРЯ) — на его подкорпусе, охватывающем период с 1800 по 1900 г. (по дате создания текста). Репрезентативность этого подкорпуса внешнему пользователю оценить довольно трудно. Общий объем выбранного подкорпуса: 55 271 174 слова². Мы исходим из того, что этот подкорпус репрезентативен для большинства стилистических изменений, поскольку, например, диахронический корпус немецкого языка «Deutsches Textarchiv» (<http://www.deutschestextarchiv.de/>), включающий тексты с 1600 по 1900 г., в целом содержит 150 млн словоупотреблений, то есть в среднем на столетие приходится порядка 40 млн словоупотреблений.

В качестве объекта исследования выбрана группа единиц, которые в традиционной русской грамматике называются, «вводными словами и вводными сочетаниями слов, содержащими оценку степени достоверности сообщения» [Грамматика русского языка 1954: 153–154]: *конечно, разумеется, само собой разумеется, без сомнения, без всякого сомнения, вне всякого сомнения, вполне очевидно, судя по всему, по всей видимости, по-видимому, видимо, как кажется, казалось бы, как представляется, думается, наверно, как бы, вероятно, по всей вероятности, пожалуй, поистине, действительно*. Выбор этой группы единиц, которые далее называются вводными словами и выражениями эпистемической модальности, объясняется тем, что они в подавляющем количестве употреблений входят в модальную рамку высказывания. Отметим, что в более поздней «Русской грамматике» эта группа вводных слов описана не вполне удачно [Русская грамматика 1980].

К этой группе вводных слов и выражений относятся также формы: *бесспорно, очевидно, само собой, как видно, возможно, видимо, как бы* и некоторые другие, имеющие значительное количество недискурсивных употреблений, ср. *Это очевидно / бесспорно; Все произошло само собой; Как видно из окна, солнце скрылось*. Исследование контекстов подобных форм требует ручной фильтрации дискурсивных и недискурсивных употреблений, что представляет существенную проблему при большом объеме корпуса³. По этой причине такие выражения в проведенном эксперименте не исследовались.

Другая причина, по которой некоторые из дискурсивных единиц рассматриваемой группы не учитывались в эксперименте, — это низкая частотность употребления в НКРЯ. Так, выражение *как представляется* насчитывает всего 9 вхождений, *по всей видимости* — 9 вхождений, *вполне очевидно* — 7 вхождений. Низкая частота употребления не дает возможности сделать правдоподобные выводы о временном распределении.

² Все приводимые в статье статистические данные, если это не оговорено особо, приводятся по Национальному корпусу русского языка.

³ Отметим, что ориентация на пунктуационное выделение дискурсивных употреблений некорректна, поскольку нынешняя пунктуационная норма в отношении вводных слов в XIX в. не соблюдалась (точнее — только складывалась). Кроме того, далеко не все дискурсивные слова обособляются на письме.

Семантика усложнения модуса высказывания может также передаваться синтаксическими конструкциями с глаголами ментальной сферы. Ср. (я) *думаю, что...*; (я) *считаю, что...*; (я) *полагаю, что...*; (мне) *кажется, что...* Некоторые из них были привлечены для данного эксперимента с целью сопоставления с результатами, полученными для дискурсивных слов. Отсутствие корреляции между этими данными свидетельствовало бы о ложности исходной гипотезы.

Далее обсуждаются результаты проведенного эксперимента для указанной группы дискурсивных слов, разбираются контрпримеры, проводится анализ сопоставимых по семантике синтаксических конструкций, а также для некоторых единиц рассматриваются особенности частотного распределения по конкретным авторам.

2. Частотное распределение основной группы вводных слов эпистемической модальности

Результаты эксперимента на выбранном подкорпусе НКРЯ показывают, что количество вхождений соответствующих форм возрастает к середине XIX в., а затем либо держится на том же уровне, либо несколько падает. Ниже на **рис. 1** и **рис. 2** отражена динамика частоты употребления для вводных слов *вероятно, в сущности, действительно, пожалуй*⁴ (**рис. 1**), *по-видимому, конечно и разумеется* (**рис. 2**).

Большая сглаженность кривых для лексем *в сущности* и *по-видимому* объясняется другой размерностью: общее количество словоупотреблений *конечно* и *действительно* (25 931 и 15 782 соответственно) столь велико, что выбор одинаковой размерности существенно сглаживает кривые распределения менее частотных единиц (*в сущности* — 3052, *по-видимому* — 4824).

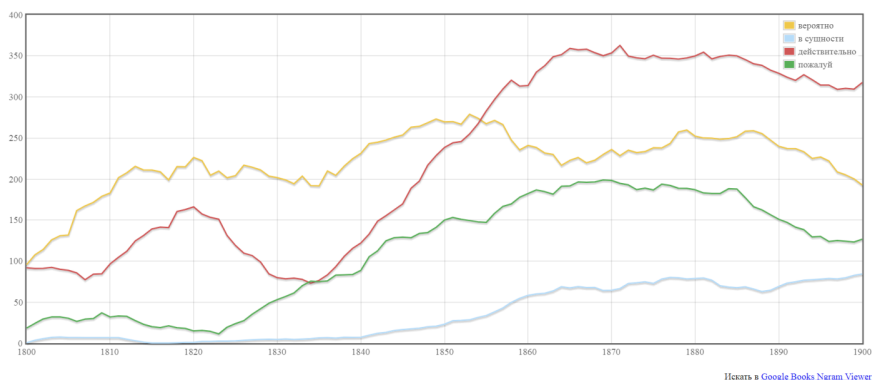
Аналогичное распределение по годам обнаруживают следующие вводные слова и сочетания слов: *само собой разумеется, без всякого сомнения, как кажется, казалось бы, думается, наверно, по всей вероятности, действительно, (как) надо полагать, (как) надо думать, как оказалось*. «Аналогичным» такое распределение является в том смысле, что в течение XIX века наблюдается общее увеличение частоты употребления соответствующих единиц, при этом варьирование в частоте может быть весьма значительным. Так, точка максимума может достигаться в середине XIX века, в 70–80 годах, с небольшими отклонениями продолжаться до конца века. Кроме того, возможны случаи, когда единица имеет несколько пиков частотности.

⁴ Дискурсивное слово *пожалуй* имеет недискурсивный омоним, ср. — *Пожалуй, мой друг, сюда поближе и скажи что такое?* Однако частота его употреблений незначительна по сравнению с дискурсивным вариантом.

Распределение по годам (частота на миллион словоформ)

Слова или сочетания слов, через запятую:

Годы с по со сглаживанием



[Показать таблицу](#)

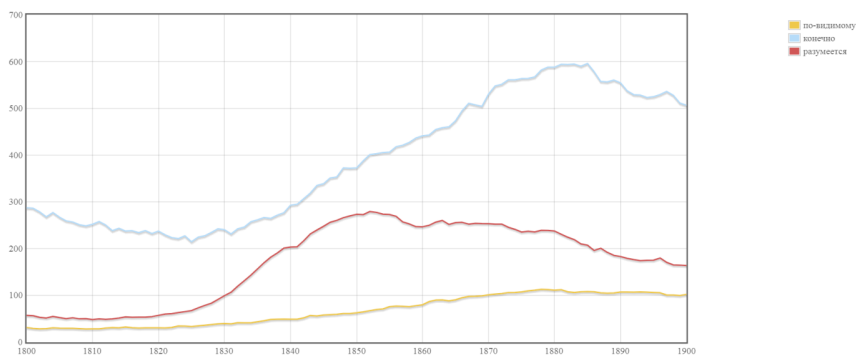
Читать в [Google Books Ngram Viewer](#)

Рис. 1. Частотное распределение по годам вводных слов *вероятно, в сущности, действительно и пожалуй*

Распределение по годам (частота на миллион словоформ)

Слова или сочетания слов, через запятую:

Годы с по со сглаживанием



[Показать таблицу](#)

Читать в [Google Books Ngram Viewer](#)

Рис. 2. Частотное распределение по годам вводных слов *по-видимому, конечно и разумеется*

Тип кривой отражает характер изменения частоты единицы — выпуклые кривые соответствуют относительно постепенному возрастанию частоты, а вогнутые — более динамичному, резкому. В основном на [рис. 1](#) и [рис. 2](#) представлены выпуклые кривые, за исключением распределения единиц *конечно* и *действительно*, которое сочетает вогнутые и выпуклые части. Хороший пример чисто вогнутой кривой представлен в распределении частоты выражения *казалось бы* на [рис. 3](#).

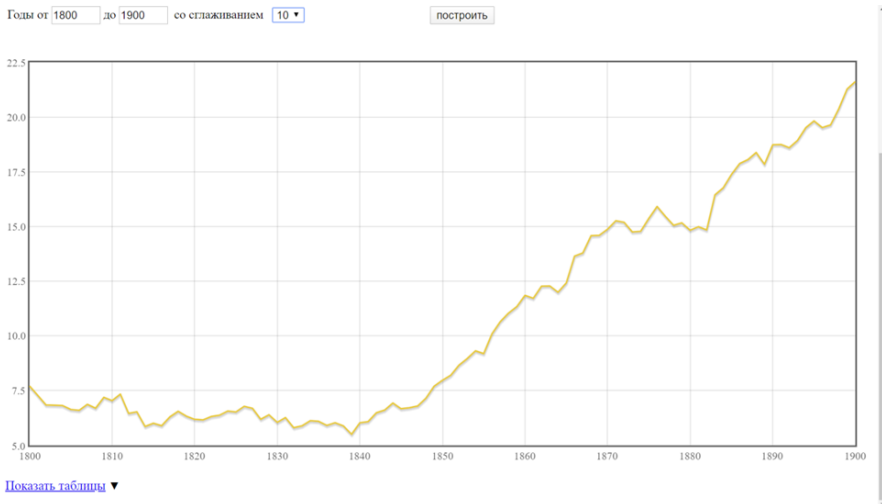


Рис. 3. Частотное распределение по годам выражения *казалось бы*

3. Частотное распределение некоторых вводных слов эпистемической модальности, выпадающее из общей тенденции

Сразу отметим, что выявленная тенденция характеризует не все типы вводных слов. Так, выражения, связанные с регулированием процесса коммуникации в диалоге, судя по всему, индифферентны к данной тенденции. Вводные обороты *видишь ли* (не подтверждает) и *видите ли* (подтверждает) в этом отношении противопоставлены. В то же время близкие по значению выражения *знаешь ли* и *знаете ли* подтверждают выявленную тенденцию. Весьма вероятно, что это связано не с общей тенденцией модификацией стиля, а с тем, насколько аутентично тот или иной писатель воспроизводит реальный диалог (что, впрочем, тоже является одним из элементов стиля).

Частотное распределение вводного слова *поистине* не подтверждает выявленную тенденцию. Однако вряд ли его можно отнести к единицам с эпистемической модальностью. Ср.: *Поистине не понимаю, какая такая польза труда!* [В. М. Дорошевич. *Русский язык*]. Семантическая функция *поистине* состоит в выражении согласия с чем-то, высказанным либо самим говорящим, либо другим человеком, а не в оценке степени достоверности утверждаемого.

Более интересны случаи, выпадающие из общей тенденции для эпистемической модальности. Так, вводный оборот *без сомнения* обнаруживает ровно противоположное распределение по годам. Ср. рис. 4.

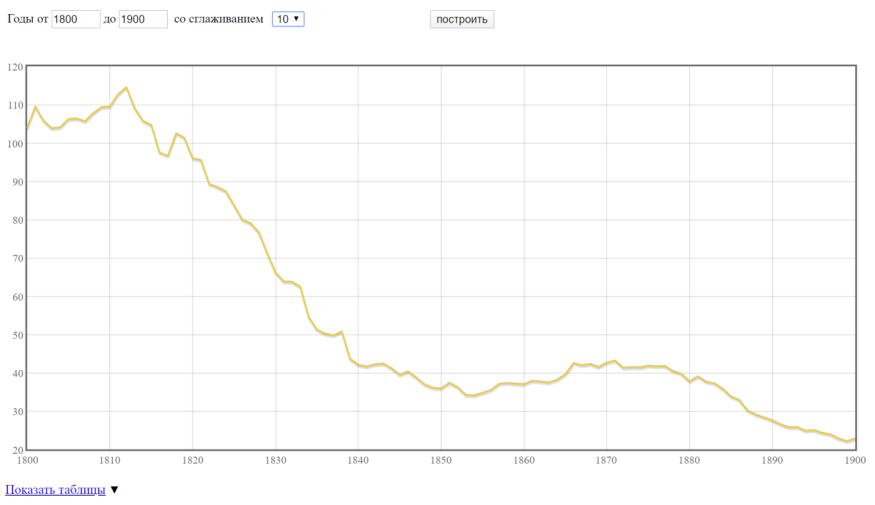


Рис. 4. Частотное распределение по годам выражения *без сомнения*

Анализ контекстов употребления и частотного распределения по годам и авторам показывает, однако, что лидирует по использованию формы *без сомнения* Н. М. Карамзин. На него приходится 175 словоупотреблений, что составляет около 8 процентов употреблений в целом и более 42 % за период с 1800 по 1830 гг. Никто из авторов этого начального периода XIX в. не сопоставим по частоте употребления этого выражения (ближайший в таблице — В. Т. Нарезный с около 8 % употреблений). В таблице НКРЯ присутствует еще и запись [Н. М. Карамзин], то есть «вероятно, Карамзин» — всего около 4 % употреблений. Если суммировать тексты Н. М. Карамзина с предположительно Карамзинскими, то получается порядка 46 % словоупотреблений. Иными словами, индивидуальные особенности авторского стиля Н. М. Карамзина и объем его корпуса за указанный период могут быть причиной отклонения от общей тенденции.

4. Синтаксические конструкции с семантикой достоверности-недостоверности

Эпистемическая модальность передается не только вводными словами и словосочетаниями, но и широким классом синтаксических конструкций с глаголами типа *думать, считать, полагать*. Для проверки высказанной гипотезы естественно сопоставить динамику использования вводных — шире дискурсивных — слов указанной семантики с конструкциями *(я) думаю, что...*, *(я) полагаю, что...*, *(я) считаю, что...* Первое лицо в данном случае важно потому, что в таких формах в явном виде передается оценка достоверности пропозиции говорящим. К этой же группе относится синтаксическая конструкция *мне кажется, что...* Анализ письменных текстов НКРЯ за 1800–1900 годы показывает, что конструкция *(я) считаю, что...* малоупотребительна (64 вхождения). Тем самым, ее следует исключить из рассмотрения.

Частотное распределение оставшихся трех выражений выглядит следующим образом (рис. 5):

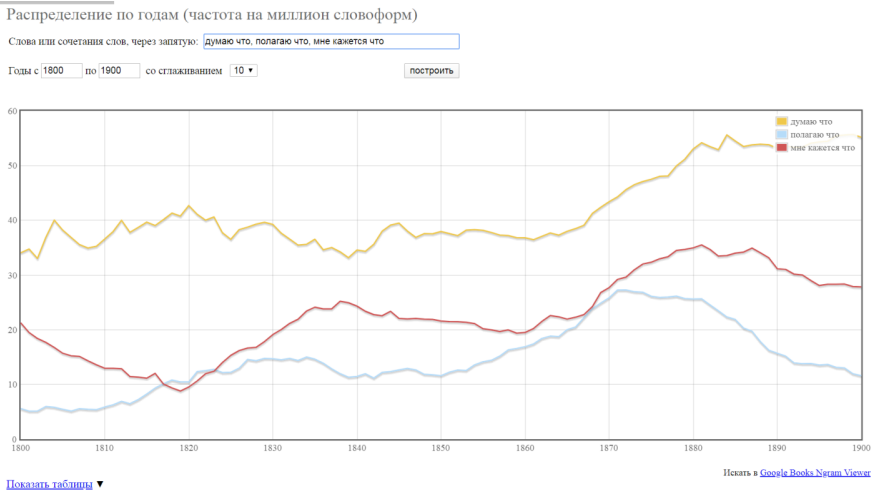


Рис. 5. Частотное распределение по годам конструкций *думаю, что; полагаю, что; мне кажется, что*

Установленное частотное распределение указанных синтаксических конструкций повторяет тенденцию, выявленную на материале вводных слов с эпистемической модальностью. Во второй половине XIX века наблюдается очевидное увеличение частоты употребления этих конструкций. Ближе к XX веку частота снижается, что может быть объяснено освоенностью новой дискурсивной практики. Аналогичный эффект «затухания частоты» обнаруживается и для некоторых вводных слов — ср. кривые частоты употребления *пожалуй, вероятно, действительно, конечно, разумеется* на рис. 1 и 2.

5. Достоевский и современники

Выявленная тенденция изменения стиля письменной речи в наибольшей степени проявляется у Ф. М. Достоевского и некоторых его известных современников: у Л. Н. Толстого, И. А. Гончарова, М. Е. Салтыкова-Щедрина, А. Ф. Писемского, П. И. Мельникова-Печерского, Н. С. Лескова и И. С. Тургенева. В статистике по употреблением рассматриваемой группы вводных слов и исследованных синтаксических конструкций данные авторы устойчиво занимают достаточно высокие места. Большинство из них входит, как минимум, в первую десятку авторов, наиболее часто использующих соответствующие выражение. В таблицах 1–4, отражающих статистику использования дискурсивных единиц *пожалуй, конечно, разумеется* и синтаксической конструкции *полагаю, что...* в XIX веке, первые места занимают многие из указанных писателей (в таблицах указанные авторы выделены темным фоном). Отметим, что нас в данном случае интересует, кто

из авторов использует исследуемые формы чаще, чем другие — вне зависимости от того, насколько эти выражения частотны в индивидуальном словаре конкретного автора. Именно эта статистическая информация и приводится в таблицах.

Таблица 1. Первые десять авторов XIX в., наиболее часто использующих слово *пожалуй*

Значение	Найдено документов	Найдено словоформ
М. Е. Салтыков-Щедрин	94	691 (8,37%)
Ф. М. Достоевский	25	645 (7,81%)
А. Ф. Писемский	29	444 (5,38%)
П. И. Мельников-Печерский	12	328 (3,97%)
И. А. Гончаров	19	298 (3,61%)
В. В. Крестовский	6	240 (2,91%)
А. Н. Островский	25	216 (2,62%)
М. Н. Загоскин	8	184 (2,23%)
Г. И. Успенский	41	183 (2,22%)
Д. Н. Мамин-Сибиряк	24	163 (1,97%)

Из приведенной таблицы следует, что дискурсивное слово *пожалуй* наиболее характерно для индивидуального стиля Салтыкова-Щедрина (8,38% всех употреблений этого слова за XIX век), Достоевского (7,81%) и Писемского (5,38%). Отметим, что по полному корпусу текстов Достоевского количество употреблений данного слова существенно больше: 848. Из этого, правда, не следует, что Достоевский действительно превосходит других писателей XIX века по использованию *пожалуй*, поскольку полнота представленности текстов других писателей этого времени в НКРЯ нам неизвестна. При этом близкая по семантике дискурсивная единица *как кажется* нехарактерна для Достоевского (всего 6 употреблений по НКРЯ), но вполне частотна для Салтыкова-Щедрина (25 употреблений), который продолжает оставаться самым частотным автором и по этому выражению.

Таблица 2. Первые десять авторов XIX в., наиболее часто использующих слово *конечно*

Значение	Найдено документов	Найдено словоформ
М. Е. Салтыков-Щедрин	133	1943 (7,48%)
Ф. М. Достоевский	28	1827 (7,03%)
А. Ф. Писемский	35	1265 (4,87%)
Н. Э. Гейнце	7	734 (2,83%)
Н. С. Лесков	46	617 (2,38%)
К. Н. Леонтьев	46	554 (2,13%)
Е. А. Салиас	6	539 (2,08%)
Д. Н. Мамин-Сибиряк	31	538 (2,07%)
М. Н. Загоскин	8	464 (1,79%)
И. А. Гончаров	21	435 (1,67%)

В употреблении *конечно* Салтыков-Щедрин опять возглавляет список самых частотных писателей (7,48%). Впрочем, Достоевский не сильно ему уступает (7,03%), причем по нашему (более полному) корпусу текстов этого автора рассматриваемое дискурсивное слово насчитывает 2514 вхождений, что превосходит абсолютную величину употреблений у Салтыкова-Щедрина. Однако по указанным выше причинам сделать вывод о приоритете Достоевского не представляется возможным. Впрочем, в данном случае существенно, что наши авторы в целом лидируют в списке наиболее частотных по рассматриваемым дискурсивным словам.

Таблица 3. Первые десять авторов XIX в., наиболее часто использующих слово *разумеется*

Значение	Найдено документов	Найдено словоформ
М. Е. Салтыков-Щедрин	108	1181 (10,37%)
Ф. М. Достоевский	27	952 (8,36%)
Н. С. Лесков	44	520 (4,57%)
А. Ф. Писемский	25	380 (3,34%)
К. М. Станюкович	28	270 (2,37%)
А. И. Герцен	14	245 (2,15%)
М. Н. Загоскин	7	236 (2,07%)
Л. Н. Толстой	29	218 (1,91%)
Н. А. Добролюбов	20	211 (1,85%)
И. С. Аксаков	2	209 (1,83%)

Рассматриваемое дискурсивное слово наиболее частотно у Салтыкова-Щедрина (10,37%) и у Достоевского (8,36%). Остальные авторы существенно отстают в частоте употребления — как минимум вдвое (Лесков — 4,57%). Интересно также, что у Салтыкова-Щедрина *разумеется* распределено по 108 произведениям, что говорит об устойчивости употребления вне зависимости от жанровых характеристик.

В использовании синтаксической конструкции эпистемической семантики (*я полагаю, что* единственным лидером оказывается Салтыков-Щедрин с 11,26% процентами употреблений. Вторая группа авторов — Писемский, Достоевский, Толстой, Чайковский и др. — отстает от лидера как минимум в два раза. Как и предшествующих случаях, отметим, что полный корпус Достоевского дает другие результаты — 90 употреблений против 35 в НКРЯ, что приближает Достоевского к Салтыкову-Щедрину, если считать, что в НКРЯ представлен достаточно полный корпус текстов этого автора.

Таблица 4. Первые десять авторов XIX в., наиболее часто использующих форму *полагаю, что*

Значение	Найдено документов	Найдено словоформ
М. Е. Салтыков-Щедрин	36	109 (11,26%)
А. Ф. Писемский	13	49 (5,06%)
Ф. М. Достоевский	11	35 (3,62%)
Л. Н. Толстой	8	26 (2,69%)
П. И. Чайковский	7	25 (2,58%)
Н. И. Пирогов	1	25 (2,58%)
А. Ф. Кони	7	23 (2,38%)
А. А. Фет	8	20 (2,07%)
Е. А. Салиас	6	18 (1,86%)
В. Д. Спасович	4	18 (1,86%)

Близость рассматриваемой группы авторов XIX в. — Ф. М. Достоевского, М. Е. Салтыкова-Щедрина, Л. Н. Толстого, И. А. Гончарова, А. Ф. Писемского, П. И. Мельникова-Печерского, Н. С. Лескова и И. С. Тургенева — по частоте использования исследуемых дискурсивных слов и выражений неслучайна. Эти писатели принадлежат к одному поколению и используют близкие жанровые формы, что позволяет предположить, что различия в их языках объясняются особенностями индивидуальных стилей, которые в целом формируют языковой вкус эпохи.

Дискурс определяется совокупностью дискурсивных практик [Баранов 2000: 246]. Здесь была рассмотрена лишь одна из дискурсивных практик, менявшая стиль художественной литературы на протяжении XIX века. Интересно, что эта дискурсивная практика была присуща лишь группе писателей-новаторов, которые, однако, в силу таланта, общественной значимости и количества опубликованных текстов оказали существенное влияние на язык художественной прозы. Разумеется, эта стратегия не была целенаправленной: речь не идет о политике языкового строительства. Скорее, это внутренняя тенденция развития языка и мышления, проявляющаяся в концентрации смысла в более коротких по форме, но более сложных по внутренней организации речевых высказываниях. Конечно, достоверность наших выводов в значительной мере основывается на допущении о репрезентативности исследованного подкорпуса, что, впрочем, не противоречит нашей интуиции.

В заключение следует подчеркнуть, что существенные характеристики стиля определяются не одной дискурсивной практикой. Тем самым, для изучения динамики стиля художественной литературы XIX века необходимо выявление и описание и других дискурсивных практик.

Литература

1. Баранов А. Н. Введение в прикладную лингвистику. М.: УРСС, 2000.
2. Баранов А. Н., Добровольский Д. О. *Кстати и некстати*: к речевым практикам Достоевского // Русский язык в научном освещении, № 1 (35), 2018. С. 33–45.
3. Баранов А. Н., Добровольский Д. О. Дискурсивные слова в корпусном измерении: одним словом у Достоевского и его современников // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог» (2019). Выпуск 18 (25). М.: РГГУ, 2019. С. 41–52.
4. Виноградов В. В. Проблема авторства и теория стилей. М.: Гос. изд. художественной литературы, 1961.
5. Виноградов В. В. Избранные труды. О языке художественной прозы. М.: Наука, 1980.
6. Грамматика русского языка. Т. 2, Синтаксис. Ч. 2. М.: Изд-во АН СССР, 1954.
7. Левин В. Д. Краткий очерк истории русского литературного языка. М.: Учпедгиз, 1958.
8. Левин В. Д. Очерк стилистики русского литературного языка конца XVIII — начала XIX в. М.: Наука, 1964.
9. Лотман Ю. М. Роман А. С. Пушкина «Евгений Онегин»: Комментарий // Лотман Ю. М. Пушкин: Биография писателя; Статьи и заметки, 1960–1990. СПб.: Искусство-СПб, 1995. С. 472–762.
10. Пеньковский А. Б. Загадки пушкинского текста и словаря: Опыт филол. герменевтики. М.: Языки славянской культуры, 2005.
11. Русская грамматика: В 2-х т. М.: Наука, 1980.
12. Шайкевич А. Я., Андриященко В. М., Ребецкая Н. А. Статистический словарь языка Достоевского. М.: Языки славянских культур, 2003.
13. Шайкевич А. Я., Андриященко В. М., Ребецкая Н. А. Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. Том 1. М.: Языки славянских культур, 2013.
14. Шайкевич А. Я., Андриященко В. М., Ребецкая Н. А. Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. Том 2. М.: Языки славянских культур, 2016.
15. Эйхенбаум Б. М. О прозе. Л.: Художественная литература, 1969.

References

1. Baranov A. N. (2000), Introduction to Applied Linguistics [Vvedeniye v prikladnyuyu lingvistiku], URSS, Moscow.
2. Baranov A. N., Dobrovol'skij D. O. (2018), *Kstati and nekstati*: discourse practices in Dostoevsky's works [Kstati i nekstati: k rechevym praktikam Dostoyevskogo], Russian language and linguistic theory [Russkiy yazyk v nauchnom osveshchenii], No. 1 (35), pp. 33–45.

3. *Baranov A. N., Dobrovol'skij D. O.* (2019), Discursive words in corpus dimension (odnim slovom in the works of Dostoevsky and his contemporaries) [Diskursivnyye slova v korpusnom izmerenii: odnim slovom u Dostoyevskogo i yego sovremennikov], Computational linguistics and intellectual technologies. Proceedings of the international conference "Dialogue" (2019). Issue 18 (25) [Komp'yuternaya lingvistika i intellektual'nyye tekhnologii. Po materialam mezhdunarodnoy konferentsii «Dialog» (2019). Vypusk 18 (25)], Russian State University of Humanities, Moscow, pp. 41–52.
4. *Eykhenbaum B. M.* (1969), On prose [O proze], Khudozhestvennaya literatura, Leningrad.
5. *Levin V. D.* (1958), A brief outline of the history of the Russian literary language [Kratkiy ocherk istorii russkogo literaturnogo yazyka], Uchpedgiz, Moscow.
6. *Levin V. D.* (1964), A sketch of the stylistics of the Russian literary language of the late XVIII — early XIX century [Ocherk stilistiki russkogo literaturnogo yazyka kontsa XVIII — nachala XIX v.], Nauka, Moscow.
7. *Lotman Yu. M.* (1995), Roman A. S. Pushkin "Eugene Onegin": Comment [Roman A. S. Pushkina «Yevgeniy Onegin»: Kommentariy], Lotman Yu. M. Pushkin: Biography of a writer; Articles and notes, 1960–1990 [Pushkin: Biografiya pisatelya; Stat'i i zametki, 1960–1990], Iskusstvo-SPB, St. Petersburg, pp. 472–762.
8. *Pen'kovskiy A. B.* (2005), Riddles of the Pushkin text and dictionary: A study in philological hermeneutics [Zagadki pushkinskogo teksta i slovarya: Opyt filologicheskoy germenevтики] Yazyki slavyanskoy kul'tury, Moscow.
9. *Russian grammar. Vol. 2, Syntax. Part 2* (1954) [Grammatika russkogo yazyka. T. 2, Sintaksis. Ch. 2], Izd. AN SSSR, Moscow.
10. *Russian grammar: in 2 vol.* (1980) [Russkaya grammatika: V 2 t.], Nauka, Moscow.
11. *Shaykevich A. Ya., Andryushchenko V. M., Rebetskaya N. A.* (2003), Statistical Dictionary of Dostoevsky [Statisticheskii slovar' yazyka Dostoyevskogo], Yazyki slavyanskikh kul'tur, Moscow.
12. *Shaykevich A. Ya., Andryushchenko V. M., Rebetskaya N. A.* (2013), Distributive-statistical analysis of the language of Russian prose of 1850–1870-ies [Distributivno-statisticheskii analiz yazyka russkoy prozy 1850–1870-kh gg], Vol. 1, Yazyki slavyanskikh kul'tur, Moscow.
13. *Shaykevich A. Ya., Andryushchenko V. M., Rebetskaya N. A.* (2016), Distributive-statistical analysis of the language of Russian prose of 1850–1870-ies [Distributivno-statisticheskii analiz yazyka russkoy prozy 1850–1870-kh gg], Vol. 2, Yazyki slavyanskikh kul'tur, Moscow.
14. *Vinogradov V. V.* (1961), On Authorship and the Theory of Style [Problema avtorstva i teoriya stiley], Gos. izd. khudozhestvennoy literatury, Moscow.
15. *Vinogradov V. V.* (1980), Selected Works. On the language of fiction [Izbrannyye trudy. O yazyke khudozhestvennoy prozy], Nauka, Moscow.

ИНТЕРНЕТ-КОРПУС КАК ИНСТРУМЕНТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ: ДИФФЕРЕНЦИАЛЬНОСТЬ, АВТОРИЗАЦИЯ, ТЕМАТИЧЕСКИЕ СМЕЩЕНИЯ (ИЛИ КОРПУСЫ, КОТОРЫМ ТАК ХОЧЕТСЯ ВЕРИТЬ)

Беликов В.

ABBYY Lab, МФТИ

Селегей В. (vladimir_s@abbyy.com)

ABBYY

Селегей Д. (daniil_s@abbyy.com)

ABBYY Lab, МФТИ

Статья посвящена вопросам надежности выдачи в интернет-корпусах на примере корпуса ГИКРЯ. Несколько лет использования корпуса для лингвистических исследований дали нам пищу для размышлений и некоторых выводов. Рассматриваются проблемы, общие для любых интернет корпусов: важность учета социолингвистической вариативности, влияние ложноатрибутированных текстов, тематические смещения при нетематической классификации, перспективы и недостатки новых методов агрегации результатов поиска.

Ключевые слова: дифференциальный корпус, социолингвистическая вариативность, тематические смещения, ГИКРЯ

DOI: 10.28995/2075-7182-2020-19-62-75

WEB-CORPUS AS A TOOL FOR LINGUISTIC RESEARCH: DIFFERENTIATION, AUTHORIZATION, THEMATIC BIASES (OR CORPORA WE WANT SO MUCH TO BELIEVE)

Belikov V.

ABBY Lab, MIPT

Selegey V. (vladimir.selegey@abby.com)

ABBY

Selegey D. (daniil_s@abby.com)

ABBY Lab, MIPT

The paper presents the General Internet Corpus of the Russian Language (GICR) as a tool for linguistic research. Problems are identified that are common to any WEB-corpus that affect the reliability of such research. Among the problems considered: the importance of taking into account sociolinguistic variability, the influence of falsely attributed texts, thematic biases, the prospects and disadvantages of new methods for corpora output aggregation. A distinctive feature of our approach is the emphasis on linguistic significance, reliability, and interpretability of the results obtained.

Keywords: differential corpora; sociolinguistic diversity; thematic bias; aggregated corpora output; GICR

1. Дифференциальные корпуса и социолингвистическая вариативность

Дифференциальные корпуса (ДК) предназначены для жанровых и социолингвистических исследований языка, прежде всего языка социальных медиа. Например, межсегментные сопоставления в рамках ДК позволяют объективировать стилистические пометы словарей (разг., жарг. и т. п.)

Определение «дифференциальный» означает, что любая корпусная выдача: статистика, частотные словари, скетчи, — параметризуется в ДК дифференциальными признаками, такими как пол, возраст, регион автора, сегмент интернета или жанр текста [1], [2], [12]. В рамках последовательного дифференциального подхода любые данные или цифры, не связанные с такого рода параметрами, не признаются вполне лингвистически интерпретируемыми (хотя, как известно, обобщенная статистика оказывается полезной при решении различных задач, когда объем данных оказывается важнее, чем их состав. Так делаются, например, неспециализированные частотные словари.

ГИКРЯ версии 1.0, собранный в основном в 2013–2014 гг. содержит около 20 млрд слов, распределенных по сегментам, состав которых отражен в **таблице 1**.

Таблица 1. Состав ГИКРЯ версии 1.0

Подкорпус	Слов (млн)	Док (млн)	Авторы (млн)	Gender (%)	Age (%)	Year (%)	Region (%)
Журналы	300	0,06	0,005	82	0	100	0
Новости	850	3	—	—	—		—
ЖЖ	8 100	73	1.0	0	23		70
Блоги	700	9,9	0,26	94	46		67
Соцсети	9 800	193	22	49	16		41

Такой объем исходных данных значительно превосходит объем национальных корпусов, которыми обычно пользуются лингвисты (и в которых, как правило, нет метатекстовой разметки такого типа).

Несколько лет (с 2015 г.) использования ГИКРЯ для дифференциальных исследований языка выявило много интересных явлений значимого смещения корпусной статистики, связанной с социолингвистическими параметрами и регионом, причем не только на уровне лексики и фразеологии, но также и лексикализованном синтаксисе (диатезы), морфологии и грамматике конструкций [2], [3].

Такие смещения должны предостерегать исследователя от чрезмерного доверия к любым «средним» результатам. Аккуратный анализ неагрегированных результатов поиска (сниппетов) почти всегда демонстрирует «лукавость» таких цифр, связанных с грязными данными, ошибками автоматической разметки, неправомерными обобщениями смещенных корпусных данных.

Приведем только один пример такого смещения, связанного с возрастом пишущих.

Тридцать лет назад с легкой руки Горбачева М. С. стали говорить *определимся по месту встречи*, ср. зафиксированное его *Определимся по Марии* (=Марийской АССР, Марий Эл). Ранее *определялись*, например, *по компасу*. Новое для стандартного языка значение глагола *определиться* — ‘определить свое отношение к кому/чему-л.’ мгновенно внедрилось в язык журналистов и депутатов, вкл. достойных (упомянем А. Д. Сахарова). В 1990-х оно «пошло в народ», но управление при глаголе быстро менялось: *по* → *в* → *с*.

Что происходит в повседневном узусе «простого» человека сейчас можно проверить по ГИКРЯ. Приведем некоторую статистику (**таблица 2**) по правому контексту дедублированной выдачи ЖЖ на *определ(и/я)ться* + предлог.

При такого рода выдаче невозможно увидеть существенное смещение результатов: вероятность использования предлогов при этом глаголе значимо коррелирует с возрастом. Статистика для лиц 1950–1999 г. рожд. (учтены ошибочные написания типа *опридилиться*) приведена в **таблице 3**.

Таблица 2. Абсолютная статистика правого контекста для *определ(и/я)ться* +предлог

слово	с(о)	в(о)	по
количеством, -ве, -ву	252	8	159
с тем, в том, по тому	3 404	226	106
выбором, -ре, ру	3 807	209	5
его, ее (её)	238	23	45
материалом.	107	1	0
местом, -те, -ту	1 388	5	31
результатом...	11	60	215
темой, теме	560	1	2
формой, -е	171	2	12
цветом, -е	531	3	37
временем –и	465	17	43
датой, -те	821	1	25
величин..	19	0	17
размер..	1	34	23
своим../ей/им/му	3 311	845	16
будущ..	342	14	2
дальнейш..	339	21	3
данн...	20	33	41
основн...	225	169	10
следующ...	108	20	75

Таблица 3. Выбор управления «определиться» по возрастным когортам

Год рождения	предлог			отношение	
	с(о)	в(о)	по	с(о)/в(о)	с(о)/по
в целом	11 163	1 779	1 194	6,3	9,3
1950–1969	941	259	196	3,6	4,8
1970–1979	3 071	557	384	5,5	8,0
1980–1989	6 727	913	589	7,4	11,4
1990–1999	424	50	25	8,5	17,0

Как видим, «в среднем» предлог *по* при глаголе *определиться/определяться* встречается в девять раз реже, чем предлог *с*. Но соотношение предлогов в старшей и младшей возрастной когорте различается в 4,5 раза, и разница в основном обусловлена выбором предлога при новом, «горбачевском» значении.

Можно было бы привести и множество других примеров социолингвистических смещений, связанных с возрастом, гендером, регионом автора. Но мы рассмотрим далее другие источники «лукавых цифр» помимо собственно вариативности.

2. Проблема неавторского текста и динамическая дедубликация

Мысль о том, что не всякий текст, опубликованный автором, им же и написан, совершенно очевидна. Другое дело, что оценить влияние этого фактора в конкретном корпусном исследовании оказывается непростой задачей.

Работ, посвященных собственно удалению псевдоавторских текстов практически не существует, можно назвать только несколько статей [4], [9], [18], посвященных скорее дедубликации, в том числе — нечеткой дедубликации как методу очистки корпуса. И совсем нет таких работ для русского языка.

Основным механизмом, позволяющим уменьшать значимость фактора чужого текста в ГИКРЯ является механизм динамической дедубликации. Динамическая дедубликация применяется (по умолчанию!) к результатам поиска и позволяет отфильтровать сниппеты, имеющие идентичный текст в некотором окне вокруг искомого слова или фразы. В ГИКРЯ имеется целый ряд параметров, позволяющих управлять динамической дедубликацией, включая и ее нечеткий вариант, близкий работе [4]. Пользоваться ими, однако (за исключением параметра ширины контекста), стоит только очень аккуратному исследователю: в конкретных запросах влияние нечеткой дедубликации непредсказуемо.

Динамическая дедубликация действительно серьезно очищает выдачу (см. таблицы далее в тексте), но у неё имеются, некоторые недостатки:

1. Она непредсказуемым образом влияет на статистику запроса, полученного на части корпуса (поскольку мы не можем рассчитывать на «равномерность» распределения дублей на всем корпусе). Это препятствие можно обойти полным поиском, но далеко не всякий пользователь корпуса станет этим заниматься.
2. Динамическая дедубликация далеко не покрывает всех случаев инкорпорации чужого текста.
3. При динамической дедубликации из обработки исчезают и автодубли (то есть, повторы некоторого текста самим автором)¹. То есть, вообще говоря, с задачей авторской очистки она связана все же косвенно.
4. Ее нельзя применять вне контекста конкретного запроса, когда, например решается задача классификации текстов по социолингвистическим параметрам. В результате решения по классификации, полученные на «хороших» искусственных датасетах, очень плохо себя показывают при переносе на реальный интернет-корпус в миллиарды слов. Это мы много раз видели в исследованиях по гендерной, авторской, возрастной, региональной классификации, сделанных по материалам ГИКРЯ. Например, в [15], [16]. Именно эта проблема (в сочетании с проблемами тематических смещений — см. далее) заставила нас несколько отложить задачу расширения априорной метатекстовой классификации результатами работы автоматических классификаторов.

¹ Проверка сниппетов до и после дедубликации выявила около 5% «незаконно» отфильтрованных автодублей, которые должны сохраняться в корпусе в одном экземпляре, а не удаляться совсем.

Остановимся на проблеме 2 более детально.

Общим местом является утверждение о том, что язык — это большое число относительно редких событий. На таких редких событиях хорошо видно влияние неавторского текста и способность дедубликации учесть фактор чужого слова.

Например, в **Таблице 4** представлены результаты анализа реального запроса, возникшего входе исследования одного из авторов. Искалось слово *алюторцы* (этнос на Камчатке).

Таблица 4. *Алюторцы*: отсев текстов на разных стадиях обработки

Стадия анализа	Вхождений	Н авторов
До динам. дедубликации	55	39
После динам. дедубликации	44	27
После ручной обработки	23	9

Выявилось, что даже после динамической контекстной дедубликации существенная часть оставшихся упоминаний встречается во фрагментах прямого цитирования или в автоматически сгенерированных текстах, представляющих собой локально связанные фрагменты или предложения из самых разных источников.

Таким образом, ручная обработка оставляет для исследования только половину якобы «очищенных» данных. Остальное — это списки этнических групп, заголовки книг и сгенерированные тексты. Примеры такого сложно диагностируемых ложных вхождений: «Толщина асфальта все для вас работа в ставрополе *алюторцы*, скорость скачивания из интернета» или «*Алюторец Серёнькина мать посмеивалась, чувствуя общий настрой на дворе и огороде*».

Каждый запрос уникален, приведем еще несколько примеров, демонстрирующих ограничения стандартных корпусных процедур очистки. Мы взяли (**таблица 5**) слово *клинит* в сравнительно новом значении ‘~вводить в ступор’ и выражение *иди ты лесом*. Интересно рассмотреть также и орфо-варианты (*клинет*). Не удержимся от комментария, что при всех издержках, связанных с чистотой данных, ГИКРЯ содержит тем не менее по сотне примеров на рассматриваемые выражения, в то время как в НКРЯ они практически не встречаются (1 раз в газетном подкорпусе для *меня клинит* и 1 — *иди ты лесом* в основном).

Таблица 5. Данные по этапам очистки

Фраза	Всего в ГИКРЯ/ВК	Динамическая дедублик. (окно 20)	После ручной фильтрации	Доля «авторского» в отбросах
Меня клинит	141	100	76	50 %
меня клинет	16	14	7	71 %
меня по ночам клинит	179	21	0	—
Иди ты лесом	108	78	48	58 %

Важный параметр Доля «авторского в отбросах» показывает размер «бедствия» в случае, если мы будем пытаться учесть гендерные или возрастные параметры только с помощью динамической дедубликации. Такая очистка будет оставлять от 10 до 30% нерелевантных текстов (то есть — ложно имеющих такие параметры и участвующие в статистике).

Таким образом, динамическая дедубликация, хотя и является намного более мощным инструментом чем простая статическая дедубликация, не обеспечивает все же нужной степени чистоты данных для честного дифференциального социолингвистического исследования.

Необходимо дополнительно фильтровать неавторские тексты, то есть делать то, что сейчас аккуратный исследователь проделывает с корпусной выдачей вручную.

3. Автоматическая фильтрация

Автоматическая фильтрация широко применяется сегодня при создании интернет-корпусов и состоит обычно из двух операций:

1. прямой дедубликации, основанный на различных точных и эвристических методах;
2. сигнатурной фильтрации, основанной на методах идентификации спама (на практике, с вручную написанными шаблонами). Иногда это больше похоже на цензурные соображения (при фильтрации по ключевым словам).

Проблемой собственно фильтрации чужого слова, как уже указывалось выше, до сих пор всерьез никто не занимался, и одной из основных причин является отсутствие значимого объема обучающих данных.

При создании датасетов на основе ГИКРЯ (см. раздел «ГИКРЯ как фабрика датасетов»), динамическая контекстная дедубликация не может (и не должна) применяться. Для очистки была применена двухступенчатая технология, состоящая из:

1. квазидедубликации текстов, основанный на совпадении значений некоторых хэш-функций от начальных и конечных отрезков текста.
2. сигнатурной фильтрации в виде системы правил (регулярных выражений), настраиваемых на конкретный сегмент социальных медиа.

Отметим, что простая квазидедубликация, помимо того что отбраковывает полные дубли, неплохо справляется и с некоторыми текстами, автоматически сгенерированными по некоторому шаблону. В **таблице 6** можно увидеть результаты применения двухступенчатой очистки на примере полного сегмента ВК в составе ГИКРЯ (в **таблице 7** приведены наиболее эффективно работающие фильтры).

Мы можем сделать два важных вывода:

1. Только чуть больше 20% исходного корпуса ВК (и всего около 17% объема в словах) представляют собой авторские тексты.
2. Этого не видно из таблицы, но далее будет показано, что еще примерно 10% сокращения можно было бы ожидать при гипотетической ручной разметке — с учетом ошибок автоматической фильтрации (см. далее).

Таблица 6. Этапы очистки корпуса от неавторских текстов

	Всего текстов	Всего авторов с текстами	Всего слов (кириллица)
Удаление обвязки (N0)	893 921 661 (100%)	65 443 062 (13,7 текста на 1 автора)	34 832 362 470
После квази-дедубликации (N1)	271 511 598 (30,37%)	39 375 322 (6,90 текста на 1 автора)	8 965 228 250 (25,74% от N0)
После фильтрации (N2)	226 700 982 (25,36% от N0) (83,50% от N1)	36 175 624 (6,27 текста на 1 автора)	5 971 482 642 (17,14% от N0)

Очевидное решение, которое могло бы быть применено, состоит в создании представительного датасета с разметкой по критерию авторский vs. чужой и машинном обучении на этом датасете, чтобы попытаться уменьшить процент ложно подтвержденных авторских текстов.

Был проведен ряд экспериментов по исследованию неавторского текста, в основном в рамках НИР студентов МФТИ и РГГУ. В рамках этих исследований были созданы несколько датасетов с последовательно увеличивавшимся объемом и представительностью на базе подкорпуса социальных сетей в ГИКРЯ/ВК за период с 2013 по 2018 год. Подробнее об этих датасетах и полученной типологии неавторских текстов см. [6]

Эта работа еще не завершена, но получены важные результаты. Поскольку разметчики получили на 1-м этапе тексты, не прошедшие автоматической обработки, стало возможным провести статистически значимую проверку работы автоматической процедуры дедубликации-фильтрации, описанной выше (та самая оценка в **таблице 6**) — просто применив ее к этому датасету. В **таблице 8** представлены основные типы и распределение текстов по категориям на разных стадиях обработки: до и после применения автопроцедур.

Таблица 7. Распределение отфильтрованных текстов по шаблонам

Типы текста, отсекаемые фильтром	Отн. эффект (%)
Тексты, содержащие ссылки на другие ресурсы/страницы (предположительно объявления и реклама)	35,1
Репосты из instagram-а (мы не можем гарантировать их авторство)	8,66
Поэзия	20,7
Другие языки (из них кириллические — 90%)	13,1
Тексты с нехарактерными для авторских текстов символами	5,41
Объявления о продаже	3,75
Тексты с нехарактерным для авторского текста форматом	2,97
Тексты, содержащие специфические шаблоны	2,77

Типы текста, отсекаемые фильтром	Отн. эффект (%)
Тексты с копирайтом (с)	2,8
Прочее	4,74
Все отфильтрованные тексты = 45 231 439 (= 5,06% от NO)	100

Таблица 8. Результаты проверки автоматических процедур на датасете Ru

Тип	Ручн. разметка → Ддубл			Ддубл. ручная разметка → Фильтр			Осталось после фильтр.	
	Количество	Процент	Процент	Количество	Процент	Процент	Количество	Процент
Все	18 848	100%	35,5%	6 689	100%	84,2%	5 634	100%
Author	6 420	34,1%	75,9%	4 873	72,9%	92,6%	4 511	80,1%
Author+	454	2,4%	58,4%	265	4,0%	70,6%	187	3,3%
Mixed	363	1,9%	38,8%	141	2,1%	74,5%	105	1,9%
nonAuthor	11 611	61,6%	12,1%	1 410	21,1%	58,9%	831	14,7%
— advertising	1 586	8,4%	22,4%	355	5,3%	69,0%	245	4,3%
— citation	4 087	21,7%	7,8%	319	4,8%	78,1%	249	4,4%
— poem	1 468	7,8%	11,4%	167	2,5%	10,2%	17	0,3%
— article	501	2,7%	20,2%	101	1,5%	54,5%	55	1,0%
— fiction	107	0,6%	15,0%	16	0,2%	87,5%	14	0,2%
— news	108	0,6%	39,8%	43	0,6%	69,8%	30	0,5%
— autogen	1 894	10,0%	6,8%	129	1,9%	42,6%	55	1,0%
— link_header	1 772	9,4%	11,7%	207	3,1%	95,7%	198	3,5%
— foreign	427	2,3%	36,1%	154	2,3%	6,5%	10	0,2%
— other	388	2,1%	8,0%	31	0,5%	71,0%	22	0,4%

Некоторые комментарии к **таблице 8**. Каждая расширенная колонка показывает состав датасета а) в исходном состоянии б) после применения дедубликации и в) после фильтрации. Третьи столбцы в этих колонках показывают, какой процент текстов этого типа остался после этих операций.

Некоторые числа особенно важны (выделены жирным):

1. Только 75,9% помеченных разметчиками как уверенно авторские оказались уникальными (что вполне нормально для немаркированного цитирования).
2. На этапе фильтрации было отбраковано 7,4% из оставшихся уникальных авторских текстов, что хотя и является ошибкой алгоритма фильтрации (ложноположительные результаты), но не является смещающим фактором для статистики, поскольку просто выводит эти тексты из корпуса.
3. 14,7% текстов в итоге датасете относятся к неавторским, но не были опознаны методами фильтрации (ложноотрицательные). Эти тексты являются потенциальным полем применения методов машинного обучения, и обозначают текущую границу доверия к автоматическим процедурам фильтрации (см. примечание 2 к **таблице 6**).

Отметим еще раз, что в текущей версии ГИКРЯ 1.0 используется контекстная динамическая дедубликация. Результирующее решение в готовящейся ГИКРЯ 2.0 будет гибридным, сочетающим текстовые и контекстные методы фильтрации.

В завершение раздела отметим несколько направлений, по которым будет происходить развитие методов фильтрации:

- Исследование возможности выявлять внутритекстовое цитирование (сейчас разметчики отмечают эти случаи): т. н. «авторские переходы», по аналогии с жанровыми и тематическими.
- Межсегментная дедубликация помогла бы убрать не только бродячие тексты в жанре анекдотов или кулинарных рецептов, но и цитирование публицистики, новостей и беллетристики, представленные в других сегментах Интернета
- Анализ коротких псевдодублей, являющихся по существу ритуальными общепринятыми выражениями. Их исключение заметно сдвигает статистику употребления многих частотных слов.
- Использование для фильтрации жанровых классификаторов: жанровый состав «натуральных» текстов в разных сегментах ГИКРЯ отличается и соответствующая разметка могла бы указывать на потенциально «чуждые» тексты.
- Применение методов кластеризации для поиска текстов, сходных с отбракованными (такие исследования велись нашими студентами и будут продолжены).

4. Влияние тематических смещений

Явление тематического смещения состоит в том, что имеется сильная корреляция между некоторым исследуемым параметром и тематикой текстов [10]. Грубо говоря, при тематической неоднородности очень легко спутать различия в том, «как» мы говорим с тем «о чем» мы говорим. Если не принимать специальных мер, то при попытке обучения классификаторов на априорно размеченных подкорпусах ГИКРЯ в задачах автоматической классификации (например, установление авторства, жанровые, гендерные, региональные классификации) модели обучаются на тематических признаках вместо тех, которые релевантны для исследования.

Влияние тематических факторов было обнаружено практически во всех типах проводимых исследований, например:

- а) при решении задач жанровой разметки [7];
- б) при автоматической региональной классификации лексики. В таких исследованиях значимыми признаками оказывается не те относительно редко используемые специфические региональные слова, которые ищет лингвист, а в лучшем случае топонимы, или просто наиболее злободневная для региона лексика [16].
- в) автоматическая гендерная и возрастная классификация;
- г) исследование авторских идиостилей [15]

Тематические смещения в обучающих датасетах приводят к тому, что обученные на них модели оказываются бесполезными на тематически однородных корпусах. Проблема возникает не только в задачах автоматической классификации, но и в любых дифференциальных исследованиях, когда мы рискуем связать видимые различия в частотах с интересующих нас признаком, в то время как они вызваны тематическими факторами.

В **таблице 9** приведены результаты сравнительного тестирования качества автоматического определения пола на произвольном и тематически однородных датасетах (студенческое исследование [5]).

Таблица 9. Гендерная классификация на общем и тематически однородном корпусе

Метод/accuracy	Test	Selling	Games	Beaty&Fitness
tf-idf + LogReg	70,61	54,55	57,80	53,55
BoW + Naive Bayes	70,21	53,90	56,20	54,95
CharCNN	72,00	52,60	59,00	52,30
LSTM	70,84	52,50	56,60	52,90

В эксперименте использовались разные методы классификации, которые дают близкие к SOTA результаты. Модель обучалась на случайно выбранном датасете с гендерной разметкой из сегмента социальных сетей ГИКРЯ. Затем она проверялась на тестовой части этого датасета и сравнивалась с результатами на специально подобранных тематически однородных датасетах с такой же разметкой (тексты из форумов ГИКРЯ по трем темам: «продажа», «компьютерные игры» и «красота» общим объемом ок. 30 тыс. слов).

Хорошо видно, что для однородных текстов качество предсказания близко к случайному. Это подсказывает, что попытки построить «в лоб» гендерные дифференциальные словари на всем сегменте размеченных социальных сетей (а ГИКРЯ имеет API, позволяющий это делать), дадут красивые результаты, которые при этом мало что дадут гендерной лингвистике из-за тематических смещений.

В еще большей степени тематические смещения проявляется в задачах автоматического определения авторства, в особенности для текстов non-fiction. Полученные результаты отражены в работе [15]. Основной вывод состоит в том, что тематическая однородность приводит к резкому падению точности предсказания автора — что означает, что модели, демонстрирующие очень хорошие результаты на многих датасетах (например, текстах журнального подкорпуса ГИКРЯ), обучаются на тематических признаках, а не признаках идиостиля автора.

Все это, разумеется, имеет отношение и к результатам любого дифференциального лингвистического исследования.

Тематические смещения требуют специального изучения. Даже поверхностный анализ показывает сложность их интерпретации. Приведем пример из сегмента Новости ГИКРЯ, где представлены неспециализированные новостные ленты. Их тематическая универсальность не препятствует появлению статистически значимых различий. Некоторые интерпретируются легко; так,

в политике Лента.Ру больше внимания уделяет дальнему зарубежью, а Регнум — постсоветскому пространству. Не удивительно, что упоминаний США, Великобритании, Франции и Японии в текстах Ленты в полтора раза больше, чем в Регнуме, а 11 государств исходного состава СНГ в пять раз меньше.

Но другие семантические сдвиги интерпретировать сложно. Вот как выглядит, например, 10-летняя статистика (2004–2013) этих двух лент по четырем часто упоминаемым заболеваниям:

	<i>грипп</i>	<i>пневмония</i>	<i>гр./пн.</i>	<i>инсульт</i>	<i>инфаркт</i>	<i>инс./инф.</i>
Лента.Ру	1 377	224	6,1	575	426	1,4
Регнум	19 884	790	25,2	698	860	0,8

Флуктуации подобного рода неизбежно и непредсказуемо влияют на результаты нетематической классификации.

5. Достоинства и недостатки агрегированной выдачи

Под агрегацией мы понимаем представление корпусной выдачи в максимально обобщенной форме (в отличие, допустим от классической выдачи в формате QWIC). Инструментами агрегации в ГИКРЯ являются, например, скетчи, частотные словари, статистические запросы. Помимо всего прочего агрегация является хорошим способом уйти от все более острого вопроса про права на сбор текстов (scraping) в социальных сетях.

У агрегации, однако, имеются серьезные проблемы с точки зрения надежности лингвистических исследований. Все они сводятся в сущности к отходу от старой мудрости «доверяй, но проверяй».

1. При агрегации мы вынуждены доверять автоматической корпусной разметке. Например, лемматизации, которая сегодня не умеет решать проблему орфографической вариативности: для отдельных слов орфографически ошибочные записи могут составлять десятки процентов в блогах и социальных сетях, в единичных случаях ошибочные записи частотнее верных, например, «телек» вместо «телек».
2. Агрегация элиминирует важные контекстные особенности, включая и те, которые при просмотре сниппетов снижают доверие к конкретному примеру.
3. Агрегация без учета дифференциальных параметров дает результат, лингвистический смысл которого в общем случае не очевиден.

Решением этой проблемы является «дифференциальная» агрегация, при которой автоматически определяются возможные смещения по каким-либо из имеющихся в данных параметрам [8], либо прямое сравнение агрегированной выдачи (например скетчей) с разным набором параметров. Соответствующая функциональность уже до некоторой степени реализована в корпусе, но ее применение будет вполне эффективным только после решения вопроса о очистке данных в ГИКРЯ 2.0.

6. ГИКРЯ как фабрика датасетов

ГИКРЯ играет заметную роль в русскоязычной компьютерной лингвистике, выступая источником размеченных датасетов для лингвистических задач разного типа, например, автоматической морфологической и синтаксической разметке, анализу референциальных цепочек и восстановлению эллипсиса, задачам нетематической классификации. Назовем только 4 таких датасета, использованных для проведения технологических соревнований: датасеты с морфоразметкой для тестирования систем исправления опечаток и морфологического анализа, датасет для задачи анализа явлений гэппинга и определения авторства [14], [15], [17].

Наличие априорной метатекстовой и автоматической лингвистической разметки разного типа позволяет делать датасеты с разными параметрами, но вопрос чистоты исходного корпуса оказывается первостепенно важным.

7. Заключение

Авторы осознают, что изложенный материал имеет характер отчета. Однако резюмирующего текста, описывающего уже шестилетний опыт использования ГИКРЯ, до сих пор не публиковалось.

Мы надеемся, что изложенные проблемы — а это те проблемы, с которыми мы столкнулись при реализации идеи дифференциального корпуса — будут интересны корпусным коллегам. Не менее существенно и то, что корпусом пользуются сотни лингвистов, не всегда осознавая те серьезные подводные камни, о которых шла речь в докладе.

8. Благодарности

Авторы, представляющие команду ГИКРЯ, благодарят всех коллег, пользователей корпуса, результаты работы которых стимулировали исследование их надежности. Мы выражаем благодарность Физтех-Школе Прикладной математики и Информатики за поддержку проекта ГИКРЯ, который с 2020 года развивается усилиями лаборатории компьютерного зрения и анализа социальных медиа ABBYY Lab в составе этой Физтех-школы.

Литература

1. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.*, (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation, Web as Corpus Workshop (WAC-8).
2. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2014) Variational Corpus Statistics Using Author Profiles In Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.
3. *Benko V., Zakharov V. P.* (2016) Very Large Russian Corpora: New Opportunities and New Challenges. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”, Moscow.

4. *Benko V.* (2019) Dedublication in Large Web Corpora. In. Proc.of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff.
5. *Iglina A., Smurov I.* (2019). Nonthematic texts Classification: Gender of the Author. Publ. in MIPT_CoLing_Bachelors Diplomas.
6. *Ivoylova A., Raskin I., Selegey D.* A New Dataset to solve the task of non-author text filtration in social networks-based corpora. In Proc. of Student Workshop at Dialogue, Russian International Conference on Computational Linguistics, Moscow.
7. *Katinskaya A., Sharoff S.* (2015) Applying Multi-dimensional Analysis to a Russian Webcorpus: Searching for Evidence of Genres , in Proc. of the Workshop on Balto-Slavic Natural Language Processing associated with the International Conference RANLP, Hissar, Bulgaria.
8. *Lagutin M. B., Kuratov Y., Kopylov N.* (2016) Statistical analysis of the search results in a differential corpora. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016” Moscow
9. *Laippala V. et al.* (2020) From Web Crawl to Clean Register-Annotated Corpora, — in Proc. of 12th WaC Workshop, Marseille.
10. *Petrenz P., Webber B.* (2011). Stable Classification of Text Genres: Computational Linguistics. Vol. 37, No. 2.
11. *Piperski A., Belikov V., Kopylov N., Selegey V., Sharoff. S.* (2013) Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In : 8th WAC Workshop.
12. *Piperski A.* (2013) The General Internet Corpus of Russian and the Notion of Representativeness in Corpus Linguistics. In Proc. of Institute of Linguistics (Russian State University for the Humanities, Moscow.
13. *Pomikálek, J., Jakubíček, M., and Rychly, P.* (2012). Building a 70 billion word corpus of English from ClueWeb. In LREC, pages 502–506.
14. *Ponomareva M., Droganova K., Smurov I, and Shavrina T.*(2019). AGRR-2019: A Corpus for Gapping Resolution in Russian. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 35–43. Florence, Italy. Association for Computational Linguistics.
15. *Solonin M.* (2019) Evaluation of Authorship Attribution Methods for Russian Texts. In. Computational Linguistics and Intellectual Technologies, Moscow, Supplementary volume pp. 240–246.
16. *Sorokin A.* (2015) Automatic Regional Classification Using a Dictionary of Regional Lexics: a Preliminary Study. In Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.
17. *Sorokin A, Baytin A., Galinskaya I., Shavrina T.* (2016) SpellRuEval: the First Competition on Automatic Spelling Correction for Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016” Moscow.
18. *Wenzek G., Lachaux M. et al.* (2019) CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data (2019). arXiv:1911.00359 [cs.CL].

MODELING LEMMA FREQUENCY BANDS FOR LEXICAL COMPLEXITY ASSESSMENT OF RUSSIAN TEXTS¹

Blinova O. V. (o.blinova@spbu.ru; ovblinova@hse.ru)

Saint Petersburg State University;
HSE University, Saint Petersburg, Russia

Tarasov N. A. (tarasovn2468@yandex.ru),

Modina V. V. (st072157@student.spbu.ru),

Blekanov I. S. (i.blekanov@spbu.ru)

Saint Petersburg State University, Saint Petersburg, Russia

The paper is devoted to the problem of modeling general-language frequency using data of large Russian corpora. Our goal is to develop a methodology for forming a consolidated frequency list which in the future can be used for assessing lexical complexity of Russian texts.

We compared 4 frequency lists developed from 4 corpora (Russian National Corpus, ruTenTen11, Araneum Russicum III Maximum, Taiga). Firstly, we applied rank correlation analysis. Secondly, we used the measures “coverage” and “enrichment”. Thirdly, we applied the measure “sum of minimal frequencies”. We found that there are significant differences between the compared frequency lists both in ranking and in relative frequencies. The application of the “coverage” measure showed that frequency lists are by no means substitutable. Therefore, none of the corpora in question can be excluded when compiling a consolidated frequency list.

For a more detailed comparison of frequency lists for different frequency bands, the ranked frequency list, based on RNC data, was divided into 4 equal parts. Then 4 random samples (containing 20 lemmas from each quartile) were formed.

Due to the wide range of values, accepted by ipm measure, relative frequency values are difficult to interpret. In addition, there are no reliable thresholds separating high-frequency, mid-frequency, and low-frequency lemmas. Meanwhile, to assess the lexical complexity of texts, it is useful to have a convenient way of distributing lemmas with certain frequencies over the bands of the frequency list. Therefore, we decided to assign lemmas “Zipf-values”, which made the frequency data interpretable because the range of measure values is small.

The result of our work will be a publicly accessible reference resource called “Frequentator”, which will allow to obtain interpretable information about the frequency of Russian words.

Key words: Russian, linguistic corpora, lemma frequency lists, general-language frequency, frequency bands, low-frequency words, lexical complexity

DOI: 10.28995/2075-7182-2020-19-76-92

¹ The presented research was supported by the Russian Science Foundation, project #19-18-00525 “Understanding official Russian: the legal and linguistic issues”.

МОДЕЛИРОВАНИЕ ЗОН ЧАСТОТНОГО СЛОВАРЯ ДЛЯ ОЦЕНКИ ЛЕКСИЧЕСКОЙ СЛОЖНОСТИ РУССКИХ ТЕКСТОВ

Блинова О. В. (o.blinova@spbu.ru; ovblinova@hse.ru)

Санкт-Петербургский государственный университет;
НИУ «Высшая школа экономики», Санкт-Петербург, Россия

Тарасов Н. А. (tarasovn2468@yandex.ru),

Модина В. В. (st072157@student.spbu.ru),

Блеканов И. С. (i.blekanov@spbu.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Introduction

The study is aimed at the problem of forming a consolidated lemma frequency list based on the frequency lists of large Russian corpora. Such a list can be used to assess the lexical complexity of Russian texts (for example, it will be possible to estimate the number of low-frequency, i.e. unfamiliar, words of the text and use these values in readability formulas). Such a list should contain interpretable frequency values that will allow us to divide the frequency list into bands and distinguish between high-frequency, mid-frequency and low-frequency lemmas.

Section 1 discusses readability formulas that take into account the number of long words or (un)familiar words; it is concluded that the application of the familiarity criterion is difficult to operationalize without reference to word frequency data. **Section 2** shows that features including word frequency information successfully predict text complexity. **Section 3** discusses general-language frequency and the problem of accounting for the reader's actual language experience. **Section 4** briefly discusses approaches to identifying frequency bands. **Section 5** gives a description of four Russian corpora, whose frequency lists are involved in the comparison. Section 6 describes the methods for comparing frequency lists; **section 7** gives the results of applying the selected methods. The results indicate that there are significant differences between the compared frequency lists both in the ranks of the lemmas and in their relative frequencies, and that the frequency lists are not substitutable. **Section 8** justifies the use of the frequency measure “Zipf-value” which has a small range of values.

1. Long or unfamiliar words and texts complexity

There is a fairly long tradition of applying readability assessment methods to texts in Russian; for a review see [Reynolds 2016]. In particular, readability metrics

are used, that is, formulas where variables include the number of complex words. Complex words can be understood either as long (multicharacter or multisyllabic) units, or as unfamiliar units.

Although, as K. Collins-Thompson pointed out, “the word lists used in vocabulary-based readability measures like Dale-Chall may be thought of as a simplified language model” [Collins-Thompson 2014], see also [Crossley et al. 2019], the use of such formulas is a common method for assessing the document complexity. Presently it is used in combination with other, more sophisticated methods, for more details see, for example [Benjamin 2012]. More precisely, the number of complex (long, unfamiliar/rare/low-frequency) words of the text or the average length of words in letters or syllables is used in various text classification models as one of many features, see, e.g., [Schwarm, Ostendorf 2005].² It is clear that, with the exception of some special cases,³ the application of the familiarity criterion is difficult or impossible to operationalize without using word frequency information.⁴

2. Word frequency as a parameter for text complexity assessing

According to [Leroy, Kauchak 2014], the word frequency is closely related to both the actual word complexity (measured by how well readers can choose the correct definition of the word) and the difficulty to read.

The studies of Russian text complexity for native speakers or second language learners also show that lexical features, including information on word frequency and/or inclusion in vocabulary lists for each CEFR level (“lexical minimums”), successfully predict complexity. For instance, according to [Laposhina 2017], it is precisely these features that showed the highest correlation with complexity. In [Ivanov et al. 2018] metrics based on lexical features (including word frequency, average frequency of nouns, etc.) are evaluated as reliable, see also [Sharoff et al. 2008], [Solovyev et al. 2018].

Frequency information can be applied in various ways. The average absolute word frequency or mean log frequency [Collins-Thompson, Callan, 2005], the total frequency of content words [Inavov et al. 2018] etc. can be used as measures of lexical complexity. In addition, when assessing text complexity, one can take into account the number of words that are not included in the lists of (high)frequency words, for more details on more sophisticated models, see [Chen, Meurers 2016].

² Recent studies show that “sentence and word length measures likely do not tap directly into linguistic components related to readability” [Crossley et al. 2019]. However, it is clear that the various parameters for assessing lexical complexity are not independent of each other, in particular, according to Zipf’s law of abbreviation, the length of a word correlates with its frequency, see, for example, [Bentz, Ferrer-i-Cancho 2016].

³ These are cases with “lexical minimums” or with the results of painstaking surveys aimed at identifying familiar words.

⁴ For example, in [Batinić et al. 2016] and in “LeStCor: Levelled Study Corpus of Russian” the words included in the list of 5000 most frequent Russian words compiled by S. A. Sharoff [Sharoff, electronic resource], see also [Sharoff et al. 2013], are treated as familiar.

Lemma frequency can be estimated using frequency dictionaries or representative corpora. In this paper, we focus on the problem of the general-language frequency modeling based on data from large Russian corpora.

3. In search of general-language frequency

According to K. Collins-Thompson, “a widely-used feature of lexical difficulty for a word is thus the relative frequency of that word **in everyday usage**,⁵ as measured by its relative frequency in a **large representative corpus**, or its presence/absence in a **reference word list**” [Collins-Thompson 2014]. To assess the general-language frequency of words, one should use some “general-language corpus”, see the studies on designing and balancing corpora and corpora representativeness, e.g., [Atkins et al. 1992]. As stated in [Biber 1993: 247], a representative corpus “might contain roughly 90% conversation”.⁶

In [Chen, Meurers 2016] this problem of accounting for the actual competence of a native speaker is also discussed, cf.: “the frequency lists adopted by these studies were mostly drawn from written corpora. Spoken language was rarely taken into consideration when frequency lists were being composed. This runs the risk of the frequency values not being a faithful representation of the reader’s actual language experience, hence being suboptimal for predicting the ease of perception and retrieval”. Accordingly, when modeling the general-language frequency for Russian it would be reasonable to give greater weight to the frequency values, obtained from a spoken corpus (e.g., Corpus of Spoken Russian in the Russian National Corpus).

4. Methods for modeling general-language frequency and frequency bands

The word frequency effect studies demonstrate that high-frequency words are usually perceived and produced more efficiently and faster than low-frequency ones, see, for example, [Brysbaert et al. 2018].

Meanwhile, if we use classical techniques for text complexity prediction using frequency information, averaging over all frequency values, then the contribution of low-frequency words becomes minimal [Chen, Meurers 2016]. Therefore, we are faced with the task of identifying frequency bands that explicitly show high-frequency, low-frequency, and mid-frequency units.

⁵ See also citation from [Slioussar 2005]: “Many psycholinguists who use data on the frequency of certain words or forms are often subjected to harsh criticism. After all, such data is most often taken from frequency dictionaries, based exclusively on written texts, not oral ones. Even to a layman it is intuitively clear that the frequency of words and their forms in colloquial speech should correlate with the frequency presented in the mental lexicon”.

⁶ As far as we know, balanced corpora organized according to the indicated principle have not been created yet.

Various thresholds values (for the frequencies or ranks) are used to separate the bands.⁷ The conventional threshold value for low-frequency words in a 100 million word corpus is 5 ipm (items per million) [Lyashevskaya 2016: 236]. Different threshold values are also used for ranks. High-frequency units are the words with a rank up to 2,000 [Schmitt 2010, 69]; mid-frequency units are words with ranks from 2,000 to 8,000–9,000 [Schmitt 2010: 70]. Rare units in the New Frequency Dictionary of Russian are the lemmas with a rank of 10,000 and more [Lyashevskaya 2016: 229]. The entire frequency list can be divided into quartiles (for example, in [Zhao, Jurafsky 2009] words from the lower quartile of the ranked frequency list are considered as low-frequency ones); percentiles can also be used for this purpose, see [Bell et al. 2009].

In this paper we compare 4 frequency lists based on four Russian corpora. These corpora are of different size and composition. Our goal is to develop a methodology for creating a consolidated lemma frequency list based on the frequency lists of large Russian corpora.

5. Frequency data sources

This paper compares frequency lists derived from three large web corpora: ruTenTen11 [ruTenTen11, electronic resource], [Kilgariff et al. 2014], Araneum Russicum III Maximum [Araneum Russicum, electronic resource], [Benko 2014], Taiga [Taiga, electronic resource], [Shavrina, Shapovalova 2017] and the New Frequency Dictionary of Russian (NFDR), based on data from Russian National Corpus [RNC, electronic resource], [Lyashevskaya, Sharoff 2009].

Frequency lists were obtained from the corpora sites or from corpora creators.⁸ In the current version of the Sketch Engine, it was possible to download word lists no longer than 1,000 lines. Therefore, to obtain the most complete frequency list from ruTenTen11, frequency lists of lemmas starting with possible two-letter combinations (*ab*, *av*, *az* etc.) were downloaded. The list of possible combinations is obtained using NFDR. For single-letter lemmas, a separate search was performed.

Table 1. Frequency data sources

Corpus	Composition	Size	Analyser	Number of lemmas in the frequency list
RNC (NFDR)	genre-balanced RNC subcorpus	91,982,416 graphic words	Mystem	52,138 lemmas with relative frequency ≥ 0.4 ipm (37 occurrences)

⁷ It should also be added that low-frequency words are included into the dictionaries of rare, forgotten, uncommon and obsolete words, see, for example, [Somov 1996], [Glinkina 1998], [Ilinskaja 1989], [Rogozhnikova 1997], [Korpusnoj slovar' redkih slov, electronic resource].

⁸ The authors of this paper would like to thank Tatyana Shavrina for the opportunity to use the frequency list of the Taiga corpus.

Corpus	Composition	Size	Analyser	Number of lemmas in the frequency list
ruTenTen11	Internet: news and commercial sites, blogs, social media	near 18 billion tokens (14,553,856,113 text forms)	Treetagger	457,473 lemmas with absolute frequency ≥ 5
Araneum Russicum III Maximum	Internet: news and commercial sites, blogs, social media	15,961,200,372 words	Treetagger	8,893,947 units with absolute frequency ≥ 5
Taiga	Internet: 77% of literary texts (the articles from 33 literary magazines), 19% of naive poetry, 2% of news (from 4 popular news sites), 2% of other texts (popular science, texts of social networks, etc.)	near 5 billion words	UDPipe	2,988,610 lemmas with absolute frequency ≥ 1

6. Methods for frequency list comparison

There are a number of ways to compare frequency lists and methods for measuring the distance between them. In particular, there are measures based on geometrical notions (Euclidean distance, Manhattan distance, Cosine distance, etc.), measures based on well-known statistical tests and procedures (Chi-Square-based measures, Log-Likelihood, Spearman's ρ , etc.), information theoretic measure “perplexity”, measure of distance by keywords (Simple Maths) and others, see [Kilgarriff, Rose 1998], [Piperski 2018], [Gomaa, Fahmy 2013] and many others. We chose three measures that allowed us to look at the differences between frequency lists from different points of view (comparing ranks of lemmas, the values of relative frequencies or estimating overlap between the lists).⁹

Firstly, we applied the **rank correlation analysis**, calculating the values of the Spearman and Kendall rank correlation coefficients for pairs of frequency lists. The lists were compared by intersecting lemmas, which equalized their length.

Secondly, we applied two **measures of overlap** (“**Coverage**” and “**Enrichment**”), considered in [Baroni et al. 2009]. The Coverage measure is calculated by the formula:

$$Coverage(X, Y) = \frac{(N1 \cap N2)}{N1}, \quad (1)$$

where X, Y are the corpora, $N1$ is the number of lemmas with an absolute frequency greater than or equal to a given cutoff value in the corpus X , $N2$ is the number of lemmas with an absolute frequency greater than or equal to a given cutoff value in the corpus Y .

⁹ According to [Piperski 2018], the preferred frequency-based measure of corpus distance is Euclidean distance, as this measure is the most robust to corpus size. At the same time, to achieve the objectives of this article, it is sufficient to apply the three measures we have chosen. In addition, some measures (Spearman's ρ , Chi-Square) are commonly used, that is, their application will allow ones comparing our results with the results obtained earlier, see [Khokhlova 2016].

The Enrichment measure is calculated by the formula:

$$Enrichment(Y, X) = \frac{M2}{M1}, \tag{2}$$

where $M2$ is the number of lemmas with a frequency above the threshold in the corpus Y and below the threshold in the corpus X , $M1$ is the number of lemmas with an absolute frequency below the threshold in corpus X . As a threshold value, we (following [Baroni et al. 2009]) used the absolute frequency of 20 occurrences. This is the so-called “Sinclair threshold”. This (apparently arbitrary) threshold was chosen under the influence of J. Sinclair’s statement that an experienced lexicographer would need at least 20 occurrences of an unambiguous word to make a description of its behavior, see, for example, [Lüdeling, Kytö 2009: 818].

Thirdly, we applied the measure “Sum of Minimum Frequencies” (SMF), proposed by A. Ya. Shaikevich in [Shaikevich 2015], see also [Piperski 2017]. SMF is calculated by the formula:

$$SMF(X, Y) = \frac{\sum_{min}(pX_i, pY_i)}{\sum_{0.5}(pX_i, pY_i)}, \tag{3}$$

where pX_i is the relative frequency of the lemma in the corpus X , pY_i is the relative frequency of the lemma in the corpus Y .

7. Comparison results

The frequency lists under consideration did not undergo any special preprocessing. Table 2 shows the results of applying rank correlation analysis.

Table 2. Spearman’s ρ and Kendall’s τ values

Spearman’s ρ				Kendall’s τ			
X/Y	ruTenTen11	Taiga	NFDR	X/Y	ruTenTen11	Taiga	NFDR
Araneum	0.033	0.081	0.223	Araneum	0.022	0.006	0.157
ruTenTen11		0.071	0.828	ruTenTen11		0.048	0.648
Taiga			0.095	Taiga			0.065

The rank correlation coefficient ρ takes value > 0.7 only in the pair ruTenTen11—NFDR ($\rho = 0.828$). This can be explained by the fact that these lists are the shortest and do not contain very long low-frequency tails. In pairs of web-corpora, the correlation coefficients values do not exceed 0.3, that is, the differences in ranking across these corpora are significant.

Table 3 shows the comparison results using Coverage and Enrichment measures. Coverage is a measure of the proportion of words for which there is “enough” information in the corpus X and “enough” information in the corpus Y [Baroni et al. 2009]. In other words, this is “a (very rough) measure of the extent to which X is ‘substitutable’ with Y ” [Ibid.]. Enrichment allows one to estimate the proportion of words among those words that are attested in the corpus X , and for which there is not enough information in the corpus X , but enough information in the corpus Y [Ibid.].

Table 3. Values of the measures of overlap, threshold = 20¹⁰

Coverage				Enrichment			
X/Y	Araneum	ruTenTen11	Taiga	X/Y	Araneum	ruTenTen11	Taiga
Araneum		53	51.5	Araneum		0.9	0.2
ruTenTen11	7.8		23.1	ruTenTen11	3.4		1.9
Taiga	4.6	14.1		Taiga	13.9	0.2	

When interpreting presented metrics values, it should be taken into account that the measures are able to evaluate the ratio of frequency lists as X/Y or as Y/X . The Coverage measure has the highest value for the pairs Araneum (X)—ruTenTen11 (Y) (53) and Araneum (X)—Taiga (Y) (51.5); the proportion shows that only about half of the words above the cutoff in Araneum are also above the cutoff in ruTenTen11 and Taiga. Thus, the vocabularies of the compared web corpora are significantly different. The Enrichment values allow one to assess the extent to which the frequency lists are capable of enriching each other. The highest value measure is found for the Taiga—Araneum pair (13.9). Thus, if we consider the entire frequency range in question, the use of various web-corpora is not so beneficial.

On the whole, the assessment of the overlap allows us to conclude that the frequency lists are not substitutable, and when compiling a consolidated frequency list of lemmas, all compared frequency lists should be used.

Finally, **Table 4** shows the results of comparing all four lists using SMF measure. This measure compares relative frequencies of all intersecting elements (lemmas) in the lists in pairs.

Table 4. Values of SMF measure

X/Y	ruTenTen11	Taiga	NFDR
Araneum	0.056	0.024	0.264
ruTenTen11		0.116	0.756
Taiga			0.197

Particular attention should be paid to the results of the comparison of web corpora with NFDR. The high value we observe in the pair NFDR—ruTenTen11 (SMF = 0.756). We saw earlier that the rank correlation coefficients for this pair also take the largest value from the observed values. Significantly less similar are NFDR and Araneum (SMF = 0.264), NFDR and Taiga (SMF = 0.197). This can also be explained by the fact that the frequency lists of Araneum and Taiga contain long tails of low-frequency units.

Thus, applying three measures, we found out that there is significant discrepancy across the lists in ranking and in relative frequencies. The use of the Coverage measure showed that frequency lists are by no means substitutable. Therefore, none of the corpora in question can be excluded when compiling a consolidated frequency list.

¹⁰ We did not include NFDR in the comparison, since this list contains lemmas with a relative frequency of 0.4 ipm or more (that is, an absolute frequency ≥ 37).

8. Comparison by frequency bands

For a more detailed comparison of frequency lists by different frequency bands, we decided to proceed as follows. We divided the ranked NFDR frequency list into 4 equal parts, then, using the ranks, we formed 4 random samples (containing 20 lemmas from each quartile). For each lemma of 4 random samples, we assigned the values of relative frequencies according to all the compared lists. The data obtained for the upper and lower quartiles are presented in [Table 7](#) and [Table 8](#) below.

We see that even for lemmas from the upper quartile, there are significant differences in the ipm values according to different corpora. So, the range of ipm values for the most frequent lemma in the sample (the noun *центр* ‘centre’) is 390.80.

It is important that the **overall range of ipm values** is very significant. NFDR contains lemmas with relative frequencies from 35,801.8 (the conjunction *и* ‘and’) to 0.4 ipm, Taiga includes lemmas with a frequency from 18,710.7 (the preposition *в* ‘in, to, into’) to 0.0017 ipm. A significant number of lemmas have frequencies <1 ipm. For example, the Taiga frequency list of 2,988,608 lines contains only 28,500 lemmas with a frequency of ≥ 1 ipm (and this is less than 1/100 of the entire list). The observed proportion of rare words is a consequence of the Zipf’s law.

Due to the wide range of values, the observable values of relative frequency are **difficult to interpret**. In addition, there are no reliable thresholds separating high-frequency, mid-frequency, and low-frequency words. Meanwhile, it is useful to have a convenient way of assigning lemmas to certain frequency bands.

Therefore, we (following [[Chen, Meurers 2016](#)]) decided to use the approach from [[Van Heuven et al. 2014](#)], where a new “Zipf-value” measure of frequency is proposed. The value of this measure is calculated by the formula (4).

$$\text{Zipf-value} = \log_{10}(\text{ipm} \times 1000), \quad (4)$$

The measure has the following advantages, see [[Ibid.](#)].

1. A logarithmic scale is used.¹¹
2. The values are easy to interpret. For example, the most frequent word in NFDR *и* ‘and’ has Zipf-value equal to 7.55 (or, when rounding to an integer, 8). The word with the lowest frequency in NFDR will have a Zipf-value of 2.6 (or 3).
3. The scale allows us to separate mid-frequency words from high-frequency and low-frequency ones.
4. Zipf-values are easy to calculate if we know ipm values.

The discussed approach is not the only one possible. In [[Sharoff et al. 2017](#)] another logarithmic measure of the frequency “FClass” is proposed (see the formula (5), where $\text{freq}(\max)$ is the absolute frequency of most frequent word (MFW) in a particular corpus, $\text{freq}(w)$ is the absolute frequency of the word in a particular corpus, for which the measure value is calculated).¹²

¹¹ The values of the logarithmic frequencies are used by psycholinguists, see for example, [[Winter 2020, 95](#)].

¹² The authors would like to thank the anonymous reviewer for pointing out this measure.

$$FClass(w) = \log_2 \frac{freq(max)}{freq(w)}, \quad (5)$$

FClass measure also has a small range of values. For example, the lemma *субпопуляция* ‘subpopulation’ from the lower quartile of NFDR frequency list will take FClass values equal to 16 and 21 (see [Table 5](#)).

Table 5. FClass values

	<i>freq</i> (субпопуляция)	MFW	<i>freq</i> (max)	FClass
NFDR	37	<i>u</i> ‘and’	3,293,765.6	16
Taiga	5	<i>в</i> ‘in, to, into’	11,076,749	21
Araneum	194	<i>u</i> ‘and’	563,822,183	21

The upper FClass value can be estimated at $freq(w) = 1$, the range of measure values for the compared corpora is [0;22], or [0;23], or [0;29], see. [Table 6](#).

Table 6. Maximum FClass values

	<i>freq</i> (w)	<i>freq</i> (max)	FClass
NFDR	1	3,293,765.6	22
Taiga	1	11,076,749	23
Araneum	1	563,822,183	29
ruTenTen11	1	503,894,565	29

The range of FClass values is greater than the range of Zipf-value. FClass scale does not look like typical Likert rating scale [[Jamieson 2004](#)]. Accordingly, interpreting Zipf-values is a simpler task.

Compared frequency lists, as shown below (see [Fig. 1](#)), obey exponential law. Therefore, we can use Zipf-value as a frequency measure.

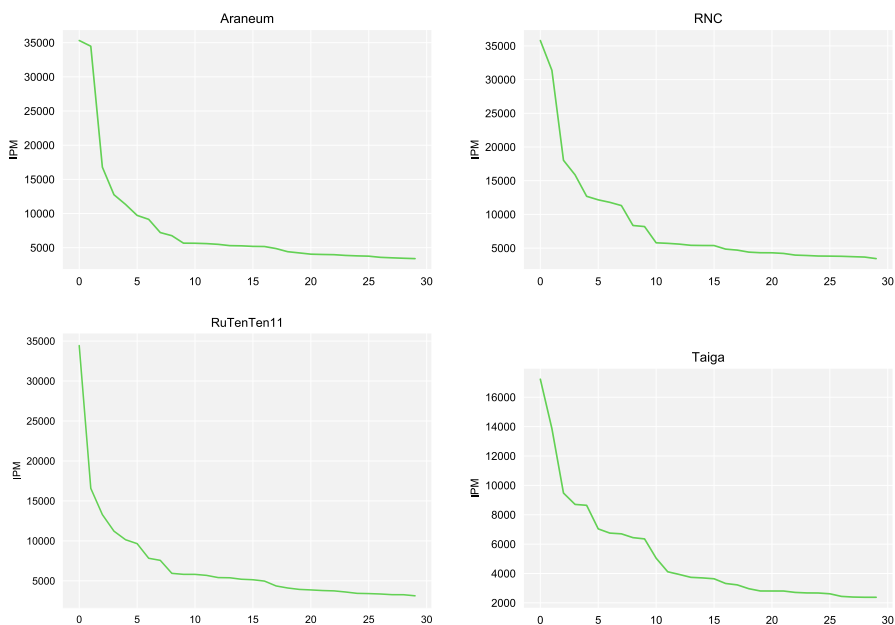


Fig. 1. Frequency distribution

Table 7 and 8 show the Zipf-values calculated for the ipm values in particular frequency lists, as well as mean values, which we will interpret. These values range from 6 (*центр*) to 2 (*субпопуляция*).

Table 7. Lemmas from the upper quartile of the NFDR list

lemma, PoS, translation	ipm				Zipf-value				
	NFDR	Taiga	Araneum	ruTenTen	NFDR	Taiga	Araneum	ruTenTen	mean
<i>центр</i> , N, ‘centre’	265.9	187.28	577.41	578.07	5	5	6	6	6
<i>участок</i> , N, ‘area; district, ward; plot’	144.2	88.32	299.34	273.94	5	5	5	5	5
<i>круглый</i> , A, ‘round’	71.9	24.96	70.43	73.15	5	4	5	5	5
<i>памятник</i> , N, ‘monument’	63.8	61.44	81.81	82.62	5	5	5	5	5
<i>превратиться</i> , V, ‘to transform into’	63.5	0.49	46.47	40.17	5	3	5	5	4
<i>чемодан</i> , N, ‘suitcase’	42.8	10.24	10.72	11.28	5	4	4	4	4
<i>туалет</i> , N, ‘toilet, lavatory’	35.1	20.33	31.46	32.12	5	4	4	5	4
<i>волшебный</i> , A, ‘magic’	28.2	12.92	37.42	31.17	4	4	5	4	4
<i>пилот</i> , N, ‘pilot’	26.9	14.51	20.36	27.5	4	4	4	4	4

lemma, PoS, translation	ipm				Zipf-value				
	NFDR	Taiga	Araneum	ruTenTen	NFDR	Taiga	Araneum	ruTenTen	mean
<i>привлечение</i> , N, 'attraction'	26	14.26	64.41	63.68	4	4	5	5	5
<i>ласково</i> , Adv, 'tenderly'	23.6	6.69	4.43	5.7	4	4	4	4	4
<i>заказывать</i> , V, 'to order'	14.5	15.94	38.5	74.1	4	4	5	5	4
<i>взорваться</i> , V, 'to implode'	14.1	1.02	5.43	6.1	4	3	4	4	4
<i>вытягивать</i> , V, 'to outstretch; to pull out'	9	9.55	4.66	11.54	4	4	4	4	4
<i>Виноградов</i> , N, 'Vinogradov'	7.9	2.87	3.8	5.73	4	3	4	4	4
<i>селедка</i> , N, 'herring'	7.3	2.14	2.78	2.11	4	3	3	3	3
<i>прибить</i> , V, 'fasten (by nailing)'	7.2	0.09	2.22	0	4	2	3	—	3
<i>растворяться</i> , V, 'to dissolve'	7.2	9.04	5.97	6.76	4	4	4	4	4
<i>овощной</i> , A, 'vegetable'	6.6	0.65	12.46	12.3	4	3	4	4	4
<i>девяностый</i> , Num, 'ninetieth'	6.1	4.47	0.04	4.51	4	4	2	4	3

It should be noted that lemmatizers assign different lemmas to the forms of Russian verbs, cf. *превратиться* (Pf)—*превращаться* (Impf), see [Lyashevskaya 2016: 228] about this problem. This is one of the reasons for discrepancies between the frequency lists. The lemma *превратиться* is present in all frequency lists, but in the Taiga list *превратиться* (Pf) has ipm = 0.49, while the lemma *превращаться* (Impf) has ipm = 55.36, which is much closer to the values demonstrated by others corpora. Similar discrepancies in the ipm values are observed for lemmas *взорваться* (*взрываться*) and *прибить* (*прибивать*).

The list of lemmas from the second quartile can be commented on in the same way as the list of lemmas from the first one. In the ruTenTen11 list the lemma *подоспеть* (Pf) 'arrive in time' was not found, but there was the lemma *подоспевать* (Impf). Lemmas from the second quartile (three of which have an average Zipf-value equal to 4, 16 have a Zipf-value equal to 3, 1 (*окрылить* 'inspire') has a Zipf-value equal to 2) for the most part can be considered as mid-frequency ones. The list of lemmas from the third quartile is also quite homogeneous: 15 out of 20 lemmas (75%) have a Zipf-value of 3.

Some low-frequency lemmas from the lower quartile (translation is given in the Table 8) cannot be found in two frequency lists of four (*послепожарный*, *тире*), or one frequency list (*несолоно*, *экономразвитие*, *напряг*, *поубавить*, *промельк*, *субпопуляция*). This fact can be explained by lemmatization errors. For instance, representations of the lemma *роздых* in various cases (except for the nominative) are present in the Araneum frequency list.

Accordingly, before the preprocessing of frequency lists for the purpose of forming a consolidated list, it is necessary to decide how to deal with such occurrences as *роздыха*, *роздыху* etc. Apparently, to such occurrences should be assigned normalized forms, and the frequencies of different word forms, related to the same lemma, should be summarized.

Table 8. Lemmas from the lower quartile of the NFDR list

lemma, PoS, translation	ipm				Zipf-value				
	NFDR	Taiga	Araneum	ruTenTen	NFDR	Taiga	Araneum	ruTenTen	mean
<i>тире</i> , N, ‘dash’	0.8	0	1.09	0	3	—	3	—	3
<i>тявкать</i> , V, ‘to yap’	0.7	0.69	0.15	0.28	3	3	2	2	3
<i>хроматин</i> , N, ‘chromatin’	0.7	0.02	0.05	0.27	3	1	2	2	2
<i>линейно</i> , Adv, ‘linearly’	0.6	0.24	0.96	1.05	3	2	3	3	3
<i>несолоно</i> , Adv, lit. ‘unsaltedly’	0.6	0.16	0.1	0	3	2	2	—	2
<i>отжимание</i> , N, ‘press-up; pressing out’	0.6	0.24	2.2	0.23	3	2	3	2	3
<i>пеленг</i> , N, ‘bearing’	0.6	0.06	0.15	0.39	3	2	2	3	2
<i>денатурация</i> , N, ‘denaturing’	0.5	0.01	0.06	0.18	3	1	2	2	2
<i>подледный</i> , A, ‘subglacial’	0.5	0.34	0.34	0	3	3	3	—	3
<i>роздых</i> , N, ‘rest’	0.5	0.15	0	0.15	3	2	—	2	2
<i>сахарок</i> , N, ‘sugar’ (diminutive)	0.5	0.11	0.08	0.17	3	2	2	2	2
<i>экономразвитие</i> , N, ‘economic development’	0.5	0.06	0.13	0	3	2	2	—	2
<i>буерак</i> , N, ‘ravine’	0.4	0.15	0.01	0.25	3	2	1	2	2
<i>втык</i> , N, ‘tongue-lashing’	0.4	0.09	0.08	0.1	3	2	2	2	2
<i>депонировать</i> , V, ‘to deposit’	0.4	0.04	0.05	0.31	3	2	2	2	2
<i>напряг</i> , N, ‘stress’	0.4	0.96	0.41	0	3	3	3	—	3
<i>послепожарный</i> , A, ‘post-fire’	0.4	0.03	0	0	3	1	—	—	2
<i>поубавить</i> , V, ‘to diminish’	0.4	0.08	0.1	0	3	2	2	—	2
<i>промельк</i> , N, ‘flash’	0.4	0.3	0.01	0	3	2	1	—	2
<i>субпопуляция</i> , N, ‘subpopulation’	0.4	0.01	0.01	0	3	1	1	—	2

Conclusion

Thus, we compared the frequency lists derived from four Russian corpora. Our aim was not comparison itself, but the development of a methodology for creating a consolidated frequency list and modeling the general-language frequency. It seems that the inclusion of Zipf-value in such a list will make the frequency data interpretable, since the range of measure values is small (the most frequent lemmas will have Zipf-values equal to 7 and 8, the least frequency lemmas will have Zipf-values equal to 1 and 2).

The result of our work¹³ will be a publicly accessible reference resource called “Frequentator” which will allow to obtain interpretable information about the frequency of Russian words. To create such a resource, it will be necessary to preprocess

¹³ The authors would like to express their sincere gratitude to anonymous reviewers for useful comments regarding the upcoming work on the consolidated frequency list formation.

the frequency lists of web corpora, detect and remove noise; perform lemmatization of occurrences that do not coincide with normalized forms; assign to each lemma a part-of-speech tag; analyze verbs and form a consolidated list. At the end, each lemma will be assigned a weighted frequency value in ipm and Zipf-value.

References

1. *Araneum Russicum III Maximum*, available at: http://ucts.uniba.sk/aranea_about/russicum.html.
2. Atkins S., Clear J., Ostler N. (1992), *Corpus Design Criteria*, Literary and Linguistic Computing, Vol. 7, № 1, pp.1–16.
3. Baroni M., Bernardini S., Ferraresi A. & Zanchetta E. (2009), The WaCky wide web: a collection of very large linguistically processed webcrawled corpora, *Language Resources and Evaluation*, 43, pp. 209–226.
4. Batinić D., Birzer S. & Zinsmeister H. (2016), Creating an extensible, levelled study corpus of Russian, Dipper S., Neubarth F., Zinsmeister H. (eds.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pp. 38–43. (Bochumer Linguistische Arbeitsberichte 16).
5. Bell A., Brenier J. M., Gregory M., Girand C., Jurafsky D. (2009), Predictability effects on durations of content and function words in conversational English, *Journal of Memory and Language*, 60, pp. 92–111.
6. Benjamin R. G. (2012), Reconstructing readability: recent developments and recommendations in the analysis of text difficulty, *Educational Psychology Review*, 24(1), pp. 63–88.
7. Benko V. (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. P. Sojka, A. Horák, I. opeček and K. Pala (Eds.). *Text, Speech and Dialogue. 17th International Conference, TSD 2014. Proceedings. LNCS 8655*. Springer International Publishing Switzerland, pp. 257–264.
8. Bentz C. & Ferrer-i-Cancho R. (2016), Zipf's law of abbreviation as a language universal, Bentz C., Jager G. & Yanovich I. (eds.) *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen, online publication system, available at: <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558>.
9. Biber D. (1993), Representativeness in Corpus Design, *Literary and Linguistic Computing*, Vol. 8, No. 4, pp. 243–257.
10. Brysbaert M., Mandera P., Keuleers E. (2018), The Word Frequency Effect in Word Processing: An Updated Review, *Current Directions in Psychological Science*, Vol. 27, Iss. 1, pp. 45–50.
11. Chen X., Meurers W. D. (2016), Characterizing Text Difficulty with Word Frequencies, *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 84–94.
12. Collins-Thompson K. (2014), Computational assessment of text readability: a survey of current and future research, François Th. and D. Bernhard (eds.), *Recent Advances in Automatic Readability Assessment and Text Simplification*. Special issue of *International Journal of Applied Linguistics*, 165:2, pp. 97–135.

13. *Collins-Thompson K., Callan J.* (2005), Predicting Reading Difficulty with Statistical Language Models, *Journal of the American Society for Information Science and Technology*, 56(13), pp. 1448–1462.
14. *Corpus dictionary of rare words* [Korpusnoj slovar' redkih slov], available at: <http://ruscorpora.ru/new/rarewords.html>.
15. *Crossley S. A., Skalicky S., Dascalu M.* (2019), Moving beyond classic readability formulas: new methods and new models, *Journal of Research in Reading*, 42, 3–4, pp. 541–561.
16. *Glinkina L. A.* (1998), Illustrated dictionary of forgotten and difficult words of Russian literature of XVIII–XIX centuries [Illjustrirovannyj slovar' zabytyh i trudnyh slov iz proizvedenij russkoj literatury XVIII–XIX vekov], Orenburgskoe knizhnoe izdatel'stvo, Orenburg.
17. *Gomaa W. H., Fahmy A. A.* (2013), A Survey of Text Similarity Approaches, *International Journal of Computer Applications*, Vol. 68, № 13, pp. 13–18.
18. *Ilinskaya N. G.* (1989), Dictionary of uncommon and archaic words [Slovar' maloupotrebiteľ'nyh i ustarevshih slov], Sovetskaja Rossija, Moscow.
19. *Ivanov V. V., Solnyshkina M. I., Solovyev V. D.* (2018), Efficiency of text readability features in Russian academic texts, *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, Vol. 17, pp. 277–287.
20. *Jamieson S.* (2004), Likert scales: how to (ab)use them, *Medical Education*, 38(12), pp. 1217–1218.
21. *Khokhlova M. V.* (2016), Large Corpora and Frequency Nouns, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”*, pp. 237–250.
22. *Kilgarriff A. et al.* (2014), The Sketch Engine: Ten Years On, *Lexicography*, Vol 1, Iss. 1, pp. 7–36.
23. *Kilgarriff A., Rose T.* (1998), Measures for corpus similarity and homogeneity, *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, Granada, Spain, pp. 46–52.
24. *Laposhina A. N.* (2017), Analysis of the relevant features for automatic readability assessment for texts in Russian as a foreign language [Analiz relevantnyh priznakov dlja avtomaticheskogo opredelenija slozhnosti russkogo teksta kak inostrannogo], *Proceedings of the International Conference “Dialogue 2017”* [Trudy Mezhdunarodnoy Konferentsii “Dialog 2017”], Bekasovo, available at <http://www.dialog-21.ru/media/3993/laposhina.pdf>.
25. *Leroy G., Kauchak D.* (2014), The effect of word familiarity on actual and perceived text difficulty, *Journal of the American Medical Informatics Association*, 21(e1), pp. e169–e172.
26. *Lüdeling A., Kytö M.* (eds.) (2009), *Corpus Linguistics: An International Handbook*, Vol. 2, De Gruyter Mouton, Berlin, Boston.
27. *Lyashevskaya O. N.* (2016), *Corpus Instruments for Russian Grammar Studies* [Korpusnye instrumenty v grammaticheskikh issledovanijah russkogo jazyka], Jazyki slavjanskoj kul'tury, Moscow.

28. *Lyashevskaya O. N., Sharoff S. A.* (2009), The frequency dictionary of modern Russian language [Častotnyj slovar' sovremennogo russkogo jazyka], csv-version, available at: <http://dict.ruslang.ru/freq.php>.
29. *Piperski A. Ch.* (2018), Corpus Size and the Robustness of Measures of Corpus Distance, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”, pp. 578–589.
30. *Piperski, A.* (2017), Sum of Minimum Frequencies as a Measure of Corpus Similarity, Presented at the Corpus Linguistics 2017, Birmingham, available at <https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper143.pdf>.
31. *Reynolds R. J.* (2016), Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories, Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 289–300.
32. *Rogozhnikova R. P.* (ed.) (1997), Rare words in the works of the authors of XIX century: Dictionary-companion [Redkie slova v proizvedenijah avtorov XIX veka: Slovar'-spravochnik], Russkie slovari, Moscow.
33. *Russian National Corpus*, available at: <http://www.ruscorpora.ru/new/>.
34. *ruTenTen11*, available at: <https://www.sketchengine.eu/rutenten-russian-corpus/>.
35. *Schmitt N.* (2010), Researching vocabulary: a vocabulary research manual, Palgrave Macmillan, Basingstoke, UK.
36. *Schwarm S. E., Ostendorf M.* (2005), Reading level assessment using support vector machines and statistical language models, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05), pp. 523–530.
37. *Shaikevich A. Ya.* (2015), Measures of lexical similarity between frequency dictionaries [Mery leksicheskogo shodstva chastotnyh slovarej], Proceedings of the International Conference “Corpus linguistics-2015” [Trudy mezhdunarodnoy nauchnoy konferentsii “Korpusnaya linguistica-2015”], Saint Petersburg, pp. 434–442.
38. *Sharoff S., Goldhahn D., Quasthoff U.* (2017), Frequency Dictionary: Russian, Quasthoff U., Fiedler S., Hallsteindóttir E. (eds.), Frequency Dictionaries 9, Leipziger Universitätsverlag.
39. *Sharoff S., Kurella S., Hartley A.* (2008), Seeking needles in the web haystack: Finding texts suitable for language learners, Proceedings of 8th Teaching and Language Corpora Conference (TaLC-8).
40. *Sharoff S., Umanskaya E., Wilson J.* (2013), A frequency dictionary of Russian: core vocabulary for learners, Routledge, NY.
41. *Sharoff S. A.* Frequency dictionary [Častotnyj slovar'], available at: <http://www.artint.ru/projects/frqlist.php>.
42. *Slioussar N. A.* (2005), Psycholinguistic survey in the structure of mental lexicon on the data of Russian verbs. [Psiholingvističeskoe issledovanie struktury mental'nogo leksikona na materiale russkih glagolov], unpublished dissertation, Saint Petersburg.
43. *Solovyev V., Ivanov V., Solnyshkina M.* (2018), Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics, Journal of Intelligent & Fuzzy Systems, Vol. 34, № 5, pp. 3049–3058.

44. *Somov V. P.* (1996), Dictionary of rare and forgotten words [Slovar' redkih i zabytyh slov], VLADOS, Moscow.
45. *Taiga Corpus*. An open-source corpus for machine learning, available at https://tatianashavrina.github.io/taiga_site/.
46. *Van Heuven W. J. B., Mander P., Keuleers, E., Brysbaert, M.* (2014), Subtlex-UK: A new and improved word frequency database for British English, *Quarterly Journal of Experimental Psychology*, 67, pp. 1176–1190.
47. *Winter B.* (2020), *Statistics for Linguists: An Introduction Using R*, Routledge, NY, London.
48. *Zhao Y., Jurafsky D.* (2009), The effect of lexical frequency and Lombard reflex on tone hyperarticulation, *Journal of Phonetics*, 37, pp. 231–247.

THE RUSSIAN LANGUAGE PIPELINE IN THE LIMA MULTILINGUAL ANALYZER

Bocharov V. V. (victor.bocharov@cea.fr),
de Chalendar G. (gael.de-chalendar@cea.fr)

CEA LIST, LASTI, Gif-sur-Yvette, France

In this paper we describe the implementation of Russian language pipeline in LIMA multilingual analyzer and the results obtained in GramEval-2020 shared task. LIMA is a modular pipeline that implements rule-based and machine learning analysis components. Russian language pipeline includes deep neural networks based modules for tokenization, sentence segmentation, part of speech tagging, lemmatization and dependency parsing. Part of speech tags, feature tags and dependency trees conform to Universal Dependencies rules.

Key words: tokenization, part of speech tagging, dependency parsing, lemmatization

DOI: 10.28995/2075-7182-2020-19-93-105

ОБРАБОТКА РУССКОГО ЯЗЫКА В МНОГОЯЗЫЧНОМ АНАЛИЗАТОРЕ LIMA

Бочаров В. В. (victor.bocharov@cea.fr),
де Шаландар Г. (gael.de-chalendar@cea.fr)

CEA LIST, LASTI, Жиф-сюр-Иветт, Франция

В этой статье описана реализация обработки текста на русском языке в анализаторе LIMA и наше участие в соревновании GramEval-2020. Анализатор LIMA — это модульная система обработки текста, включающая статистические и основанные на правилах компоненты. Обработка текста на русском языке реализована при помощи статистических моделей на основе глубоких нейронных сетей и включает токенизацию, морфологический анализ, лемматизацию и построение деревьев зависимостей. Морфологический и синтаксический анализ соответствуют правилам Universal Dependencies.

Ключевые слова: токенизация, морфологический анализ, синтаксический анализ, лемматизация

1. Introduction

The implementation of Russian language pipeline in LIMA is a part of our work on expanding the range of supported languages using machine-learning techniques and Universal Dependencies [16] corpora. LIMA is a modular multilingual toolkit that

includes a language agnostic core and a number of analysis modules sharing a common internal representation of text analysis. Our Russian language support is based on Universal Dependencies annotation of the Russian-SynTagRus corpus, recent deep neural networks models and fastText word embeddings. It includes a tokenizer (which does both word and sentence splitting), a morphological analyzer combined with a dependency parser and a lemmatizer. For our participation in GramEval-2020 shared task, we trained another model using annotated corpora supplied by the organizers.

Source code of LIMA including all mentioned components is available on GitHub¹. Trained models are published in the form of Debian packages². In the following sections, we will first describe related work and then each module in detail. We continue with the evaluation of our results in the shared task before concluding.

2. Related Work

2.1. Universal Dependencies

Universal Dependencies (UD) is an international project and a multilingual annotation framework that provides a universal inventory of linguistic categories and annotations guidelines covering tokenization, part of speech and features tagging and dependency parsing. Within the UD project, a cross-linguistically consistent treebank annotation for many languages is created. A new version of Universal Dependencies treebank collection is released twice a year. Current version UD 2.5 includes 157 treebanks for 90 languages.

There exists a wide range of software³ (editor, visualising tools, consistency checkers and libraries) that works with Universal Dependencies annotation. UDPipe (see below) is a widely known parsing pipeline that produces output following UD guidelines for many languages.

2.2. NLP pipelines and toolkits

There are many known pipeline-based natural language processing systems. GATE (general architecture for text engineering) [4] is an open-source software toolkit originally developed at the University of Sheffield in 1995. GATE includes many analysis modules (processing resources), graphical environment and an information extraction system called ANNIE (A Nearly-New Information Extraction System).

UIMA (Unstructured Information Management Architecture) [6] is an OASIS⁴ standard for content analytics developed at IBM, and Apache UIMA is an open-source implementation of this standard. DKPro (The Darmstadt Knowledge Processing

¹ <https://github.com/aymara/lima>

² <https://github.com/aymara/lima-models>

³ <https://universaldependencies.org/tools.html>

⁴ <https://www.oasis-open.org/>

Software Repository) is a collection of software components for natural language processing based on the Apache UIMA framework.

Both GATE and UIMA provide pipeline-based frameworks and analysis modules. Within GATE modules are mostly Java-developed. Apache UIMA provides both Java and C++ frameworks and annotators can be written in Java, C++, Perl, Python and TCL.

Apache OpenNLP⁵ is a machine learning library that provides analysis components for many NLP tasks: language detection, text segmentation, part of speech tagging, named entity extraction, parsing and coreference resolution. It is also a Java-based toolkit initially released in 2004.

NLTK (Natural Language Toolkit)⁶ [1] is a set of Python libraries for solving natural language processing tasks. In addition to analysis modules, NLTK includes also corpora and lexical resources available through the same installer.

spaCy⁷ is another open-source Python library offering software components for text analysis. spaCy is partially implemented using Cython⁸ and authors claim that their main focus is to provide an industrial tool that is capable to operate at large scale. AllenNLP⁹ is a framework for deep learning NLP created on top of spaCy and PyTorch machine learning library.

UDPipe [20] is an open-source tool that implements NLP tasks required to reproduce Universal Dependencies 2.0 annotations: tokenization, sentence segmentation, POS tagging, lemmatization and dependency parsing. UDPipe provides both training and annotation functionality. Training part uses only Universal Dependencies annotation without any supplementary data. UDPipe is written in C++ and bindings for Python, Perl, Java and C# are provided. Several other tools able to analyze Universal Dependencies corpora has participated to CoNLL 2017 and 2018 shared task entitled *Multilingual Parsing from Raw Text to Universal Dependencies* [8], [21]. UDPipe Future is the next version of UDPipe that is under development now.

UDify [12] is a single model for analysis of 75 languages with BERT-based encoder. It uses cased BERT-Base multilingual model pretrained on Wikipedia dumps for 104 languages¹⁰. Original research paper describes different fine-tuning strategies and their effect on high-resource and low-resource languages.

DeepPavlov is an open source framework for chatbots and virtual assistants development. It includes dependency parsing module with RuBERT-based encoder [13]. RuBERT is a monolingual BERT trained on the Russian part of Wikipedia and news data. Initial weights of RuBERT were initialized with Google's multilingual BERT.

⁵ <http://opennlp.apache.org/>

⁶ <https://www.nltk.org/>

⁷ <https://spacy.io/>

⁸ Cython—a compiled language that offers better performance and memory management for Python-like code.

⁹ <https://allennlp.org/>

¹⁰ <https://github.com/google-research/bert/blob/master/multilingual.md>

2.3. LIMA

LIMA [2] is a C++ toolkit and a pipeline-based analysis framework developed by LASTI laboratory of CEA LIST. It was designed and developed with several objectives:

- multilingualism—an ability to work with a broad spectrum of languages;
- diversity of use-cases—LIMA must be useful as a basic component for various text-processing applications such as question-answering systems, automatic summarization, etc.;
- extensibility—an architecture that makes it possible to easily add new functionality or replace the implementation of existing components;
- efficiency—LIMA must be able to process large corpora and work in an industrial context.

LIMA consists of core components defining graph-based language-independent representation of entities like linguistic analysis and processing module and a collection of modules providing several types of functionality including:

- input/output modules: source text readers and analysis writers in various formats,
- dictionary-based and OOV-words annotators,
- POS tags disambiguators,
- named-entities taggers,
- rule-based syntactic analyzers.

Processing modules are executed sequentially and the order of execution is defined by pipeline configuration. All modules have access to a shared graph-based analysis representation. Input-related modules create this representation. Next modules update and enrich analysis and output modules dump generated data into target file format. A shared analysis representation allows modules to be interchangeable to the extent that this is limited by natural dependencies between analysis steps.

Previously, LIMA had rule-based and statistical analysis components implemented for three languages (English, French, and Portuguese) under a free licence and for some other languages commercially only (German, Spanish, Mandarin Chinese and Arabic) plus some experiments (Russian, Japanese, Czech...). And although the same internal representation of the analysis was built for each of the mentioned languages, the sets of tags for morphological and syntactic categories were defined separately. To support more languages we followed Universal Dependencies 2.0 tagset for parts of speech, features and syntactic dependency labels and developed fully machine learning-based modules.

3. Analysis modules based on deep neural networks

The pipeline approach in the architecture of text analyzers together with rule-based implementations of analysis components offer high flexibility in configuration and explainability of analysis results. At the same time, the development of these rule-based modules is expensive as it involves a lot of human labor. Moreover, most of the rule-based analysis components are difficult to port from one language or domain to another. Machine learning methods are easier in portability and with recent

progress in deep neural network architectures, they surpass many rule-based methods in analysis quality. Below we describe our deep neural networks based modules.

3.1. Tokenizer

For token and sentence segmentation we adapt the character labeling approach proposed in Universal segmenter [18]. It is based on bidirectional recurrent neural networks with conditional random fields (BiRNN-CRF) and Viterbi decoder (Figure 2). The tagset consists of token segmentation tags (B—begin of token, I—inside token, E—last character of token, S—single-character token, X—outside of token) and sentence segmentation tags (T—last single character token in the sentence, U—last character of the last token in the sentence).

"	А	я	д	е	р	е	в	о	о	б	н	а	р	у	ж	и	л	!	"	-	"	А	я	с	о	б	а	к	у	!	"				
S	S	X	S	X	B	I	I	I	I	E	X	B	I	I	I	I	I	E	S	T	X	S	X	S	S	X	S	X	B	I	I	I	E	S	T

Figure 1: Characters tagging schema

Three concatenated embeddings are given on each RNN step:

- the embedding of the unigram (current character only);
- the embedding of the bigram including previous character and current one;
- the embedding of the trigram that includes previous, current and next characters.

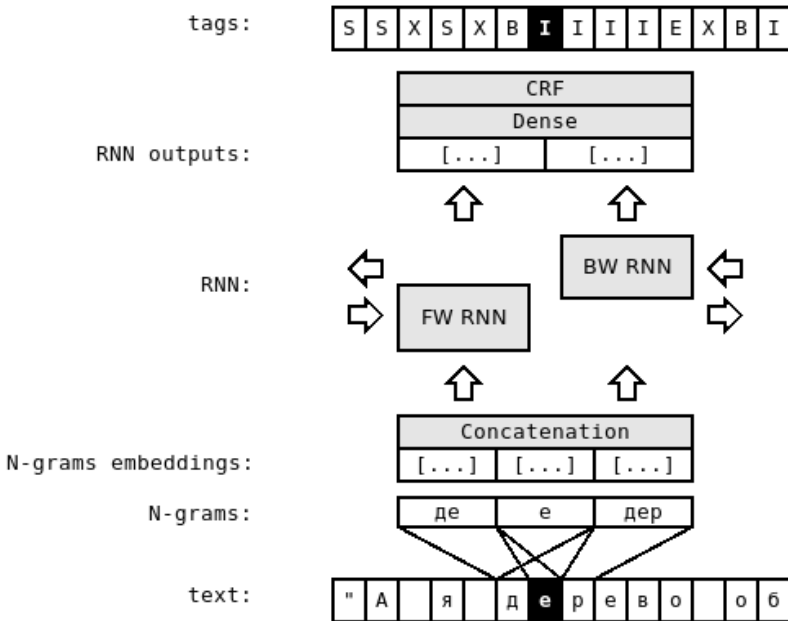


Figure 2: Tokenizer neural network

The dimension of each embedding is calculated at training time as the fourth-degree root of the number of different n-grams of given length found in the training set. This makes the model smaller for most of languages without significant quality reduction.

Table 1: N-gram embeddings dimensions for model trained on Russian-SynTagRus

	Min. frequency	Number of ngrams	Dimension
Unigram	3	153	4
Bigram	4	3,217	8
Trigram	10	12,653	12
Total			24

3.2. Morphological Tagger and Dependency Parser

The morphological tagger assigns part of speech tags and feature tags for each word in the sentence. For this purpose, we use a similar sequence labeling approach as described above for tokenization. As soon as there are many different types of tags (part of speech tags, number, gender, case, etc), a dedicated classifier is required for each type. We use single BiRNN input for all types of tags with different CRF outputs for each classifier. CRF outputs for taggers are connected to a second BiRNN layer. Remaining layers are used by the dependency parser only.

For dependency parsing, we adapted graph-based parser [11] with deep biaffine attention arcs scoring method [5]. Arc scorer is attached on top of the concatenation of the output of the same BiRNN that is used for morphological tagging and dedicated BiRNN that is used for dependency parsing only. All these tasks (i.e. morphological tagging and dependency parsing) are trained simultaneously.

The BiRNN that is shared by taggers and parser has an input that consists of pre-trained word embeddings for all words and trained word embeddings for frequent words. The sum of word embeddings of these two types is concatenated with the final state of character-level RNN for the corresponding word.

FastText word embeddings with subword information are used as pre-trained word embeddings. In fastText, model word vectors are calculated as an element-wise average of word vector and vectors of all its subwords. In case of out-of-vocabulary words the first element isn't available and subwords vectors only are used [3]. The choice of fastText instead of word2vec [15] or Glove [17] is made for two reasons: fastText provides pretrained models [7] for most languages available in the Universal Dependencies collection, and subword information gives meaningful word vectors even for OOV-words.

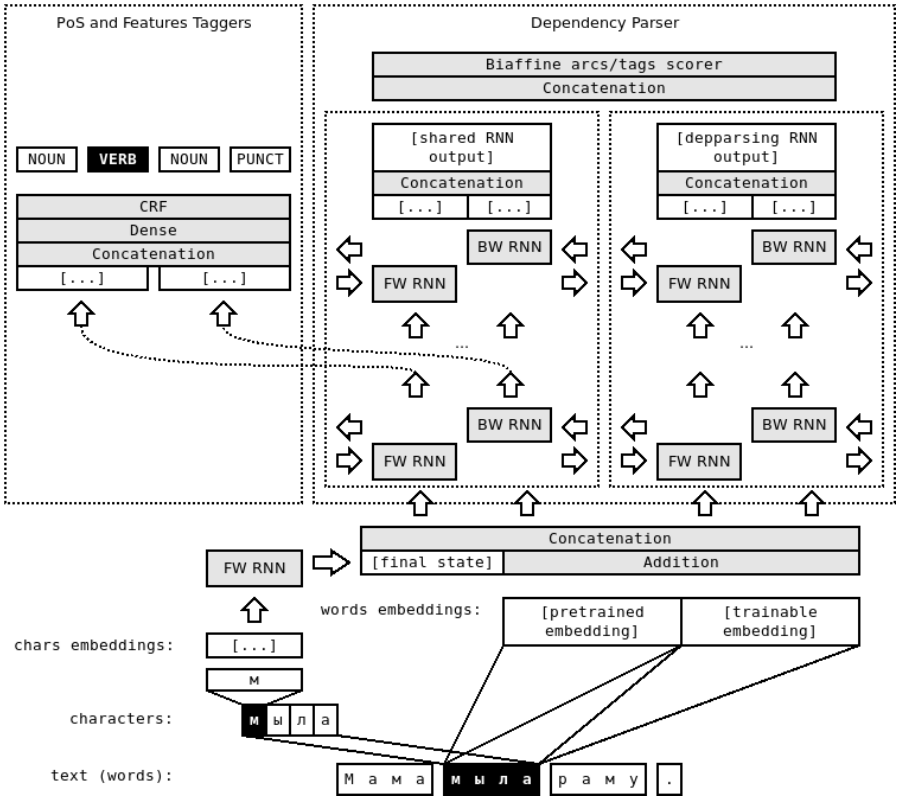


Figure 3: Taggers and dependency parser neural network

3.3. Lemmatizer

The lemmatizer uses the source form of the word and morphologic tags (part of speech and features tags) predicted on the previous step. The lemmatization task is treated as a sequence-to-sequence translation problem at the character level. Our approach is similar to the one adopted in Turku Neural Parser Pipeline [10]: surface word form together with predicted tags is given as input. Instead of encoding tags in form of strings we add them as embeddings of tags to the encoder as initial state. Seq2seq neural network with Bahdanau attention is used (Figure 4).

Lemmatization as it is described above is a context-independent task: the neighboring words aren't used to predict lemma. This makes it possible for the lemmatizer to use the dictionary of pre-lemmatized words (lemmatizer cache) to improve the analysis speed. The lemmatizer cache generated from Russian-SynTagRus training set increases test set lemmatization speed by 2.5 times.

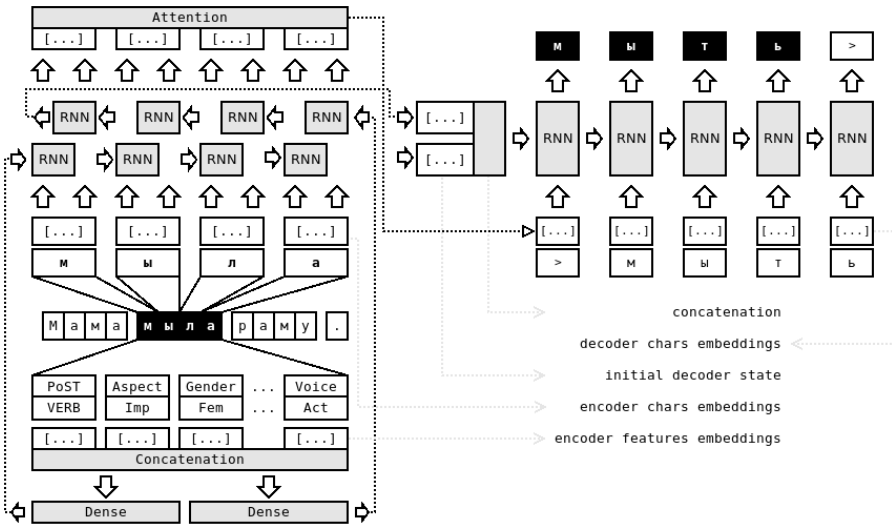


Figure 4: Lemmatizer neural network

4. Embedding Compression

The model size is an important practical issue that can limit the usage of the software on low-memory devices. The largest part of the model described above is a fast-Text embedding file that takes between 2.5 and 7.3 Gb depending on the language. Tokenizer, morphological tagger, dependency parser, and lemmatizer together take no more than 2% of the total model size. Thus the most important step toward memory footprint reduction is the compression of embeddings.

FastText embeddings file consists of four parts:

- word embedding table;
- n-grams embedding table;
- output table;
- dictionary.

The output table isn't used to calculate word embeddings and can be discarded. With remaining parts following reduction strategies are possible: dimensionality reduction, pruning and quantizing. Within the scope of this work, we have tried word and n-grams embedding tables pruning and quantizing [9].

In the case of pruning, we have removed 50% of the least frequent words and 50% of the least frequent n-grams. By default, both embedding tables and the dictionary have 2,000,000 entries each (i.e. 4,000,000 embeddings in total) and the resulting file contains 1,000,000 most frequent words and 1,000,000 most frequent n-grams.

For quantization, we used product quantization from the fastText library that consists of the approximation of real-valued vectors by the closest vector in a pre-defined set of centroids. This implementation splits each vector into several sub-vectors

(subquantizers) and maps each one to some pre-defined point. The number of these pre-defined points and the number of subquantizers are two parameters. The first one is fixed to 256 possible values (8 bits per centroid index) and the second one is variable. With $k = 1$ each real-valued element (32 bit float) of source vector is replaced by 1 byte integer value giving 4:1 compression rate. With $k = 2$ each pair of float32 values is mapped to 1 byte integer giving 8:1 compression rate. This way the embeddings file can be compressed while all entries in words and n-grams tables are preserved.

In **Table 2** we compare the effect of mentioned above compression techniques on analysis metrics. The possible combination of pruning and quantization remains for future investigation. The models we distribute include embeddings compressed with a 8:1 ratio. Embeddings with a 4:1 ratio are published separately¹¹ due to hosting size restrictions.

Table 2: The degradation of the analysis evaluation metrics with embeddings compression

	File size	UPOS	UAS	LAS
original	6.9Gb	98.34	91.87	90.20
original w/o output table	4.6Gb	98.34	91.87	90.20
50% pruning	2.7Gb	98.33 (-0.01)	91.86 (-0.01)	90.19 (-0.01)
4:1 quantization	1.2Gb	98.33 (-0.01)	91.85 (-0.02)	90.17 (-0.03)
8:1 quantization	0.6Gb	98.31 (-0.03)	91.82 (-0.05)	90.10 (-0.10)

5. Evaluation

We compare LIMA performance on Russian-SynTagRus corpus using the official CoNLL 2018 evaluation script and on GramEval-2020 [14] corpus using its official evaluation script. The use of the evaluation script from CoNLL 2018 competition is motivated by the intention to compare our results with previous works.

Both scripts evaluate part of speech tags, morphologic features, lemmatisation and syntax. Tokenization and sentence segmentation were not included into GramEval-2020 and this functionality of the analyser is evaluated only with CoNLL 2018 script on Russian-SynTagRus.

5.1. Evaluation on the Russian-SynTagRus corpus

LIMA has been evaluated in two settings: with the full analysis pipeline and with the gold tokenization and sentence segmentation. The metrics (F1 scores) are provided in **Table 3**. For comparison, we included the evaluation results of the version 1.2.0 of UDPipe (CoNLL 2018 baseline) and the results of CoNLL 2018 participants which obtained the best result on one of the evaluation metrics¹².

¹¹ <https://zenodo.org/record/3629537>

¹² CoNLL 2018 results are given according to official site. UDPipe Future evaluation metrics with gold segmentation are given according to UDify paper.

LIMA significantly outperforms UDPipe v1.2.0 in morphologic and syntactic tasks in all measurement settings but stays behind the best results of CoNLL 2018. Syntactic metrics (UAS and LAS) depends more on the quality of segmentation than part-of-speech, morphologic features and lemmata. BERT-based systems (UDify and DeepPavlov) are significantly better in syntactic metrics than all other systems.

Table 3: Performance metrics (F1) for Russian-SynTagRus corpus

	Tokens	Sentences	UPOS	UFeats	Lemmas	UAS	LAS
LIMA	99.5	98.6	98.3	96.1	96.1	91.8	90.2
UDPipe v1.2.0	99.6	98.8	97.8	93.5	96.5	87.6	85.0
CoNLL 2018 shared task results							
HIT-SCIR	99.6	98.0	98.6	93.6	95.5	93.9	92.5
NLP-Cube	99.7	98.8	98.4	96.2	92.3	92.7	90.9
Stanford	99.6	98.9	98.3	95.8	97.0	93.1	91.6
Turku NLP	99.6	98.0	98.0	96.6	98.2	93.2	91.7
UDPipe Future	99.6	98.6	98.7	97.2	97.9	93.0	91.5
Results with gold segmentation							
DeepPavlov			97.6	95.7		95.2	93.7
LIMA			98.8	96.5	94.7	92.7	91.0
UDPipe v1.2.0			98.2	93.9	96.9	88.3	85.7
UDPipe Future			99.1	97.6	98.5	93.8	92.3
UDify+Lang			99.1	97.2	96.6	95.1	93.7
Results with gold segmentation and morphology							
UDPipe v1.2.0						90.3	89.0

5.2. Evaluation on the GramEval-2020 corpus

For GramEval-2020 evaluation (Table 4), gold tokenization and sentence segmentation were given. Official evaluation scripts provide accuracy metrics for part-of-speech, morphologic features, lemmata and LAS. Overall score is an average of the four metrics mentioned above.

On GramEval-2020, metrics aren't directly comparable with metrics mentioned in previous section due to different method of calculation. Nevertheless the situation is similar: BERT-based systems (ADVance and qbic) are still better on GramEval-2020.

Table 4: Performance metrics (accuracy) for the GramEval-2020 corpus

	Overall	PoS	Feats	Lemmas	LAS
Baseline	80.4	91.0	89.6	86.4	54.5
Vocative	85.2	92.8	89.8	88.5	69.6
Lima	87.9	95.1	95.3	88.2	73.0
ADVance	90.8	95.4	95.8	93.1	78.8
qbic	91.6	95.9	96.0	93.4	81.3

5.3. Analysis speed

The analysis speed is an important characteristics of parser and it directly influences its practical applicability. There are many factors that influence analysis speed, including: analysis method, model size, underlying computation library, hardware used, compilation options etc. In the table below we provide the analysis speed comparison of several parsers. All the evaluations were performed on laptop with Intel i7-8650U CPU (4 cores, 8 threads) and 32Gb of RAM except UDPipe Future. Figures for UDPipe Future are taken from corresponding article [19]. Here are some important details regarding these measurements:

- spaCy and AllenNLP have no models for Russian language. We used models for English.
- spaCy and UDPipe run in single-thread mode. To make the comparison more meaningful we started the same analysis 8 times in parallel.
- Only dependency parsing¹³ without morphologic features have been tested for DeepPavlov.
- Although CPU we used for these experiments is capable to run up to 8 threads simultaneously, not all analyzers used all of them all the time.
- UDPipe v1.2.0 and TensorFlow library used by LIMA were compiled with “-march = native” gcc compilation option. All other software have been installed from official repositories without compilation.

Table 5: Analysis speed comparison on CPU

	Method	Threads	Speed (tok/sec)
spaCy (en_core_web_sm)	transition-based	single	2,728
spaCy (en_core_web_sm)	transition-based	multi	9,043
UDPipe v1.2.0	transition-based	single	3,000
UDPipe v1.2.0	transition-based	multi	12,000
AllenNLP (English)	RNN + graph-based	multi	250
LIMA (with cached lemmata)	RNN + graph-based	multi	430
LIMA (w/o cached lemmata)	RNN + graph-based	multi	327
UDPipe Future	RNN + graph-based	multi	517
DeepPavlov	BERT + graph-based	multi	166
UDify	BERT + graph-based	multi	108

From these data the difference in analysis speed between transition-based and graph-based parsers is clearly seen. LIMA shows an average speed comparable to other implementations of the same method.

6. Conclusion and future works

Our first implementation of neural network-based modules for LIMA obtains results at the level of the best systems that participated to CoNLL-2018. Current results

¹³ The model name is “syntax_ru_syntagrus_bert”.

in GramEval-2020 show that transformers-based models are necessary to reach today's state of the art. Anyway, LIMA is readily available for users, easily installable using simple packages.

Our future work include using BERT-like models to reach state of the art performance but also working on analysis speed to make LIMA neural network modules usable in production settings.

References

1. *Bird, S. et al.*: Natural language processing with python. (2009).
2. *Besaçon, R. et al.*: LIMA: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. Presented at the May (2010).
3. *Bojanowski, P. et al.*: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606. (2016).
4. *Cunningham, H. et al.*: GATE: A framework and graphical development environment for robust nlp tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). (2002).
5. *Dozat, T., Manning, C. D.*: Deep biaffine attention for neural dependency parsing. ArXiv. abs/1611.01734, (2016).
6. *Ferrucci, D. et al.*: Unstructured information management architecture (UIMA) version 1.0, <https://docs.oasis-open.org/uima/v1.0/uima-v1.0.html>, (2009).
7. *Grave, E. et al.*: Learning word vectors for 157 languages. In: Proceedings of the international conference on language resources and evaluation (Irec 2018). (2018).
8. *Hajič, J., Zeman, D. eds.*: Proceedings of the conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. Association for Computational Linguistics, Vancouver, Canada (2017).
9. *Joulin, A. et al.*: FastText.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651. (2016).
10. *Kanerva, J. et al.*: Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In: Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. pp. 133–142 Association for Computational Linguistics, Brussels, Belgium (2018).
11. *Kiperwasser, E., Goldberg, Y.*: Simple and accurate dependency parsing using bidirectional lstm feature representations. Transactions of the Association for Computational Linguistics. 4, 313–327 (2016).
12. *Kondratyuk, D., Straka, M.*: 75 languages, 1 model: Parsing universal dependencies universally. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). pp. 2779–2795 Association for Computational Linguistics, Hong Kong, China (2019).
13. *Kuratov, Y., Arkipov, M.*: Adaptation of deep bidirectional multilingual transformers for russian language, (2019).
14. *Lyashevskaya, O. et al.*: GramEval 2020 Shared Task: Russian Full Morphology and Dependency Parsing. In: Computational linguistics and intellectual technologies: Papers from the annual conference “dialogue”. (2020).

15. *Mikolov, T. et al.*: Efficient estimation of word representations in vector space. CoRR. abs/1301.3781, (2013).
16. *Nivre, J. et al.*: Universal dependencies v1: A multilingual treebank collection. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16). pp. 1659–1666 European Language Resources Association (ELRA), Portorož, Slovenia (2016).
17. *Pennington, J. et al.*: GloVe: Global vectors for word representation. In: Empirical methods in natural language processing (emnlp). pp. 1532–1543 (2014).
18. *Shao, Y. et al.*: Universal word segmentation: Implementation and interpretation. Transactions of the Association for Computational Linguistics. 6, 421–435 (2018).
19. *Straka, M.*: UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. pp. 197–207 Association for Computational Linguistics, Brussels, Belgium (2018).
20. *Straka, M., Strakova, J.*: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. Presented at the January (2017).
21. *Zeman, D., Hajič, J. eds*: Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. Association for Computational Linguistics, Brussels, Belgium (2018).

FULL-FLEDGED SEMANTIC ANALYSIS AS A TOOL FOR RESOLVING TRIANGLE-COPA SOCIAL SCENARIOS

Boguslavsky I. M. (bogus@iitp.ru)

A. A. Kharkevich Institute for Information Transmission
Problems, Russian Academy of Sciences, Moscow, Russia;
Universidad Politécnica de Madrid, Madrid, Spain

Dikonov V. G. (sdiconov@mail.ru),

Frolova T. I. (tfrolova@gmail.com),

Iomdin L. L. (iomdin@gmail.com),

Lazursky A. V. (lazursky@mail.ru),

Rygaev I. P. (irygaev@gmail.com),

Timoshenko S. P. (nyrestein@gmail.com)

A. A. Kharkevich Institute for Information Transmission
Problems, Russian Academy of Sciences, Moscow, Russia

Text interpretation often requires common sense knowledge and reasoning. A convenient tool for developing methods of common sense reasoning are special sets of challenge problems whose interpretation requires sophisticated reasoning. An interesting example is a recently published data set called Triangle Choice of Plausible Alternatives (Triangle-COPA), which contains 100 multiple-choice problems that test the interpretation of social scenarios. Each problem includes a statement and two alternatives. The task is to identify the more plausible alternative. For processing Triangle-COPA data we use SemETAP, a general purpose semantic analyzer. We implement the full scenario of NL understanding starting from NL texts and not from manually composed simplified logical formulas, which is a common practice in logic-based approaches to common sense reasoning. We produce Enhanced Semantic Structures of the statement and both alternatives and check which alternative manifests more semantic agreement with the statement in terms of inferences.

Keywords: Triangle-COPA, common sense reasoning, knowledge-based approach, knowledge representation, inference, Etalog language, SemETAP

DOI: 10.28995/2075-7182-2020-19-106-118

ОДИН ПОДХОД К ОПИСАНИЮ СОЦИАЛЬНЫХ СЦЕНАРИЕВ TRIANGLE-COPA НА ОСНОВЕ СЕМАНТИЧЕСКОГО АНАЛИЗА

Интерпретация текста часто требует обращения к фоновым знаниям, основанным на логике здравого смысла, и умозаключений на основе этих знаний. Удобным инструментом для разработки методов обращения с такими знаниями являются специально составленные вопросы, ответ на которые требует определенных умозаключений. Интересным примером такого набора вопросов является недавно опубликованный набор под названием Triangle Choice of Plausible Alternatives (Triangle-COPA). Он содержит 100 заданий с многовариантными ответами, требующих интерпретации социальных сценариев. Каждое задание содержит утверждение и две связанных с ним альтернативы. Задача состоит в идентификации более правдоподобной альтернативы. Для работы с этими заданиями мы используем семантический анализатор общего назначения SemETAP. Для систем, стремящихся делать логические умозаключения, основанные на здравом смысле, типично, что они оперируют с логическими формулами, составленными вручную. В отличие от этого, наш анализатор получает на вход текст на естественном языке и реализует полный сценарий обработки этого текста. Мы строим Расширенную семантическую структуру как утверждения, содержащегося в задании, так и обеих предложенных альтернатив и стремимся определить, какая из альтернатив обнаруживает большее семантическое согласование с утверждением с точки зрения возможных умозаключений.

1. Introduction

As is known, computer text understanding includes, on the one hand, linguistic analysis of the text, and on the other hand, a large area of logical and common sense reasoning. These two lines of research are pursued to a large extent independently of one another. In most cases, text analysis is performed by means of machine learning techniques and does not include logical processing of the representation obtained (cf., for example, [Ge and Mooney 2005], [Bos 2008], [Poon, Domingos 2009]). Logical and common sense reasoning, as a rule, applies to some logical form, usually hand-authored [Mueller 2006]. Artificial Intelligence researchers who model human commonsense reasoning seek to address the following fundamental tasks: represent commonsense knowledge as formal theories for use in automated reasoning systems [Gordon et al. 2011], [Hobbs et al. 2020], [Hobbs et al. 2012], [Montazeri et al. 2011], [Montazeri et al. 2012], [Morgenstern 2001], design cognitive architectures for commonsense reasoning [Meadows et al. 2014], and evaluate automatic reasoning systems [Levesque et al. 2012]; [Roemelle et al. 2012]. There are very few integrated approaches that take an NL text at the input, construct some sort of semantic representation thereof and perform a large spectrum of logical and common sense inferences.

A notable exception is a series of publications in the framework of Ontological Semantics initiated by [Nirenburg, Raskin 2004]. Our semantic analyzer [Boguslavsky 2011], [2017], [Boguslavsky et al. 2015], [2018], [2019] is being developed in the same vein.

Over the last years, two evaluation tools have been proposed to spur progress in automated commonsense reasoning. The first one, the Choice of Plausible Alternatives (COPA), tests commonsense causal reasoning in everyday situations [Roemelle et al. 2011]. It consists of one thousand binary-choice questions; and the task is to select the more plausible cause or result of a given situation. The second approach is the Winograd Schema Challenge (WSC) [Levesque et al. 2011]. It formulates the commonsense reasoning problem as a reference resolution task. Each question consists of a pair of sentences that differ in only one or two words, which switch the way the reference should be resolved. However, neither of these evaluation tools has been instrumental in advancing logical formalization of commonsense reasoning. In both cases, the best published results come from systems based on machine learning techniques.

Later, a new set of challenge problems has been proposed, called Triangle-COPA [Maslan et al. 2015]. As opposed to previous challenge problems, it was not conceived as an evaluation tool. It was specifically designed as a development test set, rather than a held-out test set for use in competitive evaluations, which means that it was intended to aid in the logical formalization of commonsense knowledge needed to correctly answer the questions. The approach adopted in Triangle-COPA differs from that seen in COPA and WSC in two important ways. First, Triangle-COPA test contains not only NL (English) questions and answers, but also their first-order logical form. This eliminates the need for sophisticated NL processing. Second, the domain of situations described in the questions was strongly constrained. This allowed researchers to concentrate their efforts on specific areas of commonsense reasoning.

Below, we will briefly describe the Triangle-COPA test set and review the existing solutions (section 2). Then we will present our approach (section 3) and describe our experiments (section 4). We will conclude in section 5.

2. Triangle-COPA

The Triangle Choice of Plausible Alternatives (Triangle-COPA) is a set of 100 challenge problems for logical formalization of commonsense psychology. All the texts describe events occurring in a simple environment whose participants are characters of a short film authored by the social psychologist Fritz Heider [Heider and Simmel 1944]. They present some brief sequence of actions and interactions between three geometrical shapes. Two triangles and a circle perform various actions in and around a room with a door. Each text consists of a statement and two alternatives connected with the statement, where one of these two is much more plausible than the other and provides a better explanation of the statement. The task is to computationally determine the correct alternative. An example of a Triangle-COPA challenge problem is as follows:

(1) Text 83

Statement: A small triangle and a big triangle are next to each other. A circle runs by and pushes the small triangle. The big triangle chases the circle. Why does the big triangle chase the circle?

Logical form: (and (approach' e1 C LT) (push' e2 C LT) (chase' e3 BT C) (seq e1 e2 e3))

Alternative 1. The big triangle is angry that the circle pushed the small triangle, so it tries to catch the circle.

Logical form: (angryAt' e4 BT C)

Alternative 2. The big triangle and circle are friends. The big triangle wants to say hello to the circle.

Logical form: (and (friend' e5 BT C) (goal' e6 e7 BT) (greet' e7 BT C)).

Notation: e1, e2,...en—events; C—Circle, BT—Big Triangle, LT—Little Triangle.

We are aware of two approaches to Triangle-COPA test published in [Gordon 2016] and [Kalluri et al. 2017].

[Gordon 2016] uses logical and probabilistic reasoning. Commonsense knowledge of actions, social relationships, intentions, and emotions are encoded as defeasible axioms in first-order logic. For example, axioms (2a)–(2d) provide possible reasons why one character would be chasing another:

- (2) (a) Chase 1: Maybe they are playing tag
 (if (and (playWith' e1 x y)
 (etcChase1 0.2 e1 x y)
 (chase' e x y))
- (b) Chase 2: Maybe one is angry at the other
 (if (and (angryAt' e1 x y)
 (etcChase2 0.2 e1 x y)
 (chase' e x y))
- (c) Chase 3: Maybe one is trying to rob the other
 (if (and (goal' e1 e2 x)
 (rob' e2 x y)
 (etcChase3 0.3 e1 e2 x y)
 (chase' e x y))
- (d) Chase 4: Maybe one is trying to scare the other
 (if (and (goal' e1 e2 x)
 (afraid' e2 y)
 (etcChase4 0.5 e1 e2 x y)
 (chase' e x y))

Axiom (2a) claims that if x plays with y , it is probable (with the 0.2 probability) that x chases y . Axiom (2b) says that chasing is also probable, and also with the 0.2 probability, if x is angry at y . With somewhat higher probability (0.3) x may chase y , if it wants to rob it (Axiom (2c)). Finally, x will chase y with the maximal probability (0.5), if x wants to scare y .

Using these axioms, possible interpretations of action sentences are identified via logical abduction (called Etcetera Abduction), backchaining to distinct sets of assumptions that logically entail the observations. These sets of assumptions are ordered by their joint probability assuming conditional independence. The axioms, 136 in total, describe commonsense knowledge in the domain of the Triangle-COPA question set. They are oriented towards specific inferences needed to infer the correct alternative for each question. Abduction, as distinct from logical deduction or induction, is a form of logical reasoning that identifies a hypothesis that, if it were true, would logically entail the given input.

In order to calculate the joint probability of events, each axiom should contain a numerical value, which characterizes the probability of the etcetera literal. Although it is supposed that in the future the probabilities will be extracted automatically out of empirical data, in the current version they are selected intuitively, which does not always look convincing. For example, our intuition does not confirm the estimation of Axiom (2d), which maintains that, if we wish to scare somebody, 50% chances are that we will chase them.

An important advantage of the approach is that the knowledge used is explicitly formulated in the form of axioms with the probabilities. Due to this, Etcetera Abduction provides an explanation of why one solution has been preferred over a different one. On the other hand, one should not disregard the fact that the axioms on which [Gordon 2016] draws were formulated specifically for Triangle-COPA questions. Judging by the examples given in Gordon’s paper, these axioms are very fragmentary and do not seek to make integral part of some balanced knowledge repository suitable for a wide range of scenarios. However, this is not a defect in the context of the paper. The goal of Gordon 2016 is to develop a new variant of abduction building on a priori probabilities and to show that, if common sense axioms are available, the method works well. As [Gordon 2016] states, “our strong performance on the Triangle-COPA test demonstrates that Etcetera Abduction is a viable approach, but our success owes much to the labor of hand-crafting the axioms necessary to solve these specific questions”.

[Kalluri et al. 2017] builds a model that deploys Bayesian inference with an action affinity lexicon to infer probabilistic affinity relations characterizing the scenario. Subsequently, the model uses the inferred affinity relations to choose the more probable alternative of two variants. As opposed to [Gordon 2016], this model does not use logical axioms nor absolute prior probabilities of events—two important components of [Gordon 2016] approach. The key concept on which [Kalluri et al. 2017] relies is the concept of affinity. This is the relation between two agents that may take one of three values—Pleasant/Unpleasant/Neutral. All the knowledge the model builds on is the static probabilistic action affinity lexicon, which links actions to the relative observation distribution of affinities. For example, action `argue_with` is assigned 0.5 Unpleasant affinity, 0.25 Pleasant and 0.25 Neutral. The approach of [Kalluri et al. 2017] does not advance much commonsense reasoning. However, it is interesting in that it shows that relatively much can be obtained if one disposes of very lightweight knowledge.

3. Our approach

We are taking quite a different stance. Of course, one can set a specific goal of solving exclusively the Triangle-COPA problem proper and, if possible, by minimal means, as it seems to be the case in [Kalluri et al. 2017]. In this case, we could be quite justified to use any kind of knowledge, however individual and case-oriented it may be, if it helps solve the problem.

We look at the problem from a different perspective. We perceive it as a particular case of a much more general task—the task of text understanding. In this vein, we have been developing a knowledge-based general semantic analyzer for Russian—SemETAP. In previous publications mentioned above, we described its various aspects and will not repeat them here. Let it only be reminded that:

- SemETAP is an option of the ETAP-4 linguistic processor and reuses its non-semantic modules (morphological analysis, syntactic dependency parsing, and normalization).
- Semantic analysis makes use of linguistic data and extralinguistic information (background knowledge). The linguistic data are provided by the Combinatorial Dictionary and the Grammar, and the background knowledge is stored in the Ontology, Repository of Individuals and the set of inference rules SemRule.
- Inference rules is a crucial component of SemETAP. We believe that the depth of understanding is growing with the number of inferences we can draw from the text. In many cases, a decomposition of the concept meaning helps produce additional inferences and thus achieve a deeper understanding.
- Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological concepts. Enhanced semantic structure (EnSemS) extends BSemS by means of a series of inferences. For the purposes of this paper, EnSemS is the most important representation, since it makes explicit all the inferences that the knowledge available permits to make from the text and the context.
- Two types of inferences are carried out: 100%-true logical entailments, and implications that implement plausible expectations.

This approach determines several distinctive features of the way we handle Triangle-COPA texts.

First, we use them as a convenient tool of formulating common sense knowledge and checking its validity. Our priority is not so much correctly answering concrete Triangle-COPA questions, but rather incorporating new knowledge into an integrated NL understanding system and testing it.

Second, in processing Triangle-COPA, we implement the full scenario of NL understanding starting from NL texts and not from manually composed simplified logical forms. This is more difficult than accepting logical forms at the input, not only because additional stages of processing are needed but also because logical forms proposed by [Maslan et al. 2015] often oversimplify the initial English text. One will clearly see that if one compares the NL text and the logical form of Alternative 1 of the text 83 above. Since our priority is developing the semantic analyzer for Russian and checking its capacity against Triangle-COPA statements, we had to translate these statements into Russian.

Third, the knowledge we acquire, both linguistic and common sense, should not be specific for the Triangle-COPA texts. We do not introduce ad hoc knowledge. The SemETAP general semantic analyzer needs knowledge that is supposed to be valid for a wide range of situations and for different purposes.

Fourth, based on this general knowledge SemETAP performs a large number of inferences, both strict and plausible, which form the Enhanced Semantic Structure (EnSemS). Such a semantic structure permits to restore implicit semantic elements and answer questions even when the answer is not directly contained in the text.

Fifth, we resolve Triangle-COPA tests as follows. For each text, we obtain EnSemSs of the Statement and both Alternatives. Then, a special algorithm selects the alternative whose EnSemS manifests higher semantic agreement with the EnSemS of the Statement. This algorithm will be explained below.

The problem of knowledge acquisition and its generality vs. specificity is worthy of some more comments, since this is where our approach differs from many other approaches to common sense reasoning. As mentioned above, our model of NL understanding proceeds from an assumption that the more inferences we can draw from the text, the deeper our understanding is. Therefore, we aim at making the maximum amount of inferences. These include both inferences motivated by the lexical meaning of words, and inferences based on world knowledge. To make such inferences, the system has to possess a large amount of knowledge. As is known, the critical bottleneck on this way is knowledge acquisition, particularly as far as general commonsense knowledge is concerned.

Many knowledge-supported projects make use of existing resources, such as WordNet, FrameNet, FreeBase, ConceptNet or YAGO [Ponzetto and Strube 2006], [2007], [Poesio et al. 2007], [Bryl et al. 2010], [Rahman and Ng 2011], [Uryupina et al. 2011], [Lee et al. 2011]. However, these resources do not provide information needed for the inferences we strive to make. Therefore, we have to acquire by ourselves most of the knowledge we need. Since common sense knowledge acquisition is an infinitely huge task, one can only advance by small incremental steps. The critical requirement is that acquired knowledge be not oriented strictly to a particular task for which it was acquired, in our case—Triangle-COPA test. It should retain its value in a wider context and for different tasks. Only in this case one could count on the incremental growth and stability of the knowledge repository. That is why we assign special importance to making knowledge as general as possible. What follows is an example illustrating this point.

Let us recall axioms (2a)–(2d) from [Gordon 2016], which describe situations of one person chasing another. We will focus on the Axiom (2b), which proclaims a direct link between the concepts AngryAt and Chase. It reads that if x is angry at y , then x will probably chase him. We conjecture that this axiom was introduced in order to resolve question 83 of the Triangle-COPA set, cited above but reproduced below for convenience:

Text 83

Statement: A small triangle and big triangle are next to each other. A circle runs by and pushes the small triangle. The big triangle chases the circle. Why does the big triangle chase the circle?

Logical form: (and (approach' e1 C LT) (push' e2 C LT) (chase' e3 BT C) (seq e1 e2 e3))

Alternative 1. The big triangle is angry that the circle pushed the small triangle, so it tries to catch the circle.

Logical form: (angryAt' e4 BT C)

Alternative 2. The big triangle and circle are friends. The big triangle wants to say hello to the circle.

Logical form: (and (friend' e5 BT C) (goal' e6 e7 BT) (greet' e7 BT C)).

To answer question 83, axiom (2b) is sufficient, but if the knowledge repository is acquired for a wider use, it should contain descriptions that give more information about the concepts and link them in a more general way. In our model, we try to include in semantic definitions all components that may be needed for inference, although of course our definitions do not pretend to be exhaustive. By way of example, we show below how concepts Chase and Anger are described in SemETAP. For readers' convenience, we represent these definitions by means of English glosses, and not in the formal Etalog language, in which they were originally written.

- (3) Chase: 'Agent **?chaser** is moving towards Agent **?victim** having the goal of coming in contact with him and perform upon him an action **?harm**, which is assessed negatively. **?victim** is moving away from **?chaser** having the goal of avoiding coming in contact with him'.
- (4) Anger: 'Agent **?experiencer** is in an intensive negative emotional state towards **?stimulus**', which may be accompanied by shouting or frowning on the part of **?experiencer**; if **?stimulus** is a living being, **?experiencer** may perform upon him an action **?harm**, which **?stimulus** assesses negatively; if **?stimulus** is an event, **?experiencer** probably feels Anger with respect to the initiator of event **?stimulus**; this event takes place both when Anger takes place and not'.

Comments:

- 1) The concept Agent used in phrases like 'Agent ?experiencer' above does not refer to the **semantic role** Agent. It means that the domain of variable ?experiencer is the Agent **ontological class**, which includes humans, organizations, etc.
- 2) The last proposition of definition (4) describes the fact that Anger is a factive predicate. This means that both (a) *Mary is angry that no one agreed with her suggestion* and (b) *Mary is not angry that no one agreed with her suggestion* imply that no one agreed with Mary's suggestion.

Definitions (3) and (4) allow for many inferences. For example, if we have the text

- (5) *The dog chased the rabbit*

we can answer a range of questions, such as: *Was the dog moving?* (yes). *What was the destination of its movement?* (the rabbit). *What was its goal?* (to get in contact with the rabbit). *What can be expected if the dog gets in contact with the rabbit?* (it will do something the rabbit will not like). *What was the goal of the rabbit?* (to avoid coming in contact with the dog). If sentence (5) is followed by something like *...but did not succeed*, we can safely infer that the dog did not get in contact with the rabbit.

Turning to text 83, we can see that definitions (3) and (4) can help discover a semantic agreement between the Statement and the Alternative 1. On the one hand, (3) implies that, when chasing, the big triangle wants to do something bad to the circle. On the other hand, (4) states that, being angry with the circle, the big triangle will also possibly do something bad to him.

Since no semantic agreement can be found between the Statement and the Alternative 2, the first one should be preferred. In more detail, the algorithm of checking semantic agreement will be discussed below.

4. Experiment

In order to clearly estimate the value of Triangle-COPA as a development set, we split our work in two stages.

At the first stage, we compiled the list of words used in the Triangle-COPA texts and provided their full description. This list contains about 300 elements, mostly describing actions, mental and emotional states, social relationships, and intentions. This description consists of a linguistic part and a semantic-ontological part. The former part includes syntactic and semantic features, government pattern, lexical functions and some other types of information, which are not specific for the semantic analyzer. The second part contains the semantic-ontological correlate of the given word (if it exists), correspondence between syntactic and semantic relations, and inference rules, which in their turn include semantic decomposition of concepts and other axioms. Importantly, at the first stage we did not consult the Triangle-COPA texts and did not know what kind of common sense axioms are needed for them.

At the second stage, we used the Triangle-COPA as a development set. Concept descriptions compiled at the previous stage were supplemented by new axioms. In doing that, we took care not to construct ad hoc axioms. Essentially, using the Triangle-COPA as a development set helps incorporate into the knowledge repository the axioms that escaped our notice before. This is how we introduced such axioms as “If a person touches another person, he/she probably wishes to attract his/her attention” or “If a person has a low estimation of some situation that takes place, he/she wants it to stop”.

After terminating each stage of knowledge preparation, we performed an experiment on selecting alternatives on the basis of this knowledge. The selection algorithm and the results of the experiment are presented below.

The algorithm starts with two EnSemSs (of the statement and of one of the alternatives) and calculates a degree of semantic similarity between them. For the purpose of the algorithm there is no difference which EnSemS represents the statement and which one the explanation hypothesis. They are treated symmetrically to find out how many propositions can be unified between the structures.

First, we identify all nodes corresponding to events marked as true facts or plausible expectations in each semantic structure and make two lists of such event nodes. Then we take a list with fewer elements and try to match each element from this list to a corresponding element in the other list. Nodes are matched in such a way that the propositions they represent are unifiable, i. e. they do not contradict each other on the assumption that they describe the same event.

Rules for unification check are the following:

- A node corresponding to an individual constant from the ontology is unifiable only to a node corresponding to the same constant.
- Nodes of different ontological classes are unifiable only if one class is a subclass of another.
- Event nodes are unifiable only if they do not contain non-unifiable arguments. By the argument here we mean any property defined in the ontology as functional, i. e. a property that can have only one value, not multiple values.

We take each functional property of the first event in the pair and check if the second event either does not have this property or their values are unifiable by the same algorithm.

An event from one list can potentially match (be unifiable) to multiple events in the other list. In such case, we select the best match by calculating a maximum number of matched functional properties, i. e. properties presented in both events and unified.

Once the list of matched pairs is produced we calculate a number of metrics:

- M1—purely a number of matched events;
- M2—M1 divided by the minimum length of the two lists;
- M3—M1 divided by the average length of the two lists.

The ideas behind M2 and M3 are the following. The minimum length of the two lists is the maximum number of events that could match in principle. For example, if there are 4 events in the first list and 20 events in the second, then no more than 4 events could match. M2 calculates the ratio of actually matched events to this maximum. But it is insensitive to the number of events in the bigger list. M2 would not change if the second list contained 120 events instead of 20. To account for that M3 was introduced. It is sensitive to the number of events in each list. Increasing either list without increasing matches will produce a smaller M3.

Once metrics for two alternative hypotheses are calculated, we select the hypothesis with a higher metric value. Our tests showed that M2 is the best metric, so we omit the other ones in the resulting table.

At the first stage the number of correctly predicted answers was not very high—53.5%, but at the second stage we made a significant improvement reaching 80,5% of correct predictions. The system produced 76 correct answers, 15 incorrect ones, and 9 random choices.

The analysis of the experiment results shows that there are three main reasons for erroneous selection.

First, in some cases our knowledge is still insufficient for the selection.

Second, in some cases, knowledge needed to resolve alternatives was not considered general enough to be introduced to the knowledge repository. An example is

Text 68:

Statement: A big triangle, small triangle, and circle are in the house. The big triangle and the circle each kiss the small triangle, wave, and then leave the room. How are the shapes related?

Alternative 1. The big triangle and the circle are parents of the small triangle.

Alternative 2. The big triangle and the small triangle are parents of the circle.

Third, an approximate manual estimation showed a better result than the real experiment run on the computer. This may suggest that the data needed for the correct selection are available in EnSemS, but the algorithm of the alternative selection may need some update, which we hope to perform in the near future.

5. Conclusion

- We obtained a reasonably good result on the Triangle-COPA test: 80.5% of questions were answered correctly. Our method implements the full scenario of text understanding, which includes morpho-syntactic and semantic analysis and a series of inferences. As opposed to other Triangle-COPA solutions, where the processing was performed on manually prepared simplified logical formulae, our analyzer applies to raw text.
- The choice of the correct alternative was done in two steps. First, Enhanced Semantic Structures of the texts to be compared were performed. These structures contain all the inferences, both strict and plausible, that can be made based on the knowledge available. Then semantic agreement between these structures was calculated, and the alternative showing higher agreement with the initial text was chosen. Importantly, the agreement is formulated in intuitively clear terms and can be checked by humans.
- Satisfactory results obtained on Triangle-COPA prove that a general scope semantic analyzer can solve specific problems, provided it is supplied with good-quality knowledge. Explicit knowledge based on the concepts meaning and common sense knowledge plays a key role.
- The knowledge repository of the semantic analyzer was incrementally enriched. The knowledge acquired during the Triangle-COPA experiments is not task-specific and can be employable in multiple scenarios.
- An important result is the discovery that common sense knowledge needed for inferences even in a very narrow domain cannot be acquired without detailed examination of the texts of the domain.

6. Acknowledgements

This work was supported by the RSF grant No. 16-18-10422-P, which is gratefully acknowledged.

References

1. *Boguslavsky I.* (2011) “Semantic Analysis Based on Linguistic and Ontological Resources”, Proceedings of the 5th International Conference on the Meaning—Text Theory. Barcelona. Igor Boguslavsky and Leo Wanner (Eds.). P. 25–36.
2. *Boguslavsky I., V. Dikonov, L. Iomdin, A. Lazursky, V. Sizov, S. Timoshenko.* (2015), Semantic Analysis and Question Answering: a System Under Development. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2015), p. 62.

3. *Boguslavsky I.* (2017), Semantic Descriptions for a Text Understanding System. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2017), p. 14–28.
4. *Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S.* (2018), Semantic analysis with inference: high spots of the football match. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”, Moscow, May 30—June 2.
5. *Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S.* (2019), Knowledge-based approach to Winograd Schema Challenge. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. Moscow, May 29—June 1.
6. *Bos J.* (2008), Wide-Coverage Semantic Analysis with Boxer. In Semantics in Text Processing. STEP 2008 Conference Proceedings. College Publications. pp. 277–286.
7. *Bryl V., Guiliano C., Serafini L., Tymoshenko K.* (2010), Using background knowledge to support coreference resolution. In Proceedings of the 19th European Conference on Artificial Intelligence, pp. 759–764.
8. *Ge R., Mooney R.* (2005), A Statistical Semantic Parser that Integrates Syntax and Semantics. Proceedings of the Ninth Conference on Computational Natural Language Learning. Ann Arbor, MI, pp. 9–16, June 2005.
9. *Gordon, A. S.* (2016), Commonsense Interpretation of Triangle Behavior, In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16).
10. *Gordon A. S., Hobbs J. R.* (2011), A Commonsense Theory of Mind-Body Interaction, Proceedings of the 10th Symposium on Logical Formalizations of Commonsense Reasoning, AAAI Spring Symposium Series.
11. *Heider F., Simmel M.* (1944), An experimental study of apparent behavior. The American Journal of Psychology 57(2):243–259.
12. *Hobbs J. R., Gordon A.* (2010), Goals in a Formal Theory of Commonsense Psychology, In A. Galton and R. Mizoguchi (eds.), Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010), IOS Press, Amsterdam, pp. 59–72.
13. *Hobbs J. R., Sagae A., Wertheim S.* (2012), Toward a Commonsense Theory of Microsociology: Interpersonal Relationships, In M. Donnelly and G. Guizzardi (eds.), Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012), IOS Press, Amsterdam, Netherlands, pp. 249–262.
14. *Kalluri P., Gervás P.* (2017), Affinity-based Interpretation of Triangle Social Scenarios. In Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART 2017), pp. 640–647.
15. *Lee H., Peirsman Y., Chang A., Chambers N., Surdeanu M., Jurafsky D.* (2011), Stanford’s multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pp. 28–34.
16. *Levesque H., Davis E., Morgenstern L.* (2012), The Winograd schema challenge. In Proceedings of the Knowledge Representation and Reasoning Conference.

17. *Maslan N., Roemmele M., Gordon A. S.* (2015), One hundred challenge problems for logical formalizations of commonsense psychology. In Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense-2015).
18. *Meadows B., Langley, P., Emery, M.* (2014), An abductive approach to understanding social interaction. *Advances in Cognitive Systems* 3:87–106.
19. *Montazeri N., Hobbs J. R.* (2011), Elaborating a Knowledge Base for Deep Lexical Semantics, In J. Bos and S. Pulman (eds.), *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, January 2011, pp. 195–204.
20. *Montazeri N., Hobbs J. R.* (2012), Axiomatizing Change-of-State Words, In M. Donnelly and G. Guizzardi (eds.), *Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012)*, IOS Press, Amsterdam, Netherlands, pp. 221–234.
21. *Morgenstern L.* (2001), MidSized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking. *Studia Logica*, April 2001, Volume 67, Issue 3, pp. 333–384.
22. *Mueller E.* (2006), *Commonsense reasoning*. Elsevier, Morgan Kaufmann Publishers.
23. *Nirenburg, S., Raskin, V.* (2004), *Ontological Semantics*. The MIT Press. Cambridge, Mass., London, England.
24. *Poesio M., Day D., Artstein R., Duncan J., Eidelman V., Giuliano C., Hall R., Hitzeman J., Jern A., Kabadjov M., Yong Wai Keong, Mann G., Moschitti A., Ponzetto S., Smith J., Steinberger J., Strube M., Jian Su, Versley Y., Xiaofeng Yang, Wick M.* (2007), ELERFED: Final report of the research group on Exploiting Lexical and Encyclopedic Resources For Entity Disambiguation. Technical report, Summer Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.
25. *Ponzetto S. P., Strube M.* (2006), Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference and Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 192–199.
26. *Ponzetto S. P., Strube M.* (2007), Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.
27. *Poon, H., & Domingos, P.* (2009), Unsupervised semantic parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Volume 1 EMNLP 09 (p.1).
28. *Rahman A., Ng V.* (2011), Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 814–824.
29. *Roemmele M., Bejan, C. Gordon A.* (2012), Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the AAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.
30. *Uryupina O., Poesio M., Giuliano C., Tymoshenko K.* (2011), Disambiguation and filtering methods in using Web knowledge for coreference resolution. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*, pp. 317–322.

GENERATING TRAINING DATA FOR WORD SENSE DISAMBIGUATION IN RUSSIAN

Bolshina A. S. (angelina_ku@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

Loukachevitch N. V. (louk_nat@mail.ru)

Kazan Federal University, Kazan, Russia

The best approaches in Word Sense Disambiguation (WSD) are supervised and rely on large amounts of hand-labelled data, which is not always available and costly to create. For the Russian language there is no sense-tagged resource of the size sufficient to train supervised word sense disambiguation algorithms. In our work we describe an approach that is used to create an automatically labelled collection based on the monosemous relatives (related unambiguous entries). The main contribution of our work is that we extracted monosemous relatives that can be located at relatively long distances from a target ambiguous word and ranked them according to the similarity measure to the target sense. The selected candidates are then used to extract training samples from the news corpus. We evaluated word sense disambiguation models based on a nearest neighbor classification on BERT and ELMo embeddings. Our work relies on the Russian wordnet RuWordNet.

Keywords: Word sense disambiguation, Russian dataset, Monosemous relatives, Automatic Dataset Collection, ELMo, BERT

DOI: 10.28995/2075-7182-2020-19-119-132

АВТОМАТИЧЕСКИЙ СБОР И РАЗМЕТКА ОБУЧАЮЩЕЙ КОЛЛЕКЦИИ ДЛЯ ЗАДАЧИ РАЗРЕШЕНИЯ ЛЕКСИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ НА РУССКОМ ЯЗЫКЕ

Большина А. С. (angelina_ku@mail.ru)

МГУ им. М. В. Ломоносова, Москва, Россия

Лукашевич Н. В. (louk_nat@mail.ru)

Казанский федеральный университет, Казань, Россия

Ключевые слова: разрешение лексической неоднозначности, обучающая коллекция текстов на русском языке, однозначные родственники, автоматический сбор текстовой коллекции, ELMo, BERT

1. Introduction

Word Sense Disambiguation (WSD) is a key task in semantic-oriented applications such as semantic text analysis, semantic information retrieval and knowledge graph construction. To achieve high performance, supervised WSD algorithms require large sense-annotated datasets. The annotation of such corpora demands considerable time and human resources, that is why supervised machine learning approaches suffer from a knowledge acquisition bottleneck.

There exist several hand-crafted sense-annotated datasets for English ([Miller et al. 1993], [Taghipour & Ng 2015]). However, not for all languages such corpora are available, and that certainly hinders the development of WSD systems for those languages. This also holds true for the Russian language.

Automatic acquisition of training samples can help to tackle this problem. Our research is focused on the monosemous relatives approach, which exploits a set of unambiguous words (or phrases) related to particular senses of polysemous word. However, as it was noted in [Martinez et al. 2006], some senses of target words do not have monosemous relatives, and the noise can be introduced by some distant relatives. In our research we tried to address these issues.

The main contribution of this study is that we have expanded a set of monosemous relatives under consideration and used word embeddings to estimate the similarity between a monosemous relative and a particular sense of a target word, which is further used in the development of the training collection. According to our knowledge, this is the first work that uses embedding similarity for selection of training contexts for the word sense disambiguation task. In order to evaluate the created training collections, we utilized contextualized word representations—ELMo [Peters et al. 2018] and BERT [Devlin et al. 2019]. We also conducted an experiment to compare the performance of the models trained on the collections with close monosemous relatives (synonyms, hypernyms and hyponyms) and more distant ones¹.

The paper is organized as follows. In section two we review the related work. Section three describes the data utilized in the research. The fourth section describes the method applied to automatically generate and annotate training collections. The procedure of creating a collection is explained in the fifth section. In the sixth section we describe a supervised word sense disambiguation algorithm trained on our collected material and also present the results obtained by four different models. Concluding remarks are provided in the seventh section.

2. Related Work

To overcome the limitations, that are caused by the lack of annotated data, several methods of generating and harvesting large train sets have been developed. There exist many techniques based on different kinds of replacements, which do not require human resources for tagging. The most popular method is that of monosemous

¹ The source code of our algorithm and experiments is publicly available at: https://github.com/loenmac/russian_wsd_data.

relatives [Leacock et al. 1998]. Usually WordNet [Miller 1995] is used as a source for such relatives. WordNet is a lexical-semantic resource for the English language that contains description of nouns, verbs, adjectives, and adverbs in form of semantic graphs. All words in those networks are grouped into sets of synonyms that are called synsets.

Monosemous relatives are those words or collocations that are related to the target ambiguous word through some connection in WordNet, but they have only one sense, i.e. belong only to one synset. Usually, synonyms are selected as relatives but in some works hypernyms and hyponyms are chosen [Przybyła 2017]. In the work [Martinez et al. 2006] distant relatives (including distant hypernyms and hyponyms) were used; the procedure of training contexts selection was based on the distance to a target word and the type of the relation connecting the target sense and a monosemous relative.

Some researchers replace the target word with named entities [Mihalcea & Moldovan 2000], some researchers substitute it with meronyms and holonyms [Seo et al. 2004]. In the article [Yuret 2007] a special algorithm was created in order to select the best replacement out of all words contained within synsets of the target word and neighboring synsets. The algorithm described in [Mihalcea 2002], that is used to construct annotated training set, is a combination of different approaches: monosemous relatives, glosses and bootstrapping. Monosemous relatives can be also used in other tasks, for example, for finding the most frequent word senses in Russian [Loukachevitch & Chetviorkin 2015].

Other methods of automatic generation of training collections for WSD exploit parallel corpora [Taghipour & Ng 2015], Wikipedia and Wiktionary [Henrich et al. 2012], topic signatures [Agirre & De Lacalle 2004]. [Pasini & Navigli 2017] created large training corpora exploiting a graph-based method that took an unannotated corpus and a semantic network as an input.

Various supervised methods including kNN, Naive Bayes, SVM, neural networks are applied to word sense disambiguation [Navigli 2009]. Recent studies have shown the effectiveness of contextualized word representations for the WSD task ([Wiedemann et al. 2019], [Kutuzov & Kuzmenko 2019]). The most widely used deep contextualized embeddings are ELMo [Peters et al. 2018] and BERT [Devlin et al. 2019]. In ELMo (Embeddings from language models) [Peters et al. 2018] context vectors are computed in an unsupervised way by two layers of bidirectional LSTM, that take character embeddings from convolutional layer as an input. Character-based token representations help to tackle the problems with out-of-vocabulary words and rich morphology. BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al. 2019] has a different type of architecture, namely multi-layer bidirectional Transformer encoder. During pre-training procedure, the model is “jointly conditioning on both left and right context in all layers” [Devlin et al. 2019]. Moreover, BERT uses WordPiece tokens, that is subword units of words, which also helps to avoid the problem of out-of-vocabulary words. Since these contextualized word embeddings imply capturing polysemy better than any other representations and, thus, fit well into the task of WSD, we employ them in our investigation.

3. Data

In our research as an underlying semantic network we exploit Russian wordnet RuWordNet [Loukachevitch et al. 2016]. It is a semantic network for Russian that has a WordNet-like structure. It is composed of 111.5 thousand of words and word combinations for the Russian language. RuWordNet has been published on the Linguistic Linked Open Data cloud [Cimiano et al. 2020] and interlinked [Loukachevitch & Gerasimova 2019] with the Collaborative Interlingual Index (CILI) [Bond et al. 2016].

RuWordNet contains 29,297 synsets for nouns. There are 63,014 monosemous and 5,892 polysemous nouns in RuWordNet. Total number of polysemous nouns' senses equals to 14,357. This resource was used to extract semantic relations (e.g. synonymy, hyponymy etc.) between a target sense of a polysemous word and all the words (phrases) connected to it, including those linked via distant paths. The sense inventory was also taken from this resource.

As a reference corpus we utilized a news corpus, that consists of one million news articles harvested from various news sources. All the texts have already been cleaned from html-elements or any other markup. The corpus consists of 24.2 million sentences, 288,1 million lemmas and 1,4 million of unique lemmas.

For evaluation of our algorithm of training data generation, we used three distinct RUSSE'18 datasets for Russian [Panchenko et al. 2018]. These datasets were created for the shared task on word sense induction for the Russian language. The first dataset is compiled from the contexts of the Russian National Corpus. The second dataset consists of the contexts from Wikipedia articles. And the last dataset is based on the Active Dictionary of the Russian Language [Apresyan et al. 2017] and contains contexts taken from the examples and illustration sections from this dictionary. All the polysemous words are nouns.

From the RUSSE dataset we excluded words whose senses are absent in RuWordNet. For example, the word *гипербола* 'hyperbole' from RUSSE'18 dataset is missing in the RuWordNet vocabulary. The word *мандарин* has two senses described in RUSSE'18: its sense 'tangerine' is included in the thesaurus, whereas its sense 'mandarin, bureaucrat' is absent, that is why we did not put it in the final test set. Some of the words like *карьер* 'quarry/a very fast gallop' and *шах* 'shah/check' do not have enough examples for one of their senses in the news corpus.

The final list of the target ambiguous words contains 30 words in total, each having two different senses. The **Appendix 1** contains the list of the target ambiguous words selected from RUSSE'18 dataset. For convenience we will call the resulting test dataset RUSSE-RuWordNet because it is a projection of RUSSE'18 sense inventory on the RuWordNet data. The entire dataset consists of 2,103 sentences, 39,311 lemmas and 12,110 unique lemmas.

We also created a small training dataset, that consists of the word sense definitions and examples of uses from Ozhegov dictionary [Ozhegov 2014] for every target polysemous word. Each sense of an ambiguous word has one definition and between 1 and 3 usage examples. This training data is utilized as a baseline for the WSD task.

4. Candidate Selection and Ranking Algorithm

The central idea of our method is based on the assumption that a training collection can be built not only with the direct relations like synonymy, hyperonymy and hyponymy but also with far more distant words, such as co-hyponyms. For example, most contexts for the word *крона* in the sense ‘krona, currency’ match the contexts of the other words denoting currency like *английский фунт* ‘pound sterling’ as they have common hypernym *валюта* ‘currency’.

The principal features of our approach are as follows:

1. We take into consideration not only the closest relatives to a target word sense, as it was done in previous works, but also more distant relatives.
2. We utilize similarity scores between a candidate monosemous relative and synsets close to a sense of a target polysemous word in order to evaluate how well this candidate can represent the sense of an ambiguous word.
3. We introduce the notion of a ***synset nest*** which is used to assess the potential of candidate’s usage contexts for displaying target sense of a polysemous word. To measure the relevance and suitability of a monosemous candidate, we exploit a thesaurus set of words similar to a target sense. The group of synonyms to a target sense and all the words from directly related synsets within 2 steps from a target word comprise ***the synset nest*** for a target sense.
4. We check similarity scores to the synset nest for both closest and further located monosemous relatives because a word described as monosemous in the thesaurus can actually have polysemous usage in a corpus. For example, Russian word *ириска* (‘toffee’) can also denote a nickname of Everton Football Club (The Toffees) [Loukachevitch 2019]. Thus, all candidate monosemous relatives should be further checked on the source corpus.
5. We propose two distinct methods of compiling a training collection based on the monosemous relatives rating.

A target word sense is a sense of a polysemous word that we want to disambiguate. Candidate monosemous relatives are unambiguous words and phrases, that can be located in up to four-step relation paths to a polysemous word. Candidate monosemous relatives are unambiguous words and phrases which can be located in up to four-step relation paths to a polysemous word. We consider only those words or word combinations, that have more than 50 occurrences in the news corpus.

A fragment of the synset nest for the word *такса* ‘dachshund’ is given below:

- (1) “*охотничий пёс, охотничья собака, пёсик, четвероногий друг, псина, собака, терьер, собачонка, борзая собака...*” / ‘*hunting dog, hunting dog, doggie, four-legged friend, dog, dog, terrier, dog, greyhound dog...*’

Our method of extracting monosemous relatives is based on comparison of distributional and thesaurus similarities. The word embedding model is utilized to select the most appropriate monosemous relatives whose context serve as a good representation of a target word sense. We used the word2vec model to extract 100 most similar words to each monosemous word from the candidates list. Thus, we collected the words that

represent a distributional set of close words with the respective cosine similarities measures. Our selection and ranking method, thus, consists of the following steps:

1. We extract all the candidate monosemous relatives within 4 steps from a target polysemous word sense s_j .
2. We compile the synset nest ns_j which consists of all closely related words to the target sense s_j , that is, for example, synonyms, hyponyms, hypernyms and cohyponyms. The synset nest ns_j consists of N_k synsets.
3. For each candidate monosemous relative r_j , we find 100 most similar words according to the word2vec model trained on a reference corpus.
4. We intersect these top-100 words with the words included in the synset nest ns_j of the target sense s_j .
5. For each word in the intersection, we take its cosine similarity weight calculated with the word2vec model and assign it to the synset it belongs to. The final weight of the synset in the synset nest s_j is determined by the maximum weight among the words $w_{k_1}^j, \dots, w_{k_i}^j$ representing this synset in the intersection.
6. The total score of the monosemous candidate r_j is the sum of the weights of all synsets from the synset nest ns_j . In such a way more scores are assigned to those candidates, that resemble a greater number of synsets from the synset nest close the target sense of the ambiguous target word. Thus, the final weight of the candidate can be defined as follows:

$$Weight_{r_j} = \sum_{k=1}^{N_k} \max [\cos(r_j, w_{k_1}^j), \dots, \cos(r_j, w_{k_i}^j)]$$

The following fragment of list of monosemous relatives with similarity scores (given in brackets) was obtained for the noun *звездика* ‘clove’:

- (2) *чёрный перец* ‘black pepper’ (7.5), *кардамон* ‘cardamom’ (6.8), *корица* ‘cinnamon’ (6.5), *имбирь* ‘ginger’ (6.4), *мускатный орех* ‘nutmeg’ (6) ...

We have also found some examples where a monosemous word is connected to a sense of a target word but received zero similarity weight. For example, the word *марля* ‘gauze’ is a cohyponym to the word *байка* in the sense ‘thick flannelette’, but it was not included in the monosemous relatives list because its distributional set of close words did not have any intersection with the synset nest.

As a result of the described procedure, all monosemous relatives are sorted by the weight they obtained. The higher rated monosemous relatives are supposed to be better candidates to represent the sense of the target word and, consequently, their contexts of use are best suited as the training examples in the WSD task. The candidate ranking algorithm identifies which monosemous relatives are most similar to the target ambiguous word’s sense. Once we have selected the monosemous candidates, we can extract from the corpus the contexts in which they occur. Then we substitute the monosemous relatives with the target ambiguous word in these texts and add them to a training collection.

5. Generating Training Data using Monosemous Relatives

The news corpus was used to extract the contexts with monosemous relatives found by the proposed algorithm. For comparison, we decided to create training collections in two ways. We compiled the first collection only with a monosemous relative from the top of the candidate rating. We wanted to obtain 1000 examples for each of the target words, but sometimes it was not possible to extract so many contexts with one particular candidate. That is why in some cases, we also took examples with words next on the candidates' list. For simplicity we call this collection Corpus-1000 because we obtained exactly 1000 examples for each sense.

As for the second collection, the training examples for the target ambiguous words were collected with the help of all respective unambiguous relatives with non-zero weight. The number of extracted contexts per a monosemous candidate is in direct proportion to its weight. Accordingly, we name this collection a balanced one because the selection of training examples was not restricted to the contexts which have only one particular monosemous relative.

The quantitative characteristics of the relations connecting the target senses and their monosemous relatives, distances between them and a proportion of monosemous relatives expressed as a phrase are given in the **Table 1**.

Table 1. Quantitative characteristics of monosemous relatives

Feature	Proportion of occurrences
Distance to a target sense	
0 (synset)	2%
1	13%
2	38%
3	31%
4	16%
Relation between a target sense and a monosemous relative	
Synonyms	2%
Hyponyms	13%
Hypernyms	11%
Cohyponyms	28%
Cohyponyms situated at three-step path	24%
Cohyponyms situated at four-step path	19%
Other	3%
Word combinations	48%

The word2vec embedding model that we used in our experiments was trained on the news corpus with the window size of 3. As a preprocessing step, we split the corpus into separate sentences, tokenized them, removed all the stop words and lemmatized the words with pymorphy2 tool [Korobov 2015]. We decided to lemmatize the train and test data, because it was shown in [Kutuzov & Kuzmenko 2019, 2], that “feeding ELMo with lemmas instead of raw tokens can improve WSD performance”. The words obtained from the word2vec model were filtered out—we removed the ones not included in the thesaurus vocabulary.

In the **Appendix 1** we present the characteristics of the two training collections: the list of the target ambiguous words selected from RUSSE'18 dataset, their senses and the number of examples per each sense respective to the collection type.

6. Experiments

We conducted several experiments to determine whether our text collection can be used as a training dataset for a WSD model. Following [Wiedemann et al. 2019], in our research we used an easily interpretable classification algorithm—non-parametric nearest neighbor classification (kNN) based on the contextualized word embeddings ELMo and BERT.

In our experiments we exploited two distinct ELMo models—the one trained by DeepPavlov on Russian WMT News and the other is RusVectōrēs [Kutuzov & Kuzmenko 2017] lemmatized ELMo model trained on Taiga Corpus [Shavrina & Shapovalova 2017]. The difference between these two models is that from the first model we extracted a vector for a whole sentence with a target word, whereas from the second model we extracted a single vector for a target ambiguous word. As for BERT, we used two models: BERT-base-multilingual-cased released by Google Research and RuBERT, which was trained on the Russian part of Wikipedia and news data by DeepPavlov [Kuratov & Arkhipov 2019]. To extract BERT contextual representations, we followed the method described by [Devlin et al. 2019] and [Wiedemann et al. 2019] and concatenated “the token representations from the top four hidden layers of the pre-trained Transformer” [Devlin et al. 2019].

The **Table 2** demonstrates the results obtained by different types of contextualized word embeddings, the training collections and model parameters. As it can clearly be seen, all the systems surpassed the quality level of the baseline solution trained on the dataset of the dictionary definitions and usage examples. So, this means that we have managed not only to collect training data sufficient to train the WSD model but also to show a good performance on the RUSSE-RuWordNet dataset.

Table 2. F1 scores for ELMo- and BERT-based WSD models (best results are marked bold): (k)—number of nearest neighbors, (1)—Corpus-1000, (2)—Balanced collection

Model	ELMo RusVectōrēs (target word)		ELMo DeepPavlov (whole sentence)		RuBERT DeepPavlov		Multilingual BERT	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1	0.794	0.797	0.752	0.758	0.735	0.75	0.67	0.662
3	0.811	0.81	0.749	0.753	0.756	0.755	0.673	0.681
5	0.819	0.81	0.748	0.756	0.771	0.769	0.667	0.682
7	0.819	0.815	0.746	0.759	0.774	0.768	0.673	0.683
9	0.816	0.821	0.747	0.753	0.769	0.774	0.677	0.688
Baseline	0.772		0.716		0.667		0.672	

The qualitative analysis of the classification errors showed that the main cause of mistakes were lexical and structural differences between the training and test sets. The examples from the test dataset were taken from the Russian National Corpus and Wikipedia, whereas the training collections were composed of news articles. Adding more data of various genres will help to diversify the training collections, thus, the training samples will have more similar representations to the test ones.

The algorithm based on ELMO pre-trained embeddings by RusVectōrēs outperformed all other models achieving 0.821 F1 score. The second-best model in the WSD task is RuBERT by DeepPavlov, followed by ELMO model by DeepPavlov. The lowest F1 score belongs to Multilingual BERT.

As for the difference in F1 scores between the Corpus-1000 and the balanced collection, we can observe the minor performance drop for the Corpus-1000 for all the models except for the RuBERT model. Corpus-1000 does not include all possible monosemous relatives, so dataset lacks contextual diversity, the balanced collection, on the contrary, is more representative with regard to the variety of contexts.

It is worth noting, that only 20 words of our dataset have close-related monosemous relatives for all their senses connected to a target word with the direct relations (synonymy, hyponymy, hypernymy), which means that the proposed expansion of paths is very useful. We evaluated the results of word sense disambiguation for these 20 words using only direct monosemous relatives and all the proposed relatives and found that the best results achieved by the RusVectōrēs ELMO model are quite similar: 0.841 (direct relatives) vs. 0.835 (all relatives).

7. Conclusion

The issue that we addressed in this article is the lack of sense-annotated training data for supervised WSD systems in Russian. In this paper we have described our algorithm of automatic collection and annotation of training data for the Russian language. Our training collections consist of the texts obtained from the news corpus and can be further replenished. The main contribution of the paper is that we have considered in the selection algorithm a wide range of monosemous relatives' types and utilized the metric based on a cosine similarity to determine the most appropriate monosemous relatives to be added to the training collection.

In order to evaluate the training collections, we applied kNN classifier to the contextualized word embeddings extracted for the target polysemous words and measured its performance on the RUSSE-RuWordNet test dataset. We have investigated the capability of different deep contextualized word representations to model polysemy. The best result was obtained with RusVectōrēs ELMO model and amounted to 0.821 F1 score.

As future work we plan to add more texts of different genres to the training collection.

8. Acknowledgements

The work of Loukachevitch N. in the current study concerns formulation of the disambiguation approach for RuWordNet data, calculation of paths between synsets, criteria for selecting contexts; this work is supported by the Russian Science Foundation grant no. 19-71-10056 financed through Kazan Federal University.

References

1. *Agirre E., De Lacalle O. L.* (2004), Publicly Available Topic Signatures for all WordNet Nominal Senses. In LREC.
2. *Bond F., Vossen P., McCrae J.P., Fellbaum C.* (2016), CILI: the Collaborative Interlingual Index. In Proceedings of the 8th Global WordNet Conference (GWC 2016), pp. 50–57.
3. *Cimiano P., Chiarcos C., McCrae J.P., Gracia J.* (2020), Linguistic Linked Open Data Cloud. In Cimiano P., et al. Linguistic Linked Data: Representation, Generation and Applications. Springer, pp. 29–41.
4. *Devlin J., Chang M.-W., Lee K., Toutanova K.* (2019), 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, pp. 4171–4186.
5. *Henrich V., Hinrichs E., Vodolazova T.* (2012), WebCAGe: A Web-Harvested Corpus Annotated with GermaNet Senses. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 387–396. Association for Computational Linguistics.
6. *Korobov M.* (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Analysis of Images, Social Networks and Texts, pp. 320–332.
7. *Kuratov Y., Arkhipov M.* (2019), Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213.
8. *Kutuzov A., Kuzmenko E.* (2017), WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham.
9. *Kutuzov A., Kuzmenko E.* (2019), To Lemmatize or not to Lemmatize: How Word Normalisation Affects ELMo Performance in Word Sense Disambiguation. In Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing, pp. 22–28.
10. *Leacock C., Miller G. A., Chodorow M.* (1998), Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics, vol. 24(1), pp. 147–165.
11. *Lopukhina A. A., Apresyan V. Yu., Iomdin B. L., Apresyan Yu. D., Boguslavsaya O. Yu., Krylova T. V., Levontina I. B., Sannikov A. V., Uryson E. V., Babaeva E. E., Glovinskaya M. Ya., Ptentsova A. V.* (2017), Active Dictionary of the Russian Language [Aktivnyj slovar' russkogo yazyka]. Publishing House Nestor-Istoria, Moscow, Vol. 3.

12. *Loukachevitch N. V.* (2019), Corpus-based Check-up for Thesaurus. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5773–5779.
13. *Loukachevitch N. V., Lashevich G., Gerasimova A. A., Ivanov V. V., Dobrov B. V.* (2016), Creating Russian WordNet by Conversion. In Proceedings of Conference on Computational linguistics and Intellectual technologies Dialog-2016, pp. 405–415.
14. *Loukachevitch N. V., Chetviorkin I.* (2015), Determining the Most Frequent Senses Using Russian Linguistic ontology RuThes. In Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015, pp. 21–27.
15. *Loukachevitch N., Gerasimova A.* (2019), Linking Russian Wordnet RuWordNet to WordNet. In Proceedings of the 10th Global Wordnet Conference (GWC 2019), pp. 64–71.
16. *Martinez D., Agirre E., Wang X.* (2006), Word Relatives in Context for Word Sense Disambiguation. In Proceedings of the Australasian Language Technology Workshop 2006, pp. 42–50.
17. *Mihalcea R.* (2002), Bootstrapping Large Sense Tagged Corpora. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), vol. 1999. Las Palmas, Canary Islands, Spain.
18. *Mihalcea R., Moldovan D. I.* (2000), An Iterative Approach to Word Sense Disambiguation. In FLAIRS Conference, pp. 219–223.
19. *Miller G.* (1995), WordNet: A Lexical Database for English. In Communications of the ACM, vol. 38(11), pp. 39–41.
20. *Miller G. A., Leacock C., Teng R., Bunker R. T.* (1993), A Semantic Concordance. In Proceedings of the workshop on Human Language Technology, pp. 303–308. Association for Computational Linguistics.
21. *Navigli R.* (2009), Word Sense Disambiguation: A survey. ACM computing surveys (CSUR), vol. 41(2), 10.
22. *Ozhegov S. I.* (2014), Explanatory Dictionary of the Russian Language [Tolkovy Slovar' Russkogo Yazyka]. Edited by Skvortsova S. I., 8, pp. 1376.
23. *Panchenko A., Lopukhina A., Ustalov D., Lopukhin K., Arefyev N., Leontyev A., Loukachevitch N.* (2018), RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Moscow, Russia. RSUH, pp. 547–564.
24. *Pasini T., Navigli R.* (2017), Train-o-matic: Large-scale Supervised Word Sense Disambiguation in Multiple Languages Without Manual Training Data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 78–88.
25. *Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L.* (2018), Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2227–2237.

26. *Przybyła P.* (2017), How Big is Big Enough? Unsupervised word sense disambiguation using a very large corpus. arXiv preprint arXiv:1710.07960.
27. *Seo H. C., Chung H., Rim H. C., Myaeng S. H., Kim S. H.* (2004), Unsupervised Word sense Disambiguation Using WordNet Relatives. Computer Speech & Language SPEC. ISS., vol. 18, no. 3, pp. 253–273.
28. *Shavrina T., Shapovalova O.* (2017), To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser. In Proceedings of “CORPORA2017”, international conference, Saint-Petersburg.
29. *Taghipour K., Ng H. T.* (2015), One Million Sense-tagged Instances for Word Sense Disambiguation and Induction. In Proceedings of the nineteenth conference on computational natural language learning, pp. 338–344.
30. *Wiedemann G., Remus, S., Chawla A., Biemann C.* (2019), Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. arXiv preprint arXiv:1909.10430.
31. *Yuret D.* (2007), KU: Word Sense Disambiguation by Substitution. In Proceedings of the 4th International Workshop on Semantic Evaluations, pp. 207–213. Association for Computational Linguistics.

Приложение 1. Appendix

Appendix 1. Target polysemous words from RUSSE-RuWordNet dataset

Polysemous word	Sense	Corpus-1000	Balanced collection
акция ₁	Share/stock	1000	1239
акция ₂	Action	1000	1314
байка ₁	Tale/story	1000	1227
байка ₂	Thick flannelette	245	245
гвоздика ₁	Carnation	1000	1314
гвоздика ₂	Cloves	1000	1154
гусеница ₁	Caterpillar	1000	1295
гусеница ₂	Track	1000	1153
капот ₁	Bonnet/hood of a car	1000	918
капот ₂	Housecoat	1000	1084
крона ₁	Top of a tree	1000	1131
крона ₂	Krona (currency)	1000	1314
рок ₁	Rock music	1000	1016
рок ₂	Destiny	1000	938
слог ₁	Syllable	1000	1047
слог ₂	Style	1000	1137
стопка ₁	Pile	1000	1258
стопка ₂	Small drinking glass	1000	1005

Polysemous word	Sense	Corpus-1000	Balanced collection
таз ₁	Pelvis	1000	1124
таз ₂	Basin	1000	1314
такса ₁	Price/charge	1000	1300
такса ₂	Dachshund	1000	1069
замок ₁	Castle	1000	1078
замок ₂	Lock	1000	947
лук ₁	Bow	1000	1286
лук ₂	Onion	1000	1267
бор ₁	Boron	1000	1292
бор ₂	Pine Forest	1000	675
дар ₁	Talent	1000	1117
дар ₂	Gift	1000	1169
двигатель ₁	Engine	1000	1310
двигатель ₂	Something that causes a process to happen	1000	1305
дедушка ₁	Old man	1000	1299
дедушка ₂	Grandfather	1000	1231
декрет ₁	Maternity leave	128	128
декрет ₂	Decree	1000	1300
дерево ₁	Tree	1000	1309
дерево ₂	Timber	1000	966
диалог ₁	Conversation	1000	1278
диалог ₂	Negotiations	1000	1300
диплом ₁	Certificate	1000	1253
диплом ₂	Diploma paper	1000	1246
доктор ₁	Doctor, physician	1000	1310
доктор ₂	Doctor, degree	1000	1300
доля ₁	Part	1000	1300
доля ₂	Destiny	1000	1300
достижение ₁	Achievement	1000	1300
достижение ₂	Reaching the level	1000	1309
жестокость ₁	Ruthlessness	1000	801
жестокость ₂	Cruelty	1000	1313
жребий ₁	Lot	1000	1280
жребий ₂	Destiny	1000	1300
затея ₁	Fun	1000	1308
затея ₂	Enterprise	1000	1309
застой ₁	Stasis	1000	758
застой ₂	Stagnation	1000	1235

Polysemous word	Sense	Corpus-1000	Balanced collection
затишье ₁	Decline in activity	1000	1206
затишье ₂	Calm	1000	1300
затмение ₁	Mental breakdown	1000	1300
затмение ₂	Eclipse	1000	1002

REFERENTIAL PHENOMENA IN SPEAKER'S KINETIC CHANNELS

Budennaya E. V. (jane.sdrv@gmail.com)

Higher School of Economics / Institute of Linguistics, RAS,
Moscow, Russia

Evdokimova A. A. (arochka@gmail.com)

Institute of Linguistics, RAS, Moscow, Russia

Nikolaeva Ju. V. (julianikk@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

Sukhova N. V. (sukhova.natalya@gmail.com)

National University of Science and Technology MISiS,
Moscow, Russia

The article addresses the relation of referential expressions and co-occurring kinetic phenomena (hand and head gestures) on the material of the RUPEX multimodal corpus. The results reflect significant differences in how individual movements and gestures are aligned with two major types of reference (full NPs vs. reduced expressions). It was initially assumed that full NPs are more often accompanied by a gesture. Our data support this hypothesis not only through the material of hand gestures, but also through head movements. Moreover, full NPs are more likely to be accompanied by downward movements in both manual and cephalic channels, as well as by metadiscourse gestures, in comparison to reduced referential units (personal and demonstrative pronouns). In addition, pronouns are more likely to be aligned with pointing hand gestures and zero reference is often accompanied by descriptive hand gestures. However, the kinetic behavior of the interlocutors is determined by a variety of factors, including the topic of the conversation, which predisposes to certain types of gestures and the relative position of the interlocutors.

Key words: reference, referential choice, hand gestures, head movements, multimodal communication, RUPEX

DOI: 10.28995/2075-7182-2020-19-133-146

1. Introduction. Referential choice and related phenomena

Referential choice, or the process of selecting a linguistic expression for a specific referent, is a challenging task in linguistics. Factors related to referential choice, such as discourse phenomena (linear and referential distance between the anaphor and its antecedent, the referent's status as protagonist), grammatical categories (i.e.

animacy, gender, grammatical role), as well as typological parameters of the language as a whole (i.e. pro-drop properties), have been actively explored since the last third of the 20th century [Givón 1983]; [Gundel et al. 1993]; [Kibrik et al. 2016]; [Hint et al. 2020]. However, the exact list of factors determining referential choice remains open.

In this regard, nonverbal communication holds promise of potential related phenomena. Compared with linguistic data, non-verbal factors of referential choice have been less studied and mostly in the context of hand gestures [Levy, Fowler 2000]; [Gullberg 2006]; [Debreslioska et al. 2013]. This article considers the issue in a wider framework of Russian multichannel communication, namely in the context of head movements (cephalic channel) and hand gestures (manual channel). To date, referential mechanisms in the Russian speaker's kinetic channels have been mainly studied in the limited context of manual gestures [Grishina 2017: 78–80]. Head movements were sometimes discussed as an additional phenomenon [ibid: 156–158], but no systematic studies of both channels within the referential context were done thus far.

The work is based on the RUPEX multichannel corpus (“Russian Pear Chats and Stories”, see the project website www.multidiscourse.ru and [Kibrik 2018] for detail). For the pilot study, session #22 was selected with a total duration of approximately 54 minutes. The session consists of three videos recording a group of interlocutors with a fixed role for each (the Narrator, the Commentator and the Reteller¹). It forms a part of a demo subcorpus annotated for four kinetic communication channels (vocal, oculomotor, manual and cephalic), as well as for referential units.

The article is structured as follows. **Section 2** discusses referential expressions in the RUPEX and what is already known about their interaction with nonverbal units of communication. **Section 3** describes principles for manual and cephalic annotation. **Sections 4** and **5** are devoted to the results of comparing head and hand movements with aligned referential expressions.

2. Referential units in the context of verbal and non-verbal factors

The RUPEX referential mark-up includes all language expressions with a specific reference. Among them, two main types of units are distinguished—anaphoric and deictic. This article is devoted to anaphoric reference. For RUPEX deictic expressions and their relation to the speaker's communicative role, see [Budennaya 2019: 132].

Anaphoric units are linguistic expressions whose interpretation depends on the previous context. Thus, the same referent can be marked by a full NP (*mal'čik uezžajet* ‘the boy leaves’), a personal pronoun (*on uezžajet* ‘he leaves’), a demonstrative pronoun (*tot uezžajet* ‘that [one] leaves’), a definite pronoun (*vse uezžajut* ‘everybody leaves’), a zero (\emptyset_{pro} *uezžajet* ‘[(s)he] leaves’), and so forth. Pronominal and zeroexpressions are combined into a more general type of reduced reference [Kibrik 2011]. It is opposed to full reference which is marked by full NPs acting as antecedents for subsequent anaphoric units: A [*tovarišč naverhu*]_i, [*on*]_i, *prodolzhaet sobirat' gruši i* [\emptyset_{pro}]_i *ničego ne vidit* ‘And [the guy]_i above, [he]_i keeps on picking pears and does not see anything’.

¹ Hereinafter N, C and R, respectively.

Both full and reduced referential expressions are marked in the RUPEX with ELAN software (<https://tla.mpi.nl/tools/tla-tools/elan/>), with zero units being attributed to the word explicitly expressed immediately thereafter (see [Budennaya 2019] for more details on the referential mark-up).

Traditionally reduced (or attenuated) anaphoric expressions code highly accessible and identifiable entities [Givón 1983]; [Ariel 1990]; [Gundel et al. 1993]. In contrast, full NPs are prototypically chosen for new and less accessible referents. Both full and reduced reference correlate with various grammatical, semantic and discourse factors, although their role varies from language to language [Kibrik et al. 2016]; [Hint et al. 2020]. From the non-verbal perspective, full NPs are more often accompanied by a gesture than reduced expressions [Levy, Fowler 2000], which holds true not only for gestures in general, but for their selected functional types as well (see, for example, [Azar et. Al. 2019] on pointing gestures). In this study, **Sections 3** and **4** will provide data on how the direction of a gesture, the referent's position in the speaker's gestural space, and the type of referential expression interact with each other in RUPEX cephalic and manual channels.

3. Methodology

3.1. Principles of cephalic and manual annotation

To compare verbal and non-verbal expressions, a unified method of annotating these phenomena in kinetic (cephalic and manual) channels was used. Within this method, all gestures were coded for directions and their phases.

Cephalic movements were tagged in accordance with the direction (down-up, right-left, forward-backward) and with the type of action (tilt, turn, slide, rotation, etc.). Since head movements are often superimposed, they were marked in three layers for accuracy: Movement A, Movement B and Displacement (the latter is not covered in the article). All communication movements performed by the neck muscles independently of other kinetic channels fell into Movement A. All movements produced under the influence of one of the other channels, as well as echo movements and adaptors, were tagged under Movement B [Kibrik, Fedorova 2020], see **Figure 1** for illustration.

When analyzing the relationship between the direction of movements and referential phenomena, the type of referential expression was marked along with the cases when the onset of the cephalic movement was ahead or lagged behind the concomitant referential expression, as shown in **Figure 2**.



Fig. 1. Sequence of Movements B

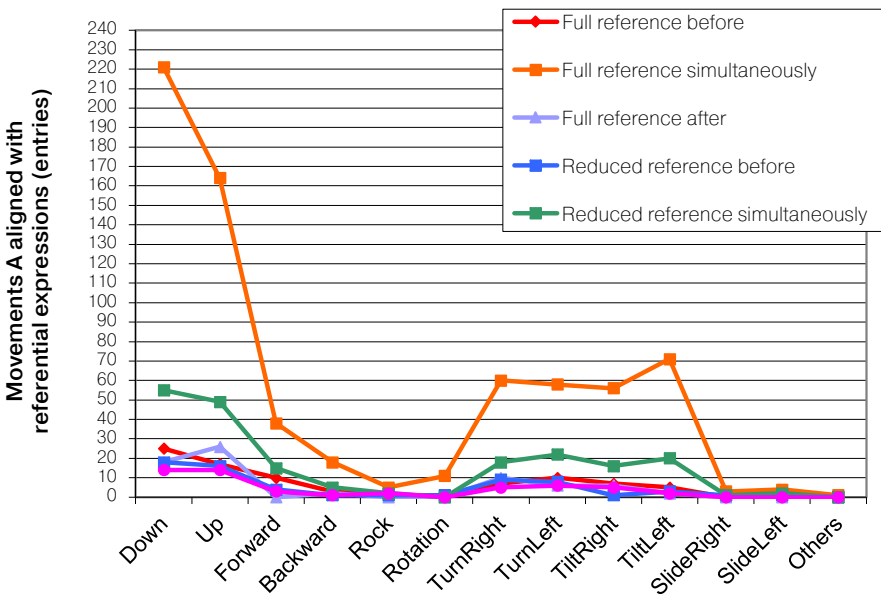


Fig. 2. Position of head movements (type A), according to concomitant referential units

The principles of the RUPEX’s manual annotation are detailed in [Litvinenko et al. 2018]. In this study, we examined only the hand movements that were marked as gestures. Each movement was coded for direction (up/down, right/left, to the center/from the center). For all movements aligned with referential expressions, both reference and functional type of gesture under discussion (see Section 3.2) were specified.

3.2. Types of cephalic and manual gestures

The functional types of hand gestures were approached as outlined in [Litvinenko et al. 2018], but pragmatic gestures were defined as a separate category, as were depictive, pointing and rhythmic gestures. This classification is based on functional grounds and can thus be applied to cephalic movements with some reservations. Since the main part of our material consists of monologue fragments (narration and retelling), only rare examples of regulatory hand movements marking the participants' interaction and conversational turn-taking were found. At the same time, head movements, if there are more than two interlocutors, often reflect how participants interact with each other and show to whom or to what the gesticulating person is paying attention. Such movements are marked with a higher amplitude and are clearly interpreted by the addressees. In this regard, the category of regulatory gestures was added for head movements [Ekman, Friesen 1969]. Some head gestures performed several functions, all such cases were considered separately.

4. Types of movement and reference

4.1. Direction of movement and type of reference in the cephalic channel

In total, 942 head movements of Movement A were aligned with referential expressions. The majority of them corresponded to downward and upward movements. Out of the three participants, only C had an equal distribution of the two types of movements, while R had a clear predominance of downward movements (30% and 17% respectively). The distribution is shown in **Figure 3**.

When compared with other parameters, a statistically significant relationship was found between the type of referential expression and the amplitude of the movement aligned. Thus, in about a half of full NPs, the speaker's head moved with a medium amplitude; in about a third, with a low amplitude, and only in about 17% of cases—with a high amplitude (χ -square, p -value < 0.01). On reduced expressions, the speaker produced movements of different amplitudes with an equal frequency (see **Figure 4**), but compared to full NPs, they moved their head significantly more often with a medium amplitude and significantly less often with a high one (χ -square, p -value < 0.01).

The above data indicate a significant correlation between downward movements and concomitant referential expressions (Pearson correlation coefficient = 0.979). That is, for all types of amplitude, downward movement A occurs more often on full NPs, rather than on reduced expressions (χ -square, p -value < 0.01; for illustration see **Figure 5**). Consequently, the movement A pattern, with the amplitude taken into account, allows one to predict the concomitant referential expression.

In Movement B type, 776 units were aligned with referential expressions. Their distribution is shown below (**Figure 6**).

Based on these data, we can assume that movements B are not affected by the type of concomitant referential expression. This provides additional evidence for movements B to be placed in a separate category.

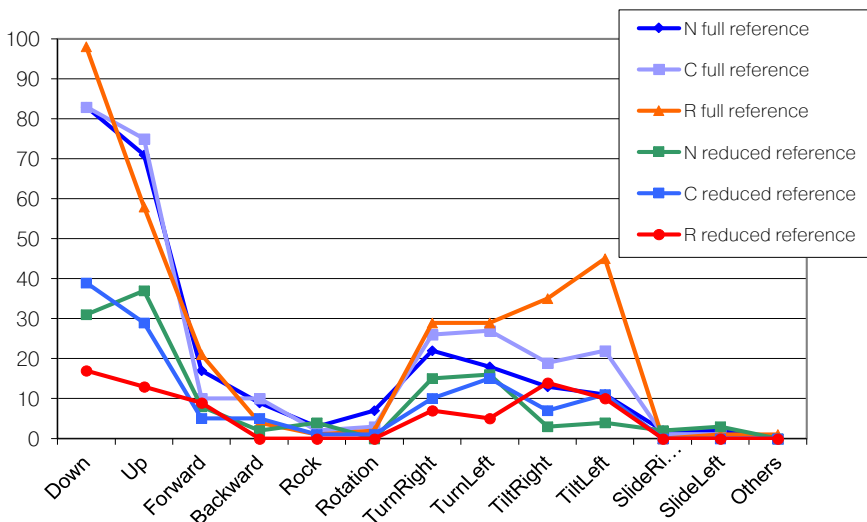


Fig. 3. Distribution of head movements (type A, %) among N/C/R referential expressions

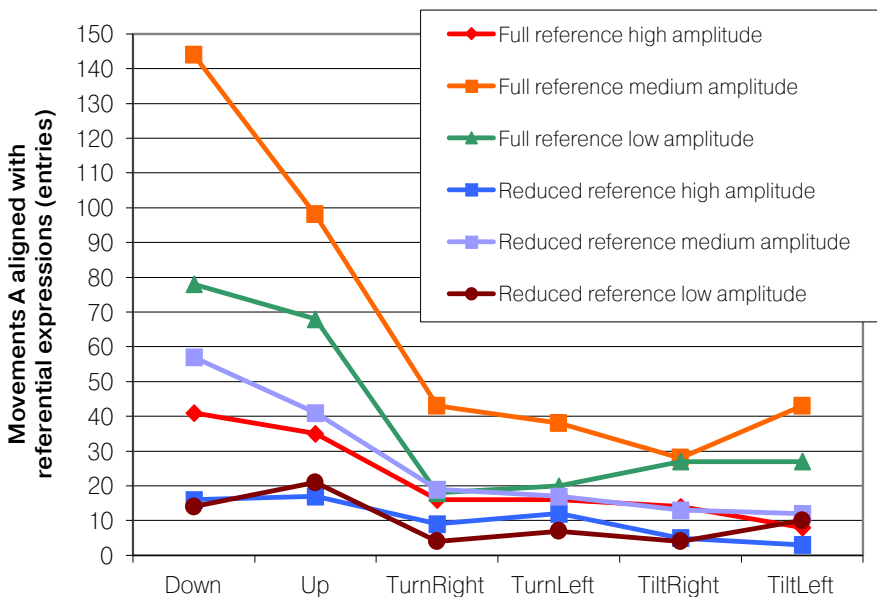


Fig. 4. Amplitude of head Movement A in relation to N/C/R referential expressions



Fig. 5 The realization of a depictive head gesture represented by a downward movement (Down) with a high amplitude on full NP *mnogo takih* 'there are many such...' The pictures show the head from the position forward moving down simultaneously with the hands moving up in a semicircle.

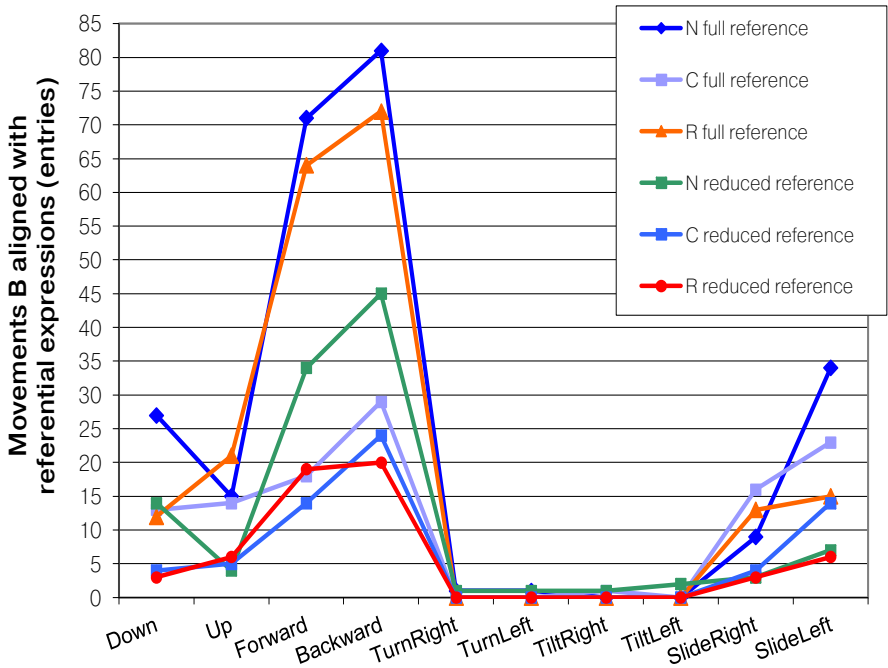


Fig. 6. Distribution of head movements (type B) in relation to referential expressions

4.2. Direction of movements and reference in the manual channel

A total of 894 hand movements was aligned with referential expressions. Their relation to the hand's direction is shown in **Figure 7**.

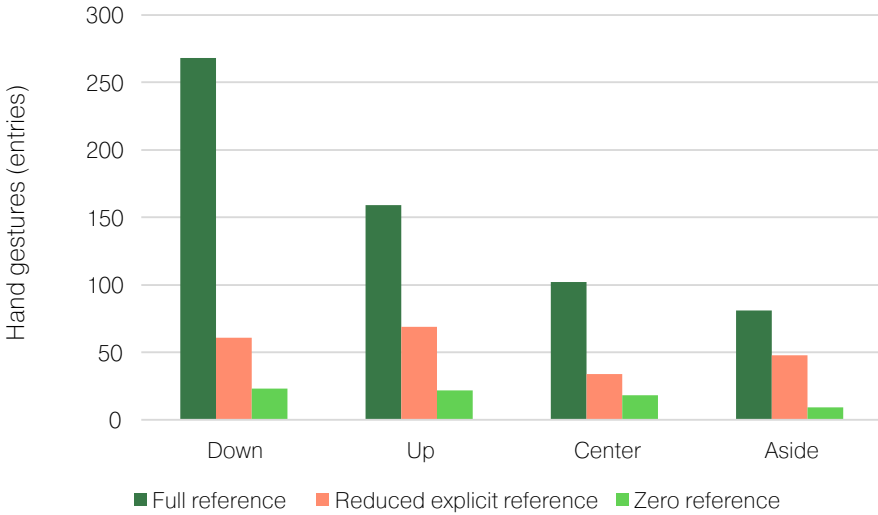


Fig. 7. Hand gesture direction within different types of reference (χ -square, $p < 0.01$)

According to our data, downward movements were produced more often than others but differed in terms of concomitant referential expressions. On full NPs, the hand was likely to go downward, as opposed to with reduced pronominal expressions, where the same direction was the least expected. In addition, zero forms were more often accompanied by movements to the center.

5. Functional types of gestures and reference

5.1. Cephalic gestures and reference

The functional types of head gestures differ from hand gestures in many ways. Although head gestures' phases can be identified similarly to their hand counterparts (i. e., preparation, stroke, retraction [Litvinenko et al. 2018]), they may fall under different functional types. For ease of comparison, only those types of head gestures which could also be distinguished in the manual channel were considered, cf. **Figure 8** and **9**.

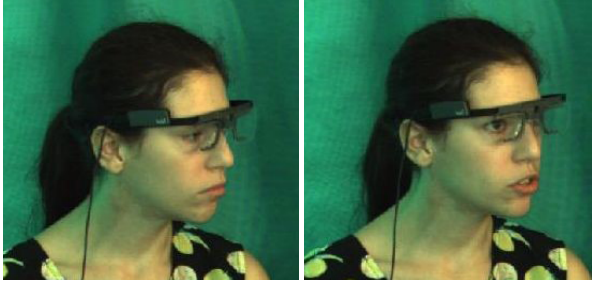


Fig. 8. The realization of a regulatory head gesture (TurnRight) aligned with the full NP *malčikam* 'to the boys'. The narrator keeps her head in the position from which it turns right as she utters the phrase.



Fig. 9. The realization of a depictive head gesture (Nod2) aligned with the a zero subject expression followed by the predicate *Ø vylezaet* '[he] crawls out'. The gesture path illustrates how the character crawls out from the fallen bike.

Given that, it was found that pragmatic and depictive head gestures were accompanied more frequently by full NPs (χ -square, p -value = 0.002).

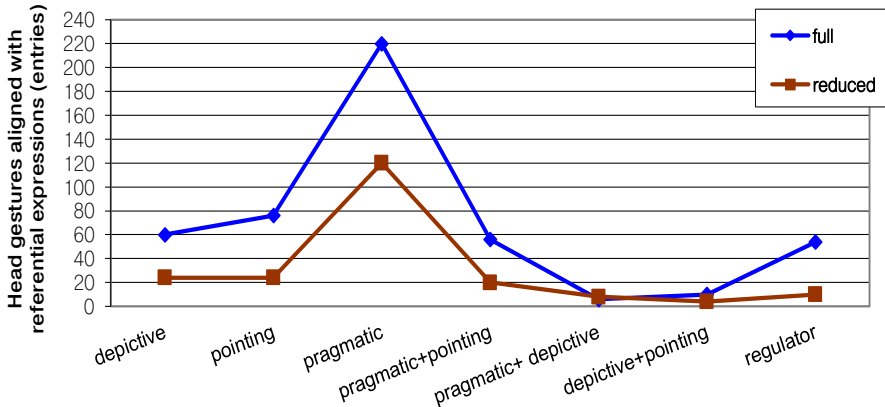


Fig. 10. Distribution of head gestures among referential expressions (22N)

Figure 10 supports the hypothesis that a person gesticulates more often on full NPs than on reduced units [Levy, Fowler 2000]; [Debreslioska et al. 2013] and confirms it for the cephalic channel, see **Figure 11** for illustration.



Figure 11. The realization of a pragmatic head gesture (Down) aligned with the full NP *pole* ‘field’.

5.2. Manual gestures and reference

In total, 425 referential expressions were accompanied hand gestures. There were no entries in this sample where zero reference coincided with a rhythmic gesture. Since the overall number of rhythmic gestures (beats) was small (38 cases), it was decided to combine them with pragmatic gestures, based on the fact that rhythmic gestures often perform metadiscourse functions [Krahmer, Swerts 2007]. The overall distribution of gestures, according to the functional type and to the type of concomitant referential expression, is presented in **Fig. 12** (a, b).

According to the above data, depictive gestures are more likely to occur on predicates with zero reference (χ -square, p -value < 0.005) and less likely to occur on explicit reduced referential expressions (personal, rare demonstrative and determinative pronouns). Pointing gestures, on the contrary, are less likely to be aligned with zero reference, but are often accompanied by explicit referential expressions, primarily by personal pronouns (cf. **Figure 13** and **14**).

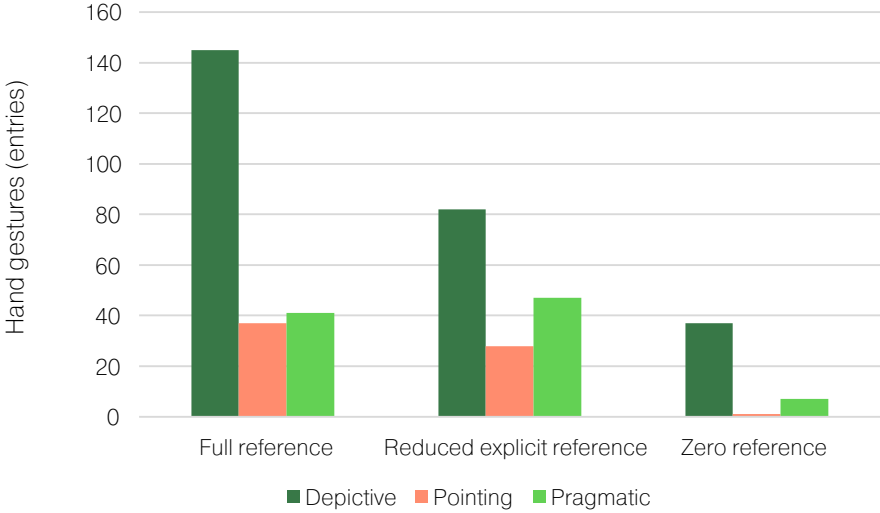


Fig. 12a. Functional types of hand gestures and types of reference (raw numbers)

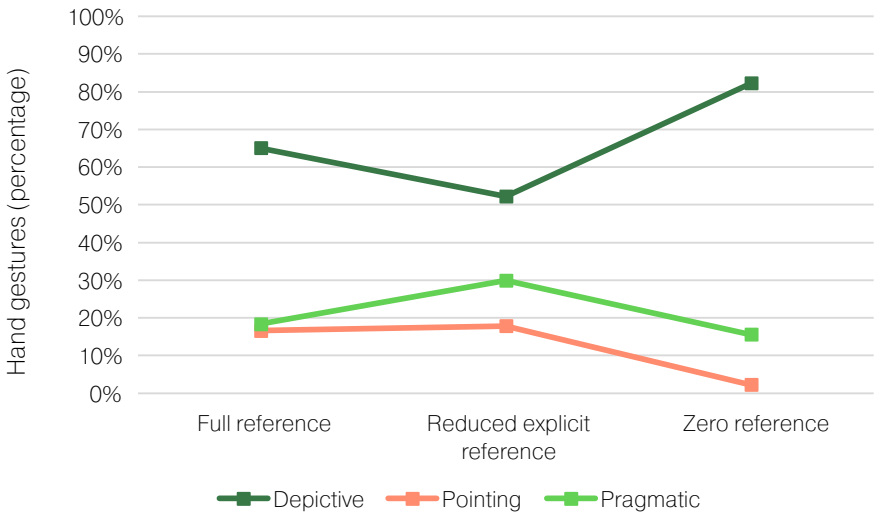


Fig. 12b. Functional types of hand gestures and types of reference (percentage)



Fig. 13. The realization of a depictive hand gesture with a downward direction which accompanies zero reference. The hand gesture occurs on the phrase *pomogajut emu podnyatsya* ‘they help him to stand up’. The narrator moves her right hand in a semicircle and then downward, even though the semantics of the word “stand up” suggests the movement “up”.

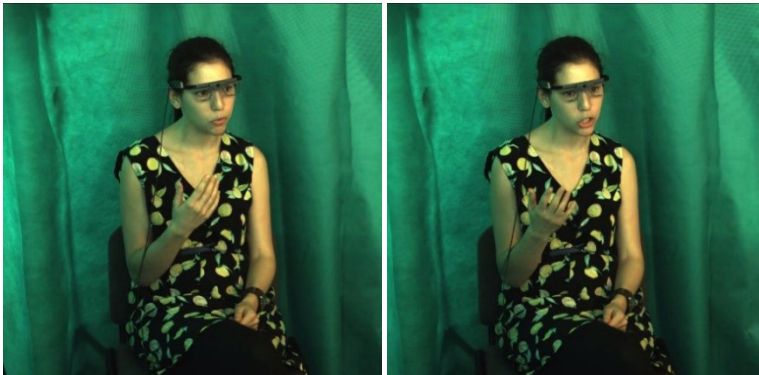


Fig. 14. The realization of a pointing hand gesture which accompanies the personal pronoun. On the phrase *čto on vot možet ne zametit’ čto-to* ‘that he can somehow miss something’ the narrator points to her face first, then she does several short movements with her hand away to the right.

6. Discussion

As the data in **Section 5** show, manual and cephalic gestures are combined in different ways with referential expressions.

There can be several explanations for the fact that zero expressions are more likely to be accompanied by depictive hand gestures. As shown in [Nikolaeva 2004], depictive gestures more often illustrate the foreground of the discourse and are often

aligned with the most dynamic events of the story's mainline. It is expected that such a description will relate to the actions of the protagonist, who is highly activated and therefore often not explicitly expressed. In addition, depictive gestures more often align with actions, and thus correlate with predicates to which zero expressions are attributed in the corpus. The data presented in **Figure 10** is consistent with the observation about the time coordination between the depictive gesture and the action performed by the protagonist [ibid]. We can assume that the detected correlation between depictive gestures and reduced referential expressions is a superficial phenomenon driven by internal factors, such as the referent's status as protagonist and topic continuity, which affect both the functional type of gesture and the type of concomitant referential expression.

Reduced reference is more likely to occur in situations of several competing antecedents. Their distinction is associated with the referent's position in the gestural space around the speaker, and thus the transition from one referent to another is accompanied by gestures in the horizontal plane. The least obvious is why zero referents are accompanied by movements of the hand or hands toward the center. Probably this tendency can be explained by specific stimulus material: the actions often mentioned are those that can be illustrated with upward gestures (for example, *sadovnik podnimaetsya po lestnice* 'the gardener goes up the stairs').

In addition, it was found that full NPs are more often accompanied not only by hand movements but also by cephalic gestures, compared to reduced expressions. Moreover, in the cephalic channel full NPs are more likely to be aligned with downward movements A and with pragmatic gestures. For Movement B, no correlation with referential expressions was detected but two directions were most often used: forward-backward and slide to the left. This indicates the adjustment of the gestural component of the interlocutors to each other and the independence of Movement B from reference. Other features can be explained by individual gestural portraits of the participants.

References

1. Ariel M. (1990), *Accessing noun-phrase antecedents*, Routledge, London.
2. Azar Z., Backus A., Özyürek A. (2019), General and language-specific factors influence reference tracking in speech and gesture in discourse, *Discourse Processes*, 56 (7), pp. 553–574.
3. Ekman P., Friesen W. V. (1969), The repertoire of nonverbal behavior: categories, origins, usage and coding, *Semiotica*, 1, pp. 49–98.
4. Budennaya E. V. (2019), Referential choice in multimodal communication, *Computational Linguistics and Intellectual Technologies [Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii "Dialog"]*, Moscow, 18 (25), pp. 125–136.
5. Debreslioska S., Özyürek A., Gullberg M., Perniss P. (2013), Gestural viewpoint signals referent accessibility, *Discourse Processes*, 50 (7), pp. 431–456.
6. Givón T. (1983), *Topic continuity in discourse: Quantified cross-language studies*, Benjamins, Amsterdam.

7. *Gullberg M.* (2006), Handling discourse: gestures, reference, tracking, and communication strategies in early L2, *Language Learning*, 56, pp. 155–196.
8. *Grishina E. A.* (2017), Russian gesticulation from a linguistic perspective (corpus studies) [Russkaja gestikuljatsija s lingvisticheskoj točki zrenija (korpusnyje issledovanija)], LRC Publishing House, Moscow.
9. *Gundel J. K., Hedberg N., Zacharski R.* (1993), Cognitive status and the form of referring expressions in discourse, *Language* 69 (2), pp. 274–307.
10. *Hint H., Nahkola T., Pajusalu R.* (2020), Pronouns as referential devices in Estonian, Finnish, and Russian, *Journal of Pragmatics* 155, pp. 43–63.
11. *Kibrik A. A.* (2011), *Reference in discourse*, Oxford University Press, Oxford.
12. *Kibrik A. A.* (2018), Russian multichannel discourse. Part II. Corpus development and avenues of research [Russkiy mul'tikanal'nyy diskurs. Chast' II. Razrabotka korpusa i napravleniya issledovaniy], *Psychological Journal [Psikhologicheskij Zhurnal]*, 39 (2), pp. 78–89.
13. *Kibrik A. A., Khudyakova M. V., Dobrov G. B., Linnik A., Zalmanov D. A.* (2016), Referential choice: predictability and its limits, *Frontiers in Psychology*, 7, p. 1429.
14. *Kibrik A. A., Fedorova O. V.* (eds.) (2020), *The MCD handbook: A practical guide to annotating multichannel discourse*, Moscow.
15. *Krahmer E., Swerts M.* (2007), The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception, *Journal of Memory and Language*, 57, pp. 396–414.
16. *Levy E., Fowler C.* (2000), Grounding references in perception, D. McNeill (ed.), *Language and gesture*, Cambridge University Press, New York, pp. 215–234.
17. *Litvinenko A. O., Kibrik A. A., Fedorova O. V., Nikolaeva J. V.* (2018), Annotating hand movements in multichannel discourse: Gestures, adaptors and manual postures, *Russian Journal of Cognitive Science*, 5 (2), pp. 4–17.
18. *Nikolaeva Y. V.* (2004), Functional and semantic aspects of illustrative gestures in Russian spoken discourse [Funksionalnye i semanticheskie osobennosti illyustrativnykh zhestov v ustnoy rechi (na materiale russkogo yazyka)], *Voprosy Jazykoznanija [Topics in the study of language]*, 4, pp. 48–64.

ЧЕМУ НАС УЧАТ ОШИБКИ: ТРУДНОСТИ ПРИ ОБРАБОТКЕ СЛОВ С ЧАСТОТНЫМИ ОРФОГРАФИЧЕСКИМИ ОШИБКАМИ¹

Чернова Д. А. (chernovadasha@yandex.ru)

СПбГУ, Санкт-Петербург, Россия

Алексеева С. В. (mail@s-alexeeva.ru)

СПбГУ, Санкт-Петербург, Россия

Слюсарь Н. А. (slioussar@gmail.com)

НИУ ВШЭ, Москва, и СПбГУ, Санкт-Петербург, Россия

DOI: 10.28995/2075-7182-2020-19-147-159

WHAT DO WE LEARN FROM MISTAKES: PROCESSING DIFFICULTIES WITH FREQUENTLY MISPELLED WORDS

Chernova D. A. (chernovadasha@yandex.ru)

St. Petersburg State University, St. Petersburg, Russia

Alexeeva S. V. (mail@s-alexeeva.ru)

St. Petersburg State University, St. Petersburg, Russia

Slioussar N. A. (slioussar@gmail.com)

HSE, Moscow, and St. Petersburg State University,
St. Petersburg, Russia

Even if we know how to spell, we often see words misspelled by other people — especially nowadays when we constantly read unedited texts on social media and in personal messages. In this paper, we present two experiments showing that the incidence of orthographic errors reduces the quality of lexical representations in the mental lexicon—even if one knows how to spell a word, repeated exposure to incorrect spellings blurs its orthographical representation and weakens the connection between form and meaning. As a result, it is more difficult to judge whether the word is spelled correctly,

¹ Исследование выполнено при поддержке РФФИ в рамках научного проекта № 18-00-0646 (18-00-00640) «Закономерности обработки лингвистической информации в ситуации многозначности: процессы активации и выбора вариантов значений».

and—more surprisingly—it takes more time to read the word even when there are no errors. We show that when all other factors are balanced the effect of misspellings is more pronounced for the words with lower frequency.

We compare our results with the only previous study addressing the problem of misspellings' influence on the processing of correctly spelled words—it was conducted on the English data. It may be interesting to explore this issue in a cross-linguistic perspective. In this study, we turn to Russian, which differs from English by a more transparent orthography. Much larger corpora of unedited texts are available for English than for Russian, but, using a different way to estimate the incidence of misspellings, we obtained similar results and could also make some novel generalizations.

In Experiment 1 we selected 44 words that are frequently misspelled and presented in two conditions (with or without spelling errors) and were distributed across two experimental lists. For every word, participants were asked to determine whether it is spelled correctly or not. The frequency of the word and the relative frequency of its misspelled occurrences significantly influenced the number of incorrect responses: not only it takes longer to read frequently misspelled words, it is also more difficult to decide whether they are spelled correctly.

In Experiment 2 we selected 30 words from the materials of Experiment 1 and for every selected word, we found a pair that is matched for length and frequency, but is rarely misspelled due to its orthographic transparency. We used a lexical decision task, presenting these 60 words in the correct spelling, as well as 60 nonwords. We used LMMs for statistics. Firstly, the word type factor was significant: it takes more time to recognize a frequently misspelled word, which replicates the results obtained for English. Secondly, the interaction between the word type factor and the frequency factor was significant: the effect of misspellings was more pronounced for the words of lower frequency. We can conclude that high frequency words have more robust representations that resist blurring more efficiently than low frequency ones.

Finally, we conducted a separate analysis showing that the number of incorrect responses in Experiment 1 correlates with RTs in Experiment 2. Thus, whether we consciously try to find an error or simply read words orthographic representations blurred due to exposure to frequent misspellings make the task more difficult.

Keywords: mental lexicon, orthographic errors, processing, Russian

1. Введение

Не все слова, которые мы знаем, мы знаем одинаково хорошо. В каких-то случаях мы не вполне уверены в значении слова, иногда у нас есть сомнения, как оно произносится или пишется. Согласно так называемой гипотезе лексического качества, Lexical Quality Hypothesis [Perfetti 1985], [2007]; а также [Nelson Taylor, Perfetti 2016]; [Perfetti, Hart, 2001]; [2002]), репрезентации таких слов в ментальном лексиконе оказываются «менее качественными»: некоторые формальные или семантические характеристики более размыты и связи между ними слабее, чем у других слов. Недостаточно четкая информация о буквенном составе слова ведет к орфографическим ошибкам.

Орфографические ошибки — страшная напасть, но мы привыкли думать, что терпение и труд помогут с ней справиться. Однако, даже если мы сами научились писать правильно то или иное слово, никто не может уберечь нас от ошибок, сделанных другими. Эта проблема стала особенно актуальной сейчас, с появлением многочисленных электронных СМИ, чатов, социальных сетей, где колоссальные массивы текстов публикуются без корректорской проверки или не предполагают ее в принципе. Если мы регулярно видим какое-то слово с ошибкой, можно предположить, что неправильное написание фиксируется в ментальном лексиконе и получает подкрепление каждый раз, когда мы сталкиваемся с ним снова. Тогда ментальные репрезентации слов, в которых часто допускают ошибки, окажутся менее качественными даже у тех, кто сам таких ошибок не делает.

Более того, согласно различным теориям, описывающим процессы обучения в целом и применительно к языку (например, [Baayen et al. 2011]; [Ramsar, Dye, McCauley 2013]; [Rescorla, Wagner 1972]), в ситуации, когда два разных набора формальных признаков связаны с одним и тем же значением, формируются более слабые связи. Это мешает запоминанию информации и затрудняет ее последующее извлечение из памяти. Предположим, что с такой точки зрения можно посмотреть и на слова, в которых часто допускают орфографические ошибки. Тогда можно ожидать, что не только написание, но и обработка таких слов при чтении окажется затруднена — даже тогда, когда эти слова написаны правильно.

Подобная гипотеза впервые была сформулирована в работе С. Рахманьян и В. Купермана [Rahmanian, Kuperman 2019], которые протестировали ее на материале английского языка. Они обратились к семимиллиардному корпусу, содержащему неотредактированные сообщения из новостных групп на платформе USENET [Shaoul, Westbury, 2013], и отобрали слова, которые чаще или реже встречаются в неправильном написании. Неопределенность, создающуюся за счет конкуренции между различными вариантами, было решено описывать при помощи такой меры, как энтропия (H) — ее апробация для лингвистических исследований была проведена в статье [Milin et al. 2009] и др. Получилось, что, например, для слова *innocent* ‘невинный’ $H=0,89$: 141 960 вхождений с правильным написанием и 62 665 вхождений (31 %) с ошибкой: *inocent*. У слова *necessary* ‘необходимый’ $H=0,14$ — ошибочный вариант *necessary* составляет всего 2 % от общего числа вхождений.

Затем было проведено два эксперимента. В первом участникам было предложено читать предложения, содержащие отобранные слова и создающие для них нейтральный контекст. Велась запись движений глаз. Во втором эксперименте использовался метод лексического решения: участники видели на экране цепочки букв и должны были как можно скорее ответить, являются ли они словами английского языка или псевдословами, нажав на одну из двух клавиш. И в первый, и во второй эксперимент целевые слова вошли в правильное написание (в качестве псевдослов использовались цепочки, не имеющие к ним отношения). В обоих экспериментах получилось, что фактор энтропии оказывает значимое влияние на скорость чтения слов — наравне с такими хорошо известными факторами, как

частотность и длина. Таким образом, даже когда слово написано правильно, наличие частотных орфографических ошибок затрудняет его обработку.

С. Рахманьян и В. Куперман также предположили, что эффект энтропии будет более сильным для слов с более высокой частотностью и для более начитанных участников. Чем чаще человек видит то или иное слово (за счет частотности слова или за счет того, что этот человек больше читает), тем лучше должна закрепиться информация о вероятностном распределении различных вариантов. Такая логика не кажется нам бесспорной: ведь у более редких слов и менее опытных читателей должно быть больше проблем с качеством лексических репрезентаций, что должно способствовать раздуванию эффекта энтропии. В любом случае, подтвердить или опровергнуть эти гипотезы пока не удалось.

Участники экспериментов проходили несколько тестов, призванных оценить общую начитанность, орфографическую и орфоэпическую грамотность и словарный запас, включая тест ART/MRT на знание фамилий авторов и названий журналов [Stanovich, West 1989], произносительный тест TOWRE [Torgeson et al. 1999] и орфографические тесты из работы [Andrews, Hersch 2010]. Однако, хотя результаты тестов коррелировали с общей скоростью чтения, ни в одном случае не удалось выявить их взаимодействия с фактором энтропии. Что касается фактора частотности, результаты двух экспериментов были противоположными: в первом эффект энтропии был больше для более частотных слов, а во втором — наоборот, для менее частотных.

Так как пока было сделано всего одно исследование на материале одного языка и некоторые результаты оказались противоречивыми, мы провели два эксперимента на русском языке. К сожалению, у нас нет корпуса неотредактированных текстов, сравнимого по объему с корпусом USENET [Shaoul, Westbury, 2013], что не позволяет с высокой точностью оценить энтропию для разных орфографических вариантов. Поэтому мы начали с того, что, опираясь на корпусные и другие источники, составили список слов, в которых особенно часто допускают орфографические ошибки (подробнее об этом см. в разделе 2). В первом эксперименте мы отобрали часть этих слов и подобрали к ним пары, в которых ошибки встречаются очень редко, сбалансировав слова попарно по длине и частотности. Мы сравнили скорость чтения в двух группах, используя метод лексического решения.

Второй эксперимент изначально задумывался как независимое исследование. Дело в том, что для русского языка есть много упражнений и проверочных заданий на знание орфографии, но нет таких тестов, как упомянутые выше для английского языка. Результат в них имеет числовое выражение, которое затем можно использовать в исследованиях, где орфографическая грамотность является одним из факторов. Так как нам был необходим такой тест для большого проекта, посвященного чтению на родном и иностранном языке, мы разработали его, используя упомянутый выше список «орфографически сложных» слов. Во время прохождения теста участник видит половину слов в правильном написании, половину — в частотном неправильном, его задача — отличить первые от вторых, нажимая на одну из двух клавиш.

Основная характеристика для тестов такого рода — достаточно дробные результаты, позволяющие эффективно ранжировать участников, и с этой точки зрения созданный нами тест оказался удачным. Однако, когда мы стали анализировать данные подробнее и задались вопросом, почему ошибки в одних словах замечаются намного чаще, чем в других, стало понятно, что результаты интересно сопоставить с первым экспериментом — особенно учитывая, что у С. Рахманьян и В. Купермана не было подобных данных для английского языка (в их работе целевые слова всегда предъявлялись в правильном написании). Ниже мы начинаем с анализа этих данных, а затем переходим к эксперименту с лексическим решением.

2. Эксперимент 1 с задачей определения ошибок

Устойчивость графического представления слов в памяти, или орфографическая зоркость, может быть оценена с помощью теста на узнавание правильно и неправильно написанных слов. Для английского языка такой инструмент, Spelling recognition test, был разработан С. Эндрус и коллегами [Andrews, Hersch, 2010]. Мы задались целью создать аналогичный тест на материале русского языка. В этой работе собранные нами данные анализируются не с точки зрения информативности теста (разброс между участниками, корреляция результатов с различными социолингвистическими переменными и др.), а чтобы выявить влияние ряда факторов на успешность распознавания правильных и ошибочных написаний.

2.1. Материал

Для первичного отбора материала мы опирались на данные различных ресурсов для подготовки к экзаменам по русскому языку, где перечислены трудные для написания слова, например, <http://slova.textologia.ru/>. Затем мы отобрали 44 слова, которые часто встречаются в неверном написании — как по данным основного подкорпуса Национального корпуса русского языка (НКРЯ, www.ruscorpora.ru), куда входят в основном редактируемые тексты, так и по данным подкорпуса «Социальные сети» корпуса «Тайга» [Shavrina, Kurmachova, n.d.], куда входят тексты нередактируемые (этот корпус больше подходит нам по содержанию, но еще не отредактирован окончательно для удаления повторяющихся фрагментов и т.д.). Ниже при анализе результатов используется информация о частотности, извлеченная из корпуса «Тайга», так как там намного шире представлены неправильные написания. В числе отобранных оказались слова с непроизносимыми согласными (*агентство, сердце, соответственно*), заимствованные слова с удвоенными согласными (*аппендицит, аккомпанемент, терраса*) и без таковых (*галерея, коридор*), слова, имеющие ошибочный вариант с эпентезой (*поскользнуться, педиатр, эспрессо, прецедент*), слова с орфограммами в приставках (*предыстория, расчет, чересчур*), корнях (*просвещение, помощник, участвовать*) и суффиксах (*серебряный, ветреный, девчонка*). Было составлено два протокола:

в первом одна случайно отобранная половина слов была предъявлена в верном написании, а другая — в частотном ошибочном, во втором протоколе — наоборот.

2.2. Участники

В исследовании на добровольной основе приняли участие 166 носителей русского языка: 136 женщин и 30 мужчин в возрасте от 18 до 42 лет.

2.3. Процедура

Участникам на экране компьютера предъявлялись слова и ставилась задача определить, написаны они верно или неверно. Слова были упорядочены по алфавиту. Участники не были ограничены во времени, скорость принятия решения не регистрировалась. Длительность эксперимента составила в среднем около 5 минут.

2.4. Анализ и результаты

Найти в стимуле определенную ошибку или удостовериться, что в нем нет ни одной ошибки, — две разные задачи, поэтому данные по правильным и неправильным написаниям надо анализировать отдельно. В силу ограниченного объема данной работы мы анализируем только данные по правильным написаниям — именно они важны в первую очередь для сравнения со вторым экспериментом и с предыдущими исследованиями. Был проведен статистический анализ с использованием смешанной логистической регрессии (GLMM). Зависимой переменной была точность ответа (верный / неверный), а независимыми — частотность ошибочных вариантов написания слова и общая его частотность в корпусе «Тайга» (фиксированные эффекты).

Был обнаружен эффект частотности слова ($b = 0,68$, $SE = 0,14$, $z = 4,68$, $p < 0,01$): чем частотнее слово, тем легче понять, что оно написано правильно (см. **Рис. 1**). Это ожидаемый результат. Более важно, что мы также обнаружили влияние частотности ошибочных вариантов ($b = -0,45$, $SE = 0,19$, $z = -2,42$, $p = 0,02$): чем частотнее у слова ошибочные написания, тем сложнее понять, что оно написано правильно (см. **Рис. 2**). Это можно объяснить изложенной выше теорией, согласно которой ошибочные варианты размывают репрезентацию слова в ментальном лексиконе. Важно подчеркнуть, что зависимость удалось установить несмотря на то, что информация об ошибочных написаниях, доступная для русского языка, намного беднее, чем та, что доступна для английского. Далее проверим, влияют ли они на *скорость* обработки правильно написанных слов, как это показали С. Рахманьян и В. Куперман для английского языка.

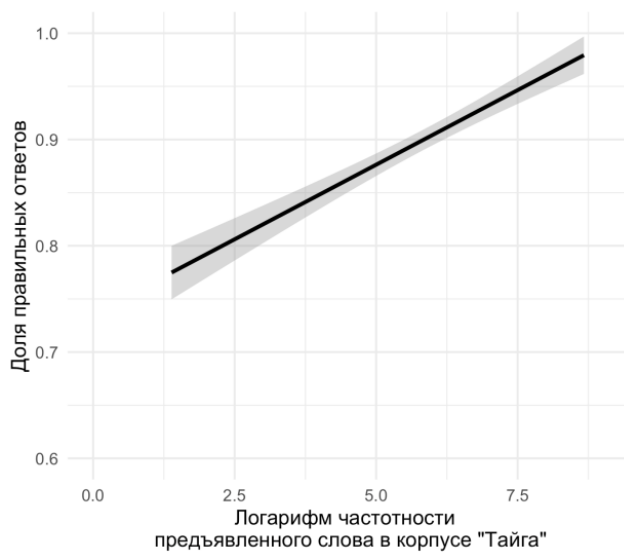


Рис. 1. Эффект частотности слова в Эксперименте 1

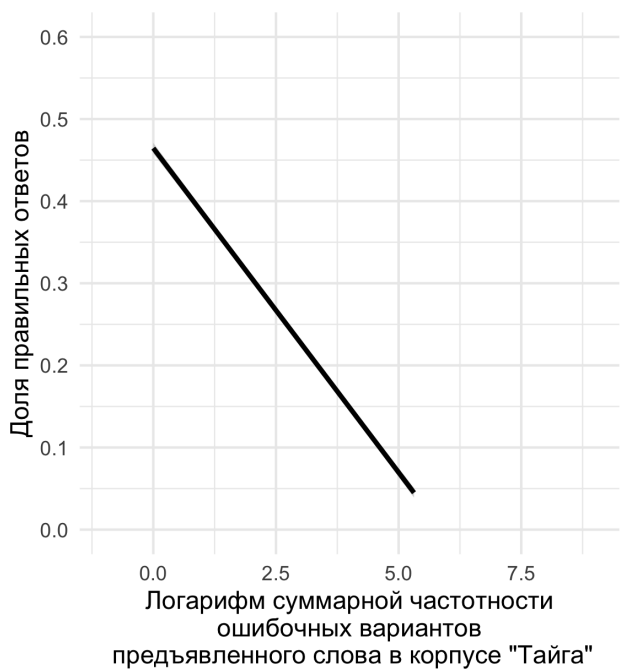


Рис. 2. Эффект частотности ошибочных вариантов в Эксперименте 1

3. Эксперимент 2 с задачей лексического решения

3.1. Материал

В качестве стимульного материала были отобраны 30 слов из числа использованных в предыдущем эксперименте. К каждому из этих слов, которые условно считались «сложными для написания», было подобрано в качестве пары «простое для написания» слово, то есть такое, где написание легко восстанавливается из звучания, а в НКРЯ и в корпусе «Тайга» не встречается ни одного неверного варианта. При этом «сложное» и «простое» слово совпадали по длине и имели минимальные отличия в частотности, например, *дилемма* (ipm 2,9) — *флагман* (ipm 2,1), *расчет* (ipm 108,9) — *клетка* (ipm 109,6)². В эксперименте все слова предъявлялись в правильном написании. Кроме этого, в стимульный материал было включено 60 псевдослов, на одну букву отличающихся от реально существующих.

3.2. Участники

В эксперименте на добровольной основе принял участие 41 носитель русского языка: 35 женщин и 6 мужчин в возрасте от 18 до 21 года, не участвовавшие в предыдущем исследовании.

3.3. Процедура

Была использована методика лексического решения: участник должен был определить, является ли предъявленная ему на экране последовательность букв словом русского языка. Эксперимент проводился на персональном компьютере с использованием программы PsychoPy [Peirce et al. 2019]. Сначала для фиксации взгляда в центре экрана на 500 мс предъявлялся символ «*», после чего также в центре экрана появлялась стимульная последовательность. Если испытуемый не принимал решения в течение 1 секунды, она сменялась символом «*» и следующим стимулом. Целевые слова и псевдослова предъявлялись в случайном порядке. Длительность эксперимента составляла около 10 минут.

3.4. Анализ данных и результаты

Никто из участников не ошибся более чем в 15% случаев, поэтому ничьи данные не были исключены из дальнейшего анализа. Мы анализировали время реакции испытуемых на целевые слова. При этом мы удалили те 149 случаев (6% данных), где испытуемые дали неверный ответ. Затем, пользуясь процедурой, описанной в работе [Ratcliff 1993], мы рассчитали для каждого

² Для подбора пары использовалась база данных StimulStat (http://stimul.cognitivestudies.ru/ru_stimul/, [Alexeeva et al. 2018]), данные о частотности лемм здесь и далее даны по словарю [Lyashevskaya, Sharov 2009].

экспериментального условия пороговые значения времени реакции (среднее время реакции плюс или минус 2,5 стандартных отклонения) и удалили все данные за пределами этих пороговых значений, а именно, 40 случаев, что составляет 1,7% от общего количества.

Для статистического анализа мы воспользовались смешанными линейными моделями (LMM)³. Значимым оказался эффект «тип слова» ($b = -0,07$, $SE = 0,03$, $t = -2,75$, $p = 0,01$): участникам исследования требовалось больше времени, чтобы распознать правильно написанное слово, имеющее частотные ошибочные варианты написания, чем правильно написанное слово, которое таких вариантов не имеет (см. **Рис. 3**). Это подтверждает общую гипотезу С. Рахманьян и В. Купермана [Rahmanian, Kuperman 2019] на новом материале.

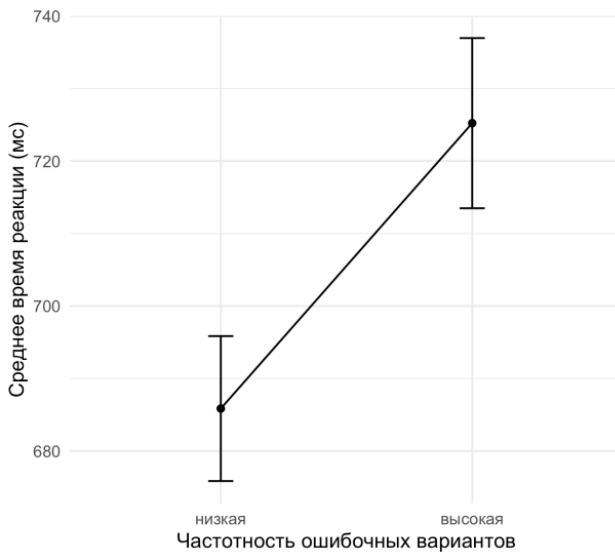


Рис. 3. Эффект наличия ошибочных вариантов в Эксперименте 2

Затем мы провели дополнительный анализ: помимо фактора «тип слова» (сложное / простое) модель включала факторы «длина слова» (количество букв) и «частотность слова» (высокая / низкая), а также взаимодействия между типом и частотностью/длиной. Разделение на высоко- и низкочастотные слова осуществлялось на основе медианного значения (средняя частотность в первой группе равнялась 88,5 ipm, а во второй — 9,5 ipm). Были обнаружены эффекты длины ($b = 0,04$, $SE = 0,006$, $t = 4,07$, $p < 0,01$) и частотности слова ($b = 0,12$, $SE = 0,026$, $t = 4,79$, $p < 0,01$): высокочастотные слова обрабатываются быстрее низкочастотных, а короткие — быстрее длинных. Так же, как и в предыдущем

³ Во всех анализах с использованием LMM зависимая переменная была введена в модель в трансформированном виде (использовалась обратная трансформация). Это было необходимо для того, чтобы так называемые остатки модели (та вариативность, которую не удалось предсказать) были распределены нормально.

анализе, был значим фактор «тип слова», т. е. наличие частотных ошибочных вариантов ($b = -0,07$, $SE = 0,026$, $t = -2,70$, $p < 0,01$). Однако более важно, что мы получили значимое взаимодействие между частотностью и типом слова ($b = -0,11$, $SE = 0,052$, $t = -2,16$, $p = 0,03$): репрезентация слова в ментальном лексиконе страдает от ошибочных вариантов сильнее, если слово низкочастотное (см. **Рис. 4**). Это противоречит гипотезе С. Рахманьян и В. Купермана [Rahmanian, Kuperman 2019] о роли частотности, но согласуется с данными, которые они получили в аналогичном эксперименте. Взаимодействие между длиной и типом слова не было значимым.

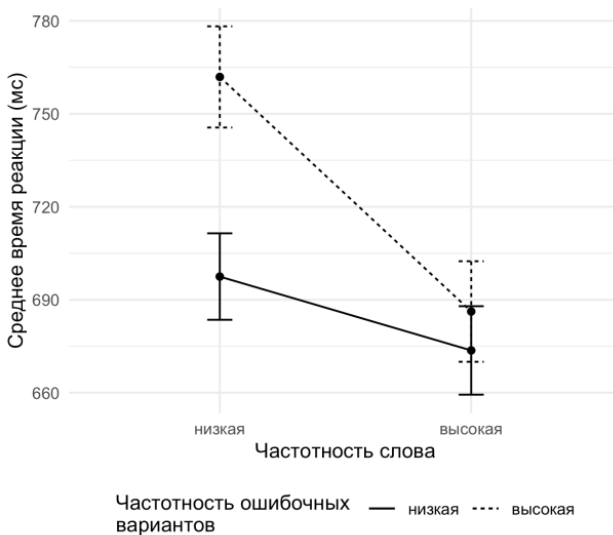


Рис. 4. Эффект частотности слова и наличия ошибочных вариантов в Эксперименте 2

Наконец мы задались вопросом, связана ли легкость распознавания правильного написания слова в первом эксперименте со временем его прочтения во втором. Для этого мы отобрали времена реакции на слова, которые имеют частотные ошибочные варианты — только они были представлены в первом эксперименте. Статистический анализ зафиксировал значимое влияние успешности определения правильного написания на скорость обработки ($b = -0,05$, $SE = 0,19$, $t = -2,82$, $p = 0,01$): чем больше правильных ответов слово получает в тесте на орфографическую грамотность, тем быстрее оно обрабатывается в эксперименте на лексическое решение (см. **Рис. 5**).

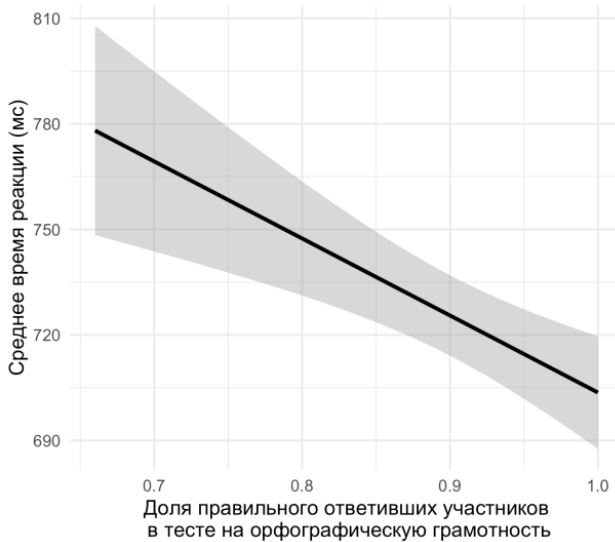


Рис. 5. Влияние успешности распознавания правильного написания в Эксперименте 1 на время реакции в Эксперименте 2

4. Заключение

Орфографическим ошибкам посвящено много исследований, но до последнего времени изучались только закономерности их возникновения при порождении. Между тем, из гипотезы лексического качества ([Perfetti 1985], [2007] и др.), согласно которой частые ошибки свидетельствуют о более размытой репрезентации слова в ментальном лексиконе, и из общих моделей обучения (например, [Baayen et al. 2011]; [Ramsar, Dye, McCauley 2013]; [Rescorla, Wagner 1972]) следует, что слова, в которых чаще допускают ошибки, должны вызывать повышенные трудности и при восприятии. Эта гипотеза впервые была протестирована С. Рахманьян и В. Куперманом [Rahmanian, Kuperman 2019] в двух экспериментах на материале английского языка. В целом она подтвердилась, однако некоторые результаты оказались противоречивыми.

Чтобы проверить эту гипотезу на другом материале и разрешить возникшие противоречия, мы провели два эксперимента на материале русского языка. Основная сложность заключалась в том, что только для английского языка есть корпуса, содержащие огромные массивы неотредактированных текстов. Тем не менее, частотности правильных и неправильных написаний, извлеченные из подкорпуса «Социальные сети» корпуса «Тайга» [Shavrina, Kurmachova, n. d.], показали значимую корреляцию с результатами первого эксперимента, в котором участникам надо было определять слова, написанные правильно и с орфографическими ошибками, и позволили эффективно разделить стимулы во втором эксперименте на две группы: орфографически сложные и простые.

Оба эксперимента подтверждают гипотезу о том, что слова, в которых чаще допускают орфографические ошибки, оказываются более сложными и при восприятии — даже когда они написаны правильно. Такие слова медленнее читаются — результат, аналогичный полученному С. Рахманьян и В. Куперманом на материале английского языка, — а оценка правильности их написания вызывает больше сложностей. Результаты двух экспериментов показали значимую корреляцию друг с другом.

С. Рахманьян и В. Куперман предположили, что этот эффект будет сильнее для высокочастотных слов, которые мы видим чаще, чем для низкочастотных, однако их эксперименты дали противоречивые результаты. Проведенный нами эксперимент свидетельствует в пользу обратной картины: эффект значимо сильнее для слов с более низкой частотностью. Это можно объяснить тем, что графический облик высокочастотных слов более устойчив в ментальном лексиконе и менее подвержен «размыванию» ошибочными вариантами написания.

Литература

1. Alexeeva S., Slioussar N., Chernova D. (2018), StimulStat: a lexical database for Russian. *Behavior Research Methods*, Vol. 50, pp. 2305–2315.
2. Andrews S., Hersch J. (2010), Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, Vol. 139, p. 299.
3. Baayen R. H., Milin P., Durdevic D. F., Hendrix P., Marelli M. (2011), An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, Vol. 118, pp. 438–481.
4. Lyashevskaya O., Sharov S. (2009), The frequency dictionary of modern Russian language based on Russian National Corpus [Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Nacional'nogo korpusa russkogo jazyka)]. Moscow: Azbukovnik.
5. Milin P., Kuperman, V., Kostic A., & Baayen R. H. (2009), Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in grammar: Form and acquisition*, pp. 214–252.
6. Nelson Taylor J., Perfetti C. A. (2016), Eye movements reveal readers' lexical quality and reading experience. *Reading and Writing*, Vol. 29, pp. 1069–1103.
7. 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.
8. Perfetti C. A. (1985), *Reading ability*. Oxford, Oxford University Press.
9. Perfetti C. A. (2007), *Reading ability: Lexical quality to comprehension*. *Scientific studies of reading*, Vol. 11, pp. 357–383.
10. Perfetti C. A., Hart L. (2001), The lexical basis of comprehension skill. On the consequences of meaning selection: Perspectives on resolving lexical ambiguity, American Psychological Association, Washington, DC, pp. 67–86.

11. *Perfetti C. A., Hart L.* (2002), The lexical quality hypothesis. Precursors of functional literacy, Vol. 11, pp. 67–86.
12. *Rahmanian S., Kuperman V.* (2019), Spelling errors impede recognition of correctly spelled word forms. *Scientific Studies of Reading*, Vol. 23, pp. 24–36.
13. *Ramscar M., Dye M., McCauley S. M.* (2013), Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, Vol. 89, pp. 760–793.
14. *Ratcliff R.* (1993), Methods for dealing with reaction time outliers. *Psychological Bulletin*, Vol. 114, pp. 510–532.
15. *Rescorla R. A., Wagner A. R.* (1972), A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, Vol. 2, pp. 64–99.
16. *Shaoul C., Westbury C.* (2013), A reduced redundancy USNET corpus (2005–2011). University of Alberta, 39, pp. 850–863.
17. *Shavrina T., Kurmachova Y.* (n. d.), Taiga Corpus. An open-source corpus for machine learning, available at: https://tatianashavrina.github.io/taiga_site/.
18. *Stanovich K. E., West R. F.* (1989), Exposure to print and orthographic processing. *Reading Research Quarterly*, Vol. 24, pp. 402–433.
19. *Torgeson J. K., Wagner R. K., Rashotte C. A.* (1999), Test of word reading efficiency (TOWRE). Austin, TX, ProEd.

ОБ ОСОБЕННОСТЯХ ВТОРИЧНОЙ ИМПЕРФЕКТИВАЦИИ ГЛАГОЛОВ С ПРЕФИКСОМ *ПО-* В РУССКОМ ЯЗЫКЕ¹

Чуйкова О. Ю. (ochuikova@herzen.spb.ru, oxana.chuykova@gmail.com)

РГПУ им. А. И. Герцена, Санкт-Петербург, Россия

DOI: 10.28995/2075-7182-2020-19-160-176

ON THE SECONDARY IMPERFECTIVATION OF *PO*-PERFECTIVES IN RUSSIAN

Chuikova O. Iu. (ochuikova@herzen.spb.ru, oxana.chuykova@gmail.com)

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

The paper deals with a number of characteristics of the secondary imperfectivation of *po*-perfectives in Russian. The study is based on the analysis of the level of imperfectivability of Russian perfective verbs with the prefix *po-* compared to a number of other prefixed perfective verb groups (e. g. the verbs with such perfectivizing prefixes as *na-*, *za-*, etc.) according to the Dictionary of Russian Language, the Russian National Corpus and the Russian-language Internet (Runet). It is shown that the discussed perfective verb group is specific as a whole as well as with respect to its subgroups, i. e., deperfective perfective verbs and morphologically marked Aktionsarten. *Po*-perfectives demonstrate a low average imperfectivability in comparison to corresponding figures for other prefixed verb groups. For the subgroup of deperfective (formed from perfective stems) verbs the level of imperfectivability is also unusually low. The delimitative Aktionsart shows a higher imperfectivability than other morphologically marked Aktionsarten do. Possible explanations for the peculiarities of imperfectivability of *po*-perfectives rather confirm than contradict the hypothesis about the regularity of the secondary imperfectivation in Russian

Key words: Russian language, verbal, morphology, aspect, perfective verbs, imperfectivation, Aktionsarten, corpus study

¹ Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-012-00143 «Суффиксальная имперфективация приставочных перфективов в современном русском языке».

1. Вводные замечания

В работе приводятся наблюдения, касающиеся имперфективируемости глагольных лексем с префиксом *по-*. Анализ показывает, что рассматриваемая префиксальная группа демонстрирует ряд особенностей по сравнению с рядом других префиксальных групп в русском языке. Основанием для сравнения служат данные по группам глаголов с префиксами *про-*, *у-* *с(о)-*, изложенные в статье [Горбова 2019а], а также полученные по той же методике данные для *на-* и *за-* глаголов.

2. Методика исследования и количественные данные

Для определения степени имперфективируемости перфективных глаголов с префиксом *по-* применялась методика, изложенная в работе [Горбова 2019а]. По Малому академическому словарю (далее — МАС) [Евгеньева (ред.) 1981–1984] методом сплошной выборки был получен список перфективных глагольных лексем² с соответствующим префиксом. Далее наличие имперфективного коррелята, образованного посредством суффиксальной имперфективации (вторичного имперфектива — НСВ₂), определяется по трем источникам. В первую очередь, осуществляется поиск соответствующего коррелята НСВ в МАС. В случае отсутствия искомого глагола в словаре, привлекаются данные Национального корпуса русского языка (НКРЯ) и русскоязычного сегмента сети Интернет (путем запросов в поисковых системах Google и Яндекс). Обращение к обоим источникам реализуется во всех случаях отсутствия глагола НСВ в МАС³, что обусловлено вероятностью отсутствия примеров употребления лексем в Интернете даже при наличии ее в НКРЯ. Далее оценивается возрастание уровня имперфективируемости при учете данных НКРЯ и русскоязычного Интернета (суммируются данные о количестве НСВ₂ в МАС и НКРЯ и в МАС и Google/Яндекс).

Коллекция выборки в рамках префиксальной группы формируется по тем же принципам, что и в [Горбова 2019а]. В первую очередь, из состава исходной выборки исключается и далее рассматривается отдельно группа отыменных перфективов. Обособление данной группы продиктовано тем обстоятельством, что используемая здесь исследовательская процедура представляет собой модификацию методики исследования, предложенной В. С. Храковским при изучении перфективов с префиксом *про-* [Храковский 2018], где отыменные перфективы остаются за пределами рассмотрения. Далее в составе полученной (максимальной) выборки выделяются подмножества, в которых, по предварительным наблюдениям, фиксируется существенно больший или

² Следует отметить, что под «лексемой» понимается каждое лексическое значение, фиксируемое в словаре. В случае вхождения в одну словарную статью более одного лексического значения имперфективируемость определяется для каждого из них отдельно.

³ В отличие от процедуры, описанной в статье [Горбова 2019а], где обращение к сети Интернет производилось лишь в случаях, когда поиск по НКРЯ не давал результатов или для искомого глагола НСВ было получено менее пяти вхождений.

меньший уровень имперфективируемости: отперфективные перфективы, глаголы морфемно характеризованных способов действия (далее — СД) и глаголы на *-и(зи)рывать* (так, например, согласно наблюдениям в [Горбова 2019а], отперфективные перфективы, как и отыменные, характеризуются повышенным уровнем имперфективации, а глаголы на *-и(зи)рывать* и глаголы ряда морфемно характеризованных СД — сниженным). Подмножество глаголов, полученное при изъятии из максимальной выборки перечисленных специализированных выборов, представляет собой минимальную выборку.

Для группы перфективных глаголов с префиксом *по-* можно выделить четыре СД (согласно классификации, приведенной в [Зализняк, Шмелев 2000]): аттенуативный (*поразвлечься, поубавиться*), делимитативный (*покашлять, порезвиться*), дистрибутивный (*понатаскать, поубивать* (о всех, многих)), ингрессивный (*побежать, помчаться*). Глаголы всех перечисленных способов действия, согласно МАС, являются непарными глаголами СВ. В рамках настоящего исследования принято решение считать результатом имперфективации делимитативов глаголы НСВ, относящиеся к прерывисто-смягчительному СД (в том числе лексемы, зафиксированные в МАС). Аргументы в пользу такого решения см. в [Пазельская, Татевосов 2008: 364]; [Федотов, Чуйкова 2013]; [Горбова 2019b]⁴.

Отметим, что в некоторых случаях наблюдается пересечение выборов. Так, глаголы аттенуативного СД полностью, а глаголы дистрибутивного СД — частично, входят в состав отперфективных глаголов, ряд глаголов на *-и(зи)рывать* относится к делимитативному СД.

Приведем коллекцию выборов, демонстрирующих уровень имперфективируемости *по-*глаголов (соответствующие данные приведены также в **Таблицах 1–3** в **Приложении** к настоящей статье):

2.1. По-глаголы

1. **Отыменные** *по-*перфективы (*повысить*): доля НСВ₂ в рамках МАС — 87,5 % (28 из 32); то же в МАС + НКРЯ — 87,5 % (28 из 32); то же в МАС + Яндекс и Google — 90,63 % (29 из 32); доля в исходной выборке — 1,8 % (32 из 1781).

⁴ Случаи объединения глаголов делимитативного и прерывисто-смягчительного СД в видовую пару обнаруживаются и в лексикографической практике. Так, например, в Толковом словаре русского языка [Ушаков (ред.) 1935–1940] устанавливаются пары типа *погрохотать* — *погрохотывать* (НСВ отсутствует в МАС): *погрохотать* — произвести грохот, шум с раскатами в течение некоторого времени, *погрохотывать* — (разг.) несовер. к *погрохотать*.

В то же время очевидно различие между глаголами прерывисто-смягчительного СД как НСВ₂ от делимитативов и «прототипическими» вторичными имперфективами (как, например, в паре *переписать*–*переписывать*). Образования типа *почитывать*, *покуривать* характеризуются композициональностью семантики (где префикс *по-* несет значение делимитатива, а суффикс *-ыва-/-ива-* — итеративную семантику). При этом глаголы в парах *почитать* — *почитывать*, *покурить* — *покуривать* не в полной мере удовлетворяют критерию видовой парности.

2. **Максимальная** выборка (без отыменных, далее — тах-выборка): доля НСВ₂ в рамках МАС — 22,47% (393 из 1749); то же в МАС + НКРЯ — 39,34% (688 из 1749); то же в МАС + Яндекс/Google — 65,47% (1145 из 1749).
3. **Отперфективные по-**перфективы (*позабыть*): доля НСВ₂ в рамках МАС — 12,3% (23 из 187); то же в МАС + НКРЯ — 27,81% (52 из 187); то же в МАС + Яндекс и Google — 49,73% (93 из 187); доля в тах-выборке — 10,69% (187 из 1749). Из них **аттенуативный СД** (*попривыкнуть*): доля НСВ₂ в МАС — 0% (0 из 62); то же в МАС + НКРЯ — 17,74% (11 из 62); то же в МАС + Яндекс/Google — 48,39% (30 из 62); доля в тах-выборке — 3,54% (62 из 1749).
4. Глаголы на **-и(зу)рывать** (*подискутировать*): доля НСВ₂ в рамках МАС — 0% (0 из 12); то же в МАС + НКРЯ — 0% (0 из 12); то же в МАС + Яндекс и Google — 16,67% (2 из 12); доля в тах-выборке — 0,67% (12 из 1749).
5. **Делимитативный СД** (*поахать*): доля НСВ₂ в МАС — 20,49% (117 из 571); то же в МАС + НКРЯ — 43,43% (248 из 571); то же в МАС + Яндекс/Google — 73,03% (417 из 571); доля в тах-выборке — 32,65% (571 из 1749).
6. **Дистрибутивный СД** (*полечь*): доля НСВ₂ в МАС — 0% (0 из 168); то же в МАС + НКРЯ — 15,48% (26 из 168); то же в МАС + Яндекс/Google — 36,9% (62 из 168); доля в тах-выборке — 9,61% (168 из 1749).
7. **Ингрессивный СД** (*погнать*): доля НСВ₂ в МАС — 0% (0 из 69); то же в МАС + НКРЯ — 13,04% (9 из 69); то же в МАС + Яндекс/Google — 37,68% (26 из 69); доля в тах-выборке — 3,95% (69 из 1748).
8. **Минимальная** выборка (далее — min-выборка) (без отперфективных, глаголов на *-и(зу)рывать*, СД): доля НСВ₂ в рамках МАС — 31,08% (253 из 814); то же в МАС + НКРЯ — 43,49% (354 из 814); то же в МАС + Яндекс и Google — 68,92% (561 из 814).

Пересечения: отперфективные и дистрибутивный СД — 61, глаголы на *-и(зу)рывать* и делимитативный СД — 11.

Для сопоставления приведем также аналогичные данные для глаголов с префиксами *за-* и *на-*:

2.2. За-глаголы

1. **Отыменные за-**перфективы (*заглубить*): доля НСВ₂ в рамках МАС — 70% (42 из 60); то же в МАС + НКРЯ — 73,33% (44 из 60); то же в МАС + Яндекс и Google — 93,3% (56 из 60); доля в исходной выборке — 2,78% (60 из 2157).
2. **Мах-**выборка (без отыменных): доля НСВ₂ в рамках МАС — 58,27% (1222 из 2097); то же в МАС + НКРЯ — 64,95% (1362 из 2097); то же в МАС + Яндекс и Google — 80,01% (1679 из 2097).
3. **Отперфективные за-**перфективы (*закупить*): доля НСВ₂ в рамках МАС — 90,47% (95 из 105); то же в МАС + НКРЯ — 94,28% (99 из 105); то же в МАС + Яндекс и Google — 98,09% (103 из 105); доля в тах-выборке — 5,01% (105 из 2097).
4. Глаголы на **-и(зу)рывать** (*замаскировать*): доля НСВ₂ в рамках МАС — 18,18% (12 из 66); то же в МАС + НКРЯ — 19,7% (13 из 66); то же в МАС

+ Яндекс и Google — 68,18% (45 из 66); доля в тах-выборке — 3,15% (66 из 2097).

5. **Начинательный СД** (*заалеть*): доля НСВ₂ в рамках МАС — 2,29% (11 из 480); то же в МАС + НКРЯ — 8,96% (43 из 480); то же в МАС + Яндекс и Google — 39,58% (190 из 480); доля тах-выборке — 22,89% (480 из 2097).
6. **Чрезмерно-длительный СД** (с циркумфиксом *за-...-ся: забегаться*): доля НСВ₂ в рамках МАС — 50% (26 из 52); то же в МАС + НКРЯ — 57,69% (30 из 52); то же в МАС + Яндекс и Google — 84,62% (44 из 52); доля в тах-выборке — 2,48% (52 из 2097).
7. **Min-выборка**: доля НСВ₂ в рамках МАС — 76,95% (1078 из 1401); то же в МАС + НКРЯ — 84,01% (1177 из 1401); то же в МАС + Яндекс и Google — 92,72% (1299 из 1401).

Пересечения: глаголы на *-и(зи)рывать* и начинательный СД — 6, глаголы на *-и(зи)рывать* и чрезмерно-длительный СД — 1.

2.3. На-глаголы

1. **Отыменные на-перфективы** (*наводнить*): доля НСВ₂ в рамках МАС — 88,89% (8 из 9); то же в МАС + НКРЯ — 88,89% (8 из 9); то же в МАС + Яндекс и Google — 88,89% (8 из 9); доля в исходной выборке — 0,61% (9 из 1471).
2. **Мах-выборка** (без отыменных): доля НСВ₂ в рамках МАС — 61,97% (906 из 1462); то же в МАС + НКРЯ — 68,47% (1001 из 1462); то же в МАС + Яндекс и Google — 85,09% (1244 из 1462).
3. **Отперфективные на-перфективы** (*накупить*): доля НСВ₂ в рамках МАС — 88,16% (67 из 76); то же в МАС + НКРЯ — 92,11% (70 из 76); то же в МАС + Яндекс и Google — 93,42% (71 из 76); доля тах-выборке — 5,2% (76 из 1462).
4. Глаголы на *-и(зи)рывать* (*набальзамировать*): доля НСВ₂ в рамках МАС — 25% (4 из 16); то же в МАС + НКРЯ — 25% (4 из 16); то же в МАС + Яндекс и Google — 56,25% (9 из 16); доля тах-выборке — 1,09% (16 из 1462).
5. **Кумулятивный СД** (*наворовать*): доля НСВ₂ в рамках МАС — 61,93% (257 из 415); то же в МАС + НКРЯ — 67,47% (280 из 415); то же в МАС + Яндекс и Google — 83,61% (347 из 415); доля в тах-выборке — 28,39% (415 из 1462).
6. **Сатуративный СД** (с циркумфиксом *на-...-ся: нарезать*): доля НСВ₂ в рамках МАС — 20,54% (38 из 185); то же в МАС + НКРЯ — 29,19% (54 из 185); то же в МАС + Яндекс и Google — 70,27% (130 из 185); доля в тах-выборке — 12,65% (185 из 1462).
7. **Min-выборка**: доля НСВ₂ в рамках МАС — 70,02% (549 из 784); то же в МАС + НКРЯ — 76,91% (603 из 784); то же в МАС + Яндекс и Google — 88,9% (697 из 784).

Пересечения: отперфективные и кумулятивный СД — 8, глаголы на *-и(зи)рывать* и сатуративный СД — 6.

3. Анализ результатов

В рамках настоящей работы остановимся на таких признаках, как общая имперфективируемость префиксальной группы (раздел 3.1), имперфективируемость глаголов, относящихся к морфемно характеризованным СД (раздел 3.2), и отперфективных *по*-глаголов (раздел 3.3). В отношении отыменных глаголов и глаголов на *-и(зи)ровать* данная префиксальная группа не демонстрирует отличительных особенностей: как и в других приставочных группах, для отыменных *по*-глаголов наблюдается более высокий, а для глаголов на *-и(зи)ровать* — сниженный уровень имперфективируемости относительно данных по префиксальной группе в целом.

3.1. Общая тенденция возрастания уровня имперфективируемости, отмеченная для ряда приставочных групп, наблюдается и в случае с *по*-глаголами. В то же время в целом можно говорить о сниженной имперфективируемости перфективных *по*-глаголов в сравнении с рядом других приставочных групп, например, *про-* *с(о)-*, *у-* [Горбова 2019а], *на-*, *за-*, что наблюдается как по максимальной выборке, включающей в себя все специализированные выборки (кроме отыменных), см. рис. 1, так и по минимальной выборке, см. рис. 2.

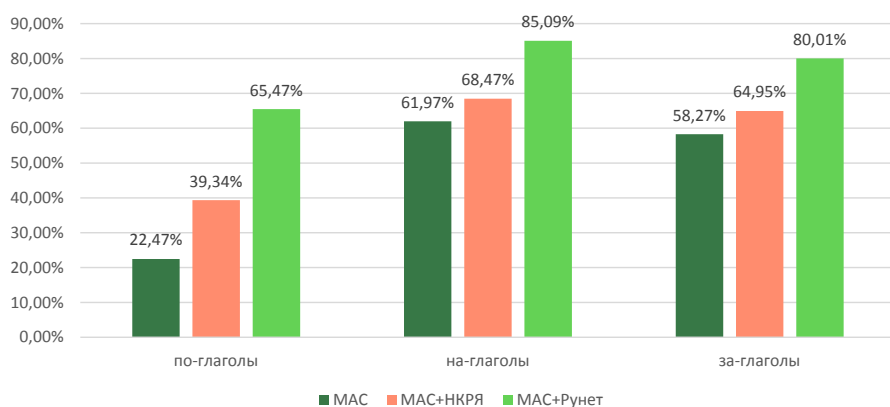


Рис. 1. Уровни имперфективируемости *по*-, *на*- и *за*-глаголов по max-выборке

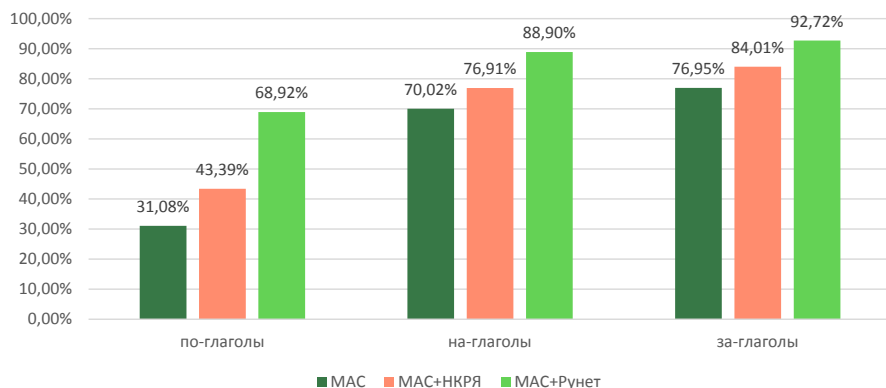


Рис. 2. Уровни имперфективируемости *по-*, *на-* и *за-*глаголов по *min*-выборке

Как видно из **рис. 1** и **2** (а также в сопоставлении с данными для *про-*, *со-* и *и-* глаголов), уровень имперфективируемости *по-* глаголов по всем трем источникам оказывается ниже, чем у других рассмотренных префиксальных групп. Можно выдвинуть некоторые предположения о причинах ограниченной способности *по-* перфективов к образованию НСВ₂. Упомянем неоднократно отмечаемую в литературе продуктивность префикса *по-* (см., например, [Черткова 1996]; [Dickey 2006]) и, в частности, его широкое использование при образовании перфективных коррелятов для бесприставочных глаголов НСВ. Так, например, для 248 глаголов (30,13 %) в составе минимальной выборки дается толкование через соответствующий симплекс: *сов.* к 'НСВ₁', НСВ₂ для таких перфективов с префиксом *по-* в МАС, как правило, отсутствует (в НКРЯ и рунете удалось обнаружить соответственно 38 и 134 не зафиксированных в МАС НСВ₂, входящих в видовые тройки: *белить* — *побелить* — *побеливать*, *венчать* — *повенчать* — *повенчивать*, *худеть* — *похудеть* — *похудевать* и т. д.).

Для *по-* глаголов различие между имперфективируемостью по МАС и показателями, полученными при привлечении данных НКРЯ и рунета, больше, чем по другим приставочным группам. Таким образом, для перфективных *по-* глаголов можно констатировать наибольшую долю отсутствующих в МАС неконвенциональных имперфективов, то есть при минимальных словарных показателях здесь также наблюдается стремление к регулярности, то есть к образованию НСВ₂.

3.2. Посредством префикса *по-* образуются глаголы следующих СД: делимитативного (*поахать*, *побегать*, *погрустить*), дистрибутивного (*повыгнать*, *пожечь* <все / многое или всех / многих>), аттенуативного (*попривыкнуть*, *поусомниться*) и ингрессивного (*поехать*, *пошлепать* 'пойти шлепая'). Данные об имперфективируемости перечисленных СД приведены на **рис. 3**. Для всех перечисленных СД наблюдается существенное возрастание имперфективируемости в НКРЯ и рунете.



Рис. 3. Имперфективируемость *по*-глаголов СД в сравнении с *по*-глаголами мин-выборки

Глаголы аттенуативного, дистрибутивного и ингрессивного СД демонстрируют сниженный уровень имперфективируемости в сопоставлении с данными по минимальной выборке, что соответствует общей картине, наблюдаемой для глаголов морфемно характеризованных СД [Горбова 2019а], ср. аналогичные данные для СД с префиксами *за*- и *на*- (рис. 4 и 5), где имперфективируемость СД также оказывается (в той или иной мере) ниже показателей по минимальной выборке по всем трем источникам.

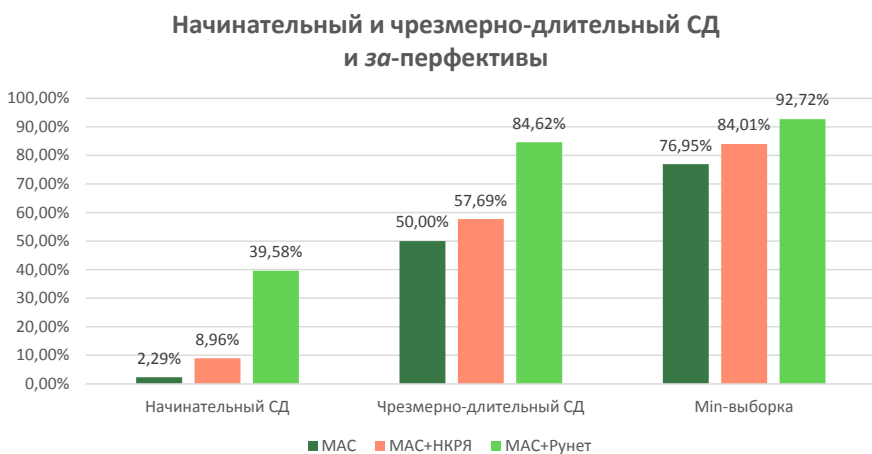


Рис. 4. Имперфективируемость *за*-глаголов СД в сравнении с *за*-глаголами мин-выборки

Кумулятивный и сатуративный СД и *на*-перфективы

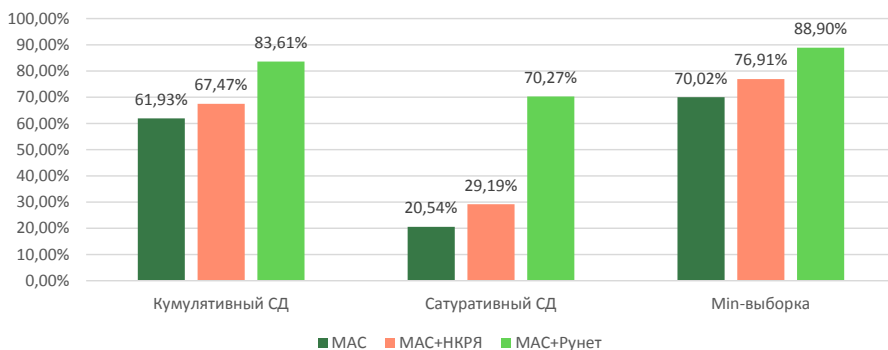


Рис. 5. Имперфективируемость *на*-глаголов СД в сравнении с *на*-глаголами мин-выборки

Как видно из **рис. 3**, глаголы делимитативного СД выделяются в ряду способов действия с префиксом *по-*: 1) это наиболее многочисленный СД (571 лексема, 32,65% в максимальной выборке), 2) имперфективируемость относящихся к данному СД глаголов по всем трем источникам выше, чем у других СД с префиксом *по-*. По данным НКРЯ уровень имперфективируемости *по*-делимитативов равен, а по данным рунета несколько превышает соответствующие показатели минимальной выборки. Глаголы прерывисто-смягчительного СД, рассматриваемые как результат имперфективации делимитативов, учтены в МАС последовательно (с регулярной пометой «разг.»), о чем может свидетельствовать резкое возрастание количества глагольных лексем, зафиксированных в НКРЯ и рунете по сравнению с имеющимися в словаре: более чем в 2 раза (22,94% — 131 лексема) по НКРЯ, и более чем в 3,5 раза (52,54% — 300 лексем) по данным рунета. Таким образом, образование прерывисто-смягчительного СД представляет собой значительно более активный процесс, чем отражает МАС. Примеры (1)–(2) демонстрируют случаи употребления отсутствующих в МАС лексем.

- (1) *Тихон Ильич, нахмуриваясь, побарабанивал пальцем по столу.*
[И. А. Бунин. Деревня (1909–1910)]
- (2) *Дедуле пора отдохнуть, мемуарчики пописывать, правнучков понячживать, на печечке опочивать.* (<https://echo.msk.ru/news/1093268-echo.html>)

Наименее многочисленный СД в составе группы глаголов с префиксом *по*- составляет редко упоминаемый в литературе аттенуативный СД. В подробной классификации способов глагольного действия, разработанной М. А. Шелякиным [2008: 141–167], указанный СД (споконателем *по-*) отсутствует. Согласно [Зализняк, Шмелев 2000: 120], глаголы аттенуативного (=смягчительного) СД образуются от приставочных глаголов СВ: *пообсохнуть, пораздумать*. Как представляется, к данному СД также следует относить ряд глаголов, образованных от неперифигированных перфективов с семельфактивным суффиксом *-ну-* и соответствующей

смягчительной семантикой: *покачнуть* 'слегка качнуть', *пошевелинуть* 'чуть-чуть шевельнуть', *пошелохнуться* 'чуть-чуть шелохнуться'. Формант *-ну-* является единственным суффиксальным средством перфективации в русском языке, выполняющим функцию, аналогичную перфективирующим префиксам. Префигированным глаголам с семельфактивным суффиксом *-ну-* относительно регулярно соответствуют вторичные имперфективы (*покачивать*, *пошевеливать*), не содержащие в своем составе семельфактивного суффикса⁵. Глаголы, содержащие в своем морфемном составе два префикса имперфективируются редко, однако примеры такой имперфективации в обследуемых источниках языкового материала также обнаруживаются, см. (3)–(4).

- (3) *Женщина чувствовала, что он нашел ее ущербинку, и сразу же в ней поубавлялось гордыни, и сразу она как-то начинала нервничать и из-за этого нервничания как бы начинала понимать, что и Вова не так уж плох, коль она сама не без изъяна.* [Вацлав Михальский. Весна в Карфагене (2001)]
- (4) *Стал уж было я поуспокаиваться, как вдруг прохожу раз по базару, вижу — сидит баба и торгует пряниками.* (https://www.e-reading.mobi/txt.php/1018140/худлит-Юмор_начала_XX_века.txt)

Следует сказать несколько слов об имперфективируемости дистрибутивного СД. Согласно классификации префиксов, предложенной в [Татевосов 2013], дистрибутивный префикс *по-* определяется как левопериферийный, что предполагает следующее: 1) «Дистрибутивный *по-* не имеет ограничений на формальную (им)перфективность основы» (с. 39), 2) «Дистрибутивный *по-* присоединяются выше, чем показатель вторичного имперфектива *-ыва-*» (с. 40) и 3) «Дистрибутивный *по-*: имперфективация невозможна» (с. 41). Из этого следует, что имперфективные основы с дистрибутивным префиксом *по-* не способны получать имперфективную трактовку. Тем не менее, как видно из рис. 3, в рассматриваемых источниках обнаруживается некоторое количество

⁵ Такие НСВ₂, по-видимому, неправомерно считать результатом имперфективации семельфактивов, в данном случае мы вероятнее всего можем наблюдать образование НСВ₂ от глаголов делимитативного СД (*пошевелить*, *покачать*), а семантическое соответствие между семельфактивами и НСВ₂ обеспечивается близостью значений делимитативного и аттенуативного СД (см. [Зализняк, Шмелев 2000: 120–121]). Возможно, семельфактивы с *-ну-* в принципе следовало бы вывести за пределы рассмотрения. Однако в данном случае возникает также вопрос о том, следует ли исключать из рассмотрения и глаголы с омонимичным суффиксом *-ну-*, не являющиеся семельфактивами (*поблкнуть*, *повиснуть*, *повянуть*, *погибнуть*, *пожухнуть*, *померзнуть* и др.). В составе приставочного имперфективного коррелята глаголов данной группы также последовательно отсутствует суффикс *-ну-*, при том, что НСВ₂ невозможно вывести к бесприставочному НСВ. Данный вопрос требует дальнейшего подробного изучения. Несмотря на дискуссионный характер данного решения, на текущем этапе исследования образования с суффиксом *-ну-* (в том числе семельфактивы) не исключались из числа рассматриваемых глаголов, а соответствующие им НСВ₂ считаются результатом их имперфективации. Следует также сказать, что немногочисленность данной группы глаголов позволяет говорить о том, что ее изъятие из выборки не привело бы к существенному изменению результатов исследования.

имперфективированных глаголов дистрибутивного СД, см. примеры (5)–(6). Все обнаруженные случаи находятся за пределами грамматической нормы.

- (5) *Сама скатерть обед накрывает и вино наливает, а затем сама бранится и посуду **побивает**.* (<https://books.google.ru/books?isbn=545752471X>)
- (6) ***Поворачивают** все что лежит и висит, как и бытовые вентиляторы, устанавливаемые на верхних этажах в каналах кухонь и с/у.* (<https://www.domkontrol.ru/forum/viewtopic.php?f=2&t=10693>)

3.3. Уровни имперфективируемости отперфективных по-глаголов по всем трем источникам оказываются существенно ниже, чем соответствующие показатели для минимальной выборки, см. **рис. 6**.

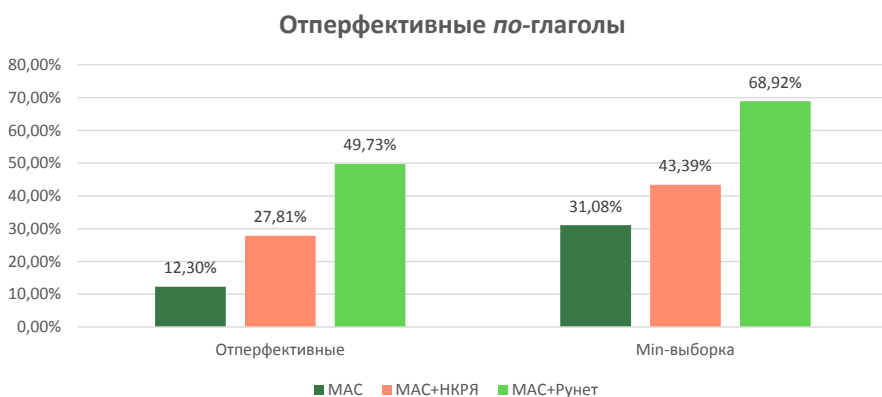


Рис. 6. Имперфективируемость отперфективных по-глаголов в сравнении с по-глаголами мин-выборки

Как правило, уровень имперфективируемости отперфективных глаголов оказывается выше, чем по приставочной группе в целом (см., например, **рис. 7** и **8** для *на-* и *за-* глаголов). Это может объясняться тем, что в отсутствие симплекса (НСВ₁) для глаголов данной группы вторичный имперфектив является единственно возможным коррелятом НСВ [Горбова 2019a].

Причиной сниженного уровня имперфективируемости отперфективных глаголов с префиксом *по-* может быть вхождение в их состав подгруппы глаголов аттенуативного СД (62 из 187–33,36%) и ряда глаголов дистрибутивного СД (61 из 187% — 32,62%), для которых характерен низкий уровень имперфективируемости. Кроме того, поиск имперфективных коррелятов для глаголов аттенуативного и дистрибутивного СД затруднен по следующим формальным причинам: 1) для дистрибутивного СД (*повытащить*) — при попытке подвергнуть глагол имперфективации получается дериват, формально совпадающий с глаголом дистрибутивного СД от имперфективной основы (*повытаскивать*); 2) для аттенуативного СД (*пообсохнуть*) — при попытке подвергнуть глагол имперфективации получается глагол, совпадающий по форме с делимитативным

СД (*пообсыхать*). Имперфективная интерпретация соответствующих дериватов от глаголов аттенуативного СД возможна в редких случаях при поддержке контекста, см. примеры (3)–(4) выше.

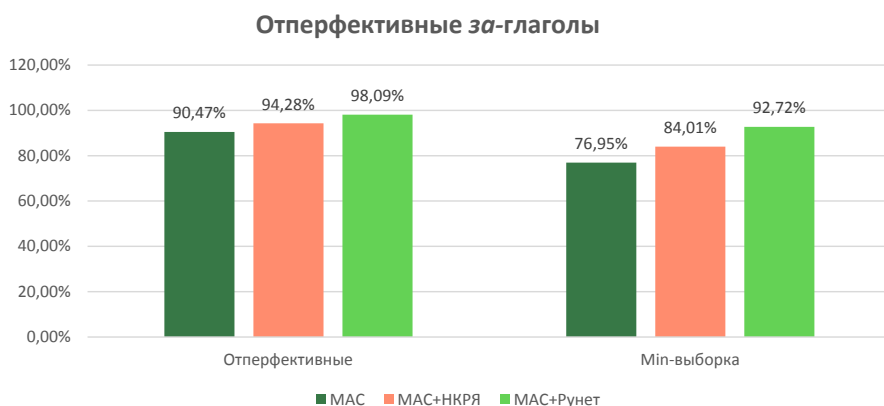


Рис. 7. Имперфективируемость отперфективных за-глаголов в сравнении с за-глаголами min-выборки

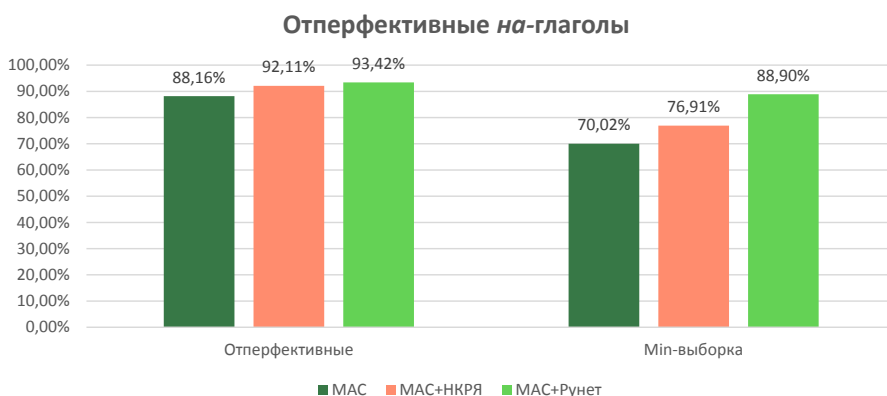


Рис. 8. Имперфективируемость отперфективных на-глаголов в сравнении с на-глаголами min-выборки

4. Выводы

Префиксальную группу *по*-глаголов составляют лексемы, значительно различающиеся по своим свойствам, в частности, в отношении их способности к вторичной имперфективации. При этом ряд особенностей отличает *по*-глаголы от других приставочных групп.

Общий уровень имперфективируемости *по*-глаголов, определяемый по минимальной (и в меньшей степени — по максимальной) выборке, значительно снижен по сравнению с другими префиксальными группами. В наименьшей степени (31,08% по минимальной выборке) НСВ₂ от перфективных *по*-глаголов отражены в МАС. В частности, для *по*-глаголов, определяемых по МАС как «сов. к 'НСВ₁'», НСВ₂ как правило отсутствует. Резкое увеличение количества НСВ₂ в НКРЯ и рунете может объясняться стремлением к регулярному образованию имперфективного коррелята.

В рамках *по*-глаголов делимитативный СД показывает более высокий уровень имперфективации относительно минимальной выборки по сравнению с другими СД как в рамках данной приставочной группы, так и за ее пределами. При этом именно делимитативный СД показывает наибольшую регулярность (32,65%) образования. Можно предположить наличие корреляции между регулярностью образования делимитативов и имперфективируемостью⁶.

Отперфективные *по*-глаголы демонстрируют более низкий уровень имперфективируемости по сравнению с другими рассмотренными приставочными группами, что может объясняться формальными причинами — вхождением в данную специализированную группу глаголов аттенуативного и дистрибутивного СД, для которых вторичная имперфективация в значительной степени затруднена.

Таким образом, можно сделать вывод, что наблюдаемые особенности имперфективации глаголов с префиксом *по*-, отличающие данную префиксальную группу от других перфективных групп, не противоречат гипотезе о регулярности и системном характере вторичной имперфективации, см. [Горбова 2017]; [2019с].

Литература

1. Chertkova M. (1996), The grammatical category of aspect in Modern Russian [Grammaticheskaya kategoriya vida v sovremennom russkom yazyke], MGU Publ., Moscow.
2. Dickey S. M. (2006), Aspectual pairs, goal orientation and *po*-delimitatives in Russian, *Glossos*, vol. 7, pp. 1–32.
3. Evgenieva A. (ed.) (1981–1984), Dictionary of Russian Language in 4 volumes [Slovar russkogo yazyka v 4-kh tomakh], 2nd ed. Academy of Science of USSR, Institute of Russian Language, Moscow, available at: <http://feb-web.ru/feb/mas/mas-abc/14/ma239217.htm>.
4. Fedotov M., Chuikova O. (2013), On the definition of limitative aspectual meaning and on the features of “delimitative” verbs in Russian [Kopredeleniiu aspektual'nogo znacheniiia delimitativa i voprosu ob osobennostiakh “delimitativnoi” derivatsiit russkogo glagola], E. I. Grekhova (ed.) From the past to the future. Collection of articles and memoirs to the 100th anniversary of prof. Yu. S. Maslov [Iz proshlogo v budushchee. Sbornik statej i vospominanij k 100-letiiu Yu. S. Maslova], SPbU Publ., St. Petersburg, pp. 153–203.

⁶ Объяснение было предложено в [Горбова, Чуйкова 2020]. Данная гипотеза объясняет также более высокий уровень имперфективируемости кумулятивных глаголов (28,39% в составе максимальной выборки) на фоне остальных СД.

5. *Gorbova E.* (2017), Aspectual formation of Russian verbs: Inflection, derivation, or a set of quasigramemes? ("sore points" of Russian aspectology revisited) [Russkoe vidoobrazovanie: slovoizmenenie, slovoklassifikaciya ili nabor kvazigrammem? (eshche raz o bolevykh tochkakh russkoj aspektologii)], Topics in the study of language [Voprosy Jazykoznanija], vol. 1, pp. 24–52.
6. *Gorbova E.* (2019a), Regularity of the Secondary Imperfectivation of Russian Prefixal Perfectives (the case of *pro-*, *u-* and *s(o)-* verbs) [Regulyarnost' vtorichnoj imperfektivacii russkikh prefigirovannykh perfektivov (na primere glagolov s prefiksami *pro-*, *u-*, *s(o)*)], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019" [Komp'yuternaya Lingvistika i Intellekual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2019"], issue 18, supplementary vol., pp. 31–44, available at: <http://www.dialog-21.ru/media/4870/-dialog2019scopusvolplus.pdf>
7. *Gorbova E.* (2019b), To the Restriction on Imperfectivation: are Russian Verbs of Perfective Aktionsarten Imperfectivable? [K ogranicheniyu na imperfektivaciyu: imperfektiviruyutsya li russkie glagoly perfektivnykh sposobov dejstviya?], Gerasimov D. V., Dmitrenko S. Yu., Zaika N. M. (eds.) Collection of articles to the 85th anniversary of V. S. Khrakovskij [Sbornik statej k 85-letiyu V. S. Khrakovskogo], Yazyki russkoj kul'tury Publ., Moscow, pp. 98–115.
8. *Gorbova E.* (2019c), Imperfectivability of Russian prefixal perfectives (the case of *pro-* and *u-* verbs) [Imperfektiviruemost' russkikh pristavochnykh perfektivov (na primere *pro-* i *u-* glagolov)], Topics in the study of language [Voprosy Jazykoznanija], vol. 1, pp. 58–74.
9. *Gorbova E., Chuikova O.* (2020), Aktionsarten and the secondary imperfectivation (the case of *po-*, *pro-*, *u-* verbs) [Sposoby dejstviya russkogo glagola i vtorichnaya imperfektivaciya (na primere pristavochnykh grupp glagolov na *po-*, *pro-*, *u-*)], Proceedings of the 7th International Aspectological Conference "Interrelation between Aspect and Adjacent Categories", May 5–8, 2020 [Materialy VII Mezhdunarodnoj aspektologicheskoy konferencii "Vzaimodejstvie aspekta so smezhnymi kategoriyami", 5–8 maya 2020 goda], Herzen State Pedagogical University of Russia Publ., St. Petersburg, pp. 136–148.
10. *Pazel'skaya, A., Tatevosov, S.* (2008), Verbal noun and the structure of Russian verb [Otglagol'noe imia i struktura russkogo glagola], Plungian V., Tatevosov S. (eds.) Verbal derivation research [Issledovaniya po glagol'noi derivatsii]. Moscow: Yazyki slavianskih kul'tur Publ., pp. 348–379.
11. *Russian National Corpus* [Nacional'nyi korpus russkogo jazyka] (2003–2019), available at: <http://www.ruscorpora.ru>.
12. *Shelyakin M.* (2008), Category of aspectuality of the Russian verb [Kategoriya aspektual'nosti russkogo glagola], URSS Publ., Moscow.
13. *Tatevosov S.* (2013), Destructive notes on the distributive Aktionsart [Destrukтивnye zametki o raspredelitel'nom sposobe dejstviya], Russian Language and Linguistic Theory [Russkij yazyk v nauchnom osveshchenii], vol. 2 (26), pp. 30–51.
14. *Ushakov D.* (ed.) (1935–1940), Explanatory Dictionary of the Russian Language in 4 volumes [Tolkovyj slovar' russkogo jazyka v 4-kh tomakh], State Publishing House of Foreign and National Dictionaries, Moscow.

15. *Xrakovskij V. S.* (2018), Aspect's "sore points" revisited once more [I opiat' o bolevykh tochkah kategorii vida], Topics in the study of language [Voprosy Jazykoznanija]. Vol. 1, pp. 105–118.
16. *Zalznjak A., Shmelev A.* (2000), Introduction to the study of Russian aspect [Vvedenie v russkuyu aspektologiyu], Yazyki russkoj kul'tury Publ., Moscow.

Приложение

Таблица 1: Данные об имперфективизируемости *по*-глаголов

	МАС	МАС+ НКРЯ	МАС+ рунет	Всего	Доля в выборке
Исходная выборка	23,64% (421)	40,2% (716)	65,91% (1174)	1781	
Глаголы в исходной выборке:					
Отыменные	87,5% (28)	87,5% (28)	90,63% (29)	32	1,8%
Максимальная выборка	22,47% (393)	39,34% (688)	65,47% (1175)	1749	98,2%
Глаголы в тах-выборке:					
Отперфективные <i>по</i> -перфективы	12,3% (23)	27,81% (52)	49,73% (93)	187	10,69%
Аттенуативный СД (полностью входит в отперфективные)	0% (0)	17,74% (11)	48,39% (30)	62	3,54%
Глаголы на <i>-и(зу)ровать</i> (из них делимитативного СД — 11)	0% (0)	0% (0)	16,67% (2)	12	0,67%
Делимитативный СД	20,49% (117)	43,43% (248)	70,03% (417)	571	32,65%
Дистрибутивный СД (из них отперфективных — 61)	0% (0)	15,48% (26)	36,9% (62)	168	9,61%
Ингрессивный СД	0% (0)	13,04% (9)	37,68% (26)	69	3,95%
Минимальная выборка	31,08% (253)	43,49% (354)	68,92% (561)	814	46,54%

Таблица 2: Данные об имперфективиремости за-глаголов

	МАС	МАС+ НКРЯ	МАС+ рунет	Всего	Доля в выборке
Исходная выборка	58,6% (1264)	65,18% (1406)	80,44% (1735)	2157	
Глаголы в исходной выборке:					
Отыменные	70% (42)	73,33% (46)	93,3% (58)	60	2,78%
Максимальная выборка	58,27% (1222)	64,95% (1362)	80,01% (1679)	2097	97,22%
Глаголы в тах-выборке:					
Отперфективные за-перфективы	90,47% (95)	94,28% (99)	98,09% (103)	105	5,01%
Глаголы на <i>-и(зи)ровать</i> (из них начинательного СД — 6, чрезмерно-дли- тельного СД — 1)	18,18% (12)	19,7% (13)	68,18% (45)	66	3,15%
Начинательный СД	2,29% (11)	8,96% (43)	39,58% (190)	480	22,89%
Чрезмерно-длительный СД	50% (26)	57,69% (30)	84,62% (44)	52	2,48%
Минимальная выборка	76,95% (1078)	84,01% (1177)	92,72% (1299)	1401	66,81%

Таблица 3: Данные об имперфективируемости *на*-глаголов

	МАС	МАС+ НКРЯ	МАС+ рунет	Всего	Доля в выборке
Исходная выборка	62,13 % (914)	68,59 % (1009)	85,11 % (1252)	1471	
Глаголы в исходной выборке:					
Отыменные	88,89 % (8)	88,89 % (8)	88,89 % (8)	9	0,61 %
Максимальна выборка	61,97 % (906)	68,47 % (1001)	85,09 % (1244)	1462	99,39 %
Глаголы в тах-выборке:					
Отперфективные <i>на</i> -перфективы (из них кумулятивного СД — 8)	88,16 % (67)	92,11 % (70)	93,42 % (71)	76	5,20 %
Глаголы на <i>-и(зи)ровать</i> (из них сатуративного СД — 6)	25 % (4)	25 % (4)	56,25 % (9)	16	1,09 %
Кумулятивный СД	61,93 % (257)	67,47 % (280)	83,61 % (347)	415	28,39 %
Сатуративный СД (из них отперфективных — 61)	20,54 % (38)	29,19 % (54)	70,27 % (130)	185	12,65 %
Минимальная выборка	70,02 % (549)	76,91 % (603)	88,9 % (697)	784	53,63 %

A SIMPLE SOLUTION FOR THE TAXONOMY ENRICHMENT TASK: DISCOVERING HYPERNYMS USING NEAREST NEIGHBOR SEARCH

Dale D. S. (dale.david@yandex.ru)

Yandex, Moscow, Russia

In this paper, we present the system we used in the Taxonomy Enrichment for the Russian Language evaluation campaign. The goal of this challenge is to predict hypernyms for the words not included in the taxonomy. Our approach was to generate and score candidate hypernyms by word embedding similarity of the input words and concepts already in the taxonomy. Despite being very simple, our system was ranked first on the verbs track.

Keywords: wordnet, hypernym, taxonomy enrichment, word2vec, nearest neighbor

DOI: 10.28995/2075-7182-2020-19-177-186

ПРОСТОЕ РЕШЕНИЕ ЗАДАЧИ ПО ОБОГАЩЕНИЮ ТАКСОНОМИИ: ВЫЯВЛЕНИЕ ГИПЕРОНИМОВ С ПОМОЩЬЮ ПОИСКА БЛИЖАЙШИХ СОСЕДЕЙ

Дале Д. С. (dale.david@yandex.ru)

Яндекс, Москва, Россия

В этой статье мы представляем систему, использованную нами в соревновании по обогащению таксономии для русского языка. Задача соревнования — предсказать гиперонимы слов, не включённых в таксономию. Для этого мы генерируем и ранжируем гиперонимы-кандидаты по сходству словных эмбедингов входных слов с эмбедингами понятий, уже включённых в таксономию. Несмотря на свою простоту, наша система достигла наилучшей точности на подзадаче поиска гиперонимов для глаголов.

Ключевые слова: wordnet, гиперонимы, обогащение таксономии, word2vec, метод ближайших соседей

1. Introduction

Hypernymy is the name for “is a” relation between words or phrases: e.g. hypernym of “whale” is “sea mammal”, and hypernyms of “sea mammal” are “mammal” and “sea creature”. Thesauri labeled with hypernymy relation can be used to solve tasks such as resolution of lexical ambiguity [11], query expansion in information retrieval [6], [11], processing questions and answers in question answering systems [6], [11], sentiment analysis and semantic similarity measurement [9], etc. One of such databases, WordNet [14] for the English language, has been in use for more than 20 years and remains a valuable source for various applications [9]. However, manually producing hypernyms for new words is time-consuming and expensive [9]. Therefore automatic discovery of hypernyms is an important problem [3], [4], [6], [9].

The evaluation campaign “Taxonomy Enrichment for the Russian Language” organized by the international conference “Dialogue 2020”¹ [15] in which we take part is aimed exactly at this problem. Its goal is to provide 10 ranked candidate hypernyms for each new word in the test set. Hypernyms should be chosen from the existing RuWordNet taxonomy [12]. The challenge consists of two separate tracks for nouns and for verbs.

We approach the problem of hypernymy discovery by exploiting the existing structure of RuWordNet. This thesaurus contains 85K (33K) terms grouped into 30K (7K) synsets for nouns (verbs), and we expect² that most new words have siblings (i.e. terms with the same hypernyms) in RuWordNet. The siblings should be semantically close to each other, so we expect that their word embeddings are also similar. Therefore, we use a weighted K-nearest-neighbor algorithm over word embeddings to retrieve potential siblings and rank their hypernyms as potential hypernyms of the query term.

This simple algorithm turned out to be unexpectedly effective, and we managed to achieve the best score for verbs track with it. In this paper, we describe it in more detail and analyze what makes our approach successful.

2. Related Work

Two important approaches to hypernymy discovery are pattern-based and distributional [3]. The pattern-based approach pioneered by Hearst [8] predicts hypernymy between words if they often co-occur in patterns like “A, such as B”. The distributional approaches make use of distributional representations of terms, such as word embeddings [7]. Another important line of work utilizes definitions of terms, instead of unstructured corpora, to propose hypernyms for the terms [9].

Biemann et al. [4] give a good overview of existing approaches for enriching lexical semantic resources with distributional data. They also provide their own system for building taxonomies based on graphs of semantically related words induced from corpora.

¹ <http://www.dialog-21.ru/evaluation/>

² It turns out to be true; see the subsection “Siblings”.

Despite the importance of the hypernymy discovery problem, the challenge “Taxonomy Enrichment for the Russian Language” [15] seems to be the first campaign for Russian or any other Slavic language that evaluates discovery of hypernyms for new terms. However, there were similar competitions for English and other European languages, most notably SemEval-2016 Task 14 [9] (enriching a taxonomy using the definitions of the new words) and SemEval-2018 Task 9 [6] (extracting hypernyms from unlabeled corpora).

Best solutions of SemEval-2018 Task 9 include CRIM [2] (pattern-based discovery and scoring query-hypernyms pairs with a neural net), 300-sparsans [1] (sparse features and formal concept analysis). In SemEval-2016 Task 14, the winning system was MSerjKu [16] (classification of query-hypernym pairs using SVM with distributional and linguistic features).

3. Task description

3.1. Goal and metrics

The task is formulated as follows: for each term (query) in the test set, one should provide a list of 10 candidate hypernyms. They are evaluated against the ground truth: human-labeled hypernyms and hypernyms of these hypernyms. All these first- and second-order hypernyms are divided into connected components, and ranking scores are evaluated relatively to these components. The scores include mean average precision³ (MAP) at the true number of hypernym components, mean reciprocal rank⁴ (MRR) at 10, and F1 score (at the top 1 prediction); the official metric is MAP. The formulas for calculation MRR and MAP were customized to treat the whole connected component of hypernyms as a single hypernym. They are available in the official repository of the competition⁵.

The task includes two separate tracks for nouns and verbs.

3.2. Provided datasets

The main piece of the provided data is the RuWordNet taxonomy, and train/test sets based on it. Some more additional datasets were proposed, but we did not use them.

RuWordNet is a collection of synsets (sets of terms with the same meaning) and relations (such as hypernymy/hyponymy) between them. Each synset consists of the terms (which may be single- or multi-word expressions), the title, and (optionally) the definition, and has a unique identifier.

³ [https://en.wikipedia.org/wiki/Evaluation_measures_\(information_retrieval\)#Mean_average_precision](https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#Mean_average_precision)

⁴ https://en.wikipedia.org/wiki/Mean_reciprocal_rank

⁵ Formulas for both MAP and MRR are in the file `evaluate.py` in <https://github.com/dialogue-evaluation/taxonomy-enrichment>.

The train set includes 25K nouns and 7K verbs with their first- and second-order hypernyms grouped into connected components. The public test set includes 762 nouns and 175 verbs, and the private test set—1525 nouns and 350 verbs without any labels.

3.3. Data split

All our experiments were evaluated on the dev set (randomly selected 2% of the training set), and on the public test set. There are two reasons, why these scores may be mismatched. The first reason is that nearly 74% nouns and 80% verbs from the training set have at least one sense in the existing RuWordNet taxonomy, whereas the test set has no intersection with the taxonomy. And some of these intersections are inconsistent: for example, **ОТКРОВЕННОСТЬ** (openness) has a sense **ЧИСТОСЕРДЕЧНЫЙ** (sincere) in the taxonomy but does not have corresponding hypernyms in the training set. To restrict the influence of this discrepancy, we exclude from the taxonomy all the terms in the dev set and their direct synonyms when we evaluate on the dev set. Second, 70% of nouns and 60% of verbs in the training set are in fact multi-word phrases such as **МУСКАТНОЕ ВИНО** (muscat wine), whereas the test sets consist only of sole words.

3.4. Siblings

Our chief hypothesis that most new words have siblings in RuWordNet proved itself true. In the training dataset, 90% of nouns and 99% of verbs have siblings. Moreover, 99.98% of training nouns and 100% of training verbs have either siblings or “cousins” (terms with common second-order hypernyms). Words without siblings include some neologisms (e.g. **ПОЛИТТЕХНОЛОГ** is the only hyponym of **ИМИДЖМЕЙКЕР**) and rare toponyms (e.g. **ЮГРА** is the only hyponym of **АВТОНОМНЫЙ ОКРУТ РФ**).

4. System description

The proposed method of predicting hypernyms is based on cosine similarity between fixed (context-independent) term embeddings. For each new term, we find its k nearest neighbors among the terms in the taxonomy and use their first- and second-order hypernyms as candidate hypernyms.

4.1. Index construction

We construct the pool of potential neighbors by taking for each RuWordNet synset its title, all its senses, and a concatenation of its title and its senses. For each of these texts, we calculate its text embedding as a weighted mean embedding of all words in it. The weights of the words in our implementation depend only on POS tags, but in a more complex setting, they could be tied e.g. to the syntactic role of the word. We L2-normalize word embeddings before aggregation in order to make representation of words more comparable to each other. We also L2-normalize sentence

embeddings after aggregation in order to make Euclidean distance between them equivalent to cosine distance and simplify neighbor search.

To extract word embeddings, we use a word2vec [13] model pretrained on the Taiga corpus [17] and published on RusVectors [10].⁶ Before lookup, we lemmatize each word and append the POS label to it. If the word is missing in the vocabulary of this model, we find all words in the vocabulary with the longest prefix matching this word and compute its embedding as the mean of their embeddings. For example, the embedding of the word **перуанка** _ NOUN (a female Peruvian) is computed as the mean of embeddings of **перуанец** _ NOUN (a male Peruvian) and **перуанский** _ ADJ (Peruvian).

As an alternative way to extract word embeddings, we use a fastText [5] model, which was pretrained on Taiga and published on RusVectors as well. It differs from the model above in two ways: it does not include POS tags, and it constructs embeddings for unseen words by averaging the embeddings of their character n-grams.

4.2. Ranking candidates

For each query term, we find its k nearest neighbors in the index (using the embeddings described above), and use all the first- and second-order hypernyms of the neighbors' synsets as answer candidates. We score occurrences of hypernyms with each particular neighbor separately and add together such scores for each hypernym candidate. The resulting prediction is the 10 candidate hypernyms with the highest total scores.

The score for each hypernym associated with a particular neighbor is calculated as

$$score = \exp(-d^\alpha) \times s^\beta \times \begin{cases} 1, & \text{for first-order hypernyms of the neighbor} \\ \gamma, & \text{for second-order hypernyms} \end{cases}$$

where s is cosine similarity between the query and the neighbor, and $d = \sqrt{2(1-s)}$ is the distance between them. The constant γ reflects the preference between first- and second-order hypernyms. This formula was constructed manually and performed no worse than our attempt to train linear scoring formulas on the training datasets. In fact, the functions $\exp(-d^\alpha)$ and s^β have similar shapes, and only one of them would suffice, but we kept both to make the formula more flexible (and as a legacy of our experiments).

In general, with this formula we try to combine the evidence from the few close neighbors with the evidence from numerous distant neighbors. The parameters α and β are tuned in order to balance these signals. High values of α and β decrease the impact of the neighbors which are far from the query, allowing to use higher values of k , i.e. evidence from more neighbors.

5. Experiments and results

After some preliminary experiments, we chose and submitted the solution with $k = 100$, $\alpha = 3$, $\beta = 5$, and $\gamma = 0.5$. When calculating text embeddings for neighbor

⁶ The model can be downloaded from <https://rusvectors.org/ru/models/>.

search, we weighted words according to POS: 1.0 for the target POS (noun and verb respectively), 0.1 for prepositions, 0.5 for other POS. But for calculating scores (i.e. s) we used uniform word weights.

Our algorithm turned out to be inefficient on nouns, with the submitted version scoring only 41.78% MAP on the private test set⁷. This is a little below the fastText baseline provided by the competition team near the deadline date. However, on verbs, our approach was more efficient and scored 44.83% MAP on the private test set, which is the best result so far.

5.1. Ablation study

In this section, we analyze the importance of different design decisions we made. The preliminary experiments were not well structured, so instead, we do an ablation study and show the effect of modifying some of our decisions. We evaluate MAP for nouns and verbs on our dev set and on the public test set. We do not report the MRR score, but its behavior is qualitatively similar to that of MAP. The results are summarized in **Table 1**.

From the table, we see that the model that we submitted performed worse than the baseline and the models of other participants on the public test set of nouns, but much better than the baseline and better than the competitors on the public test set of verbs. These results are consistent with the private test set.

We also see that some of the modifications to our model improve the MAP on a few datasets, but none of them improve the scores consistently on all the datasets.

Table 1: MAP of modified versions of the model

Model	nouns dev	nouns test	verbs dev	verbs test
Our submitted model	.4695	.4083	.2527	.4033
The best model of the competitors	—	.5590	—	.4032
The FastText baseline	—	.4343	—	.2760
$k = 30$.4570	.3871	.2407	.3937
$k = 300$.4561	.3983	.2664	.3884
$\alpha = 1$.4699	.4084	.2573	.3987
$\alpha = 0$.4216	.4093	.2587	.3909
$\beta = 1$.4415	.4083	.2514	.4023
$\beta = 0$.4216	.3639	.2466	.3799
$\gamma = 1$.4396	.3963	.2429	.3677
$\gamma = 0$.4753	.3857	.2587	.4016
FastText embeddings	.4263	.2432	.2237	.2615
s without POS weights	.4660	.4065	.2585	.4077
KNN with POS weights	.4653	.4071	.2338	.3900
Reduced index	.4627	.4121	.2671	.3645

⁷ The leaderboard is available at <https://competitions.codalab.org/competitions/22168#results>

All three parts of the ranking formula turned out to be useful: setting α or β to 0 or γ to 1 (effectively disabling parts of the formula) made the MAP scores deteriorate. When we changed the POS weighting scheme, MAP decreased in most cases as well. Replacing word2vec embeddings with FastText embeddings trained on the same Taiga corpus led to dramatically deteriorating performance.

One more subtle distinction of the proposed algorithm from the baseline is that it uses different terms of the synset separately in the search index. To validate this decision, we created an alternative index, when all entries in a synset are concatenated together before calculating an embedding and including it in the KNN index. This modification led to a visible increase in the test score for nouns, but the score for verbs dropped dramatically, so we decided not to submit this version.

6. Analysis

In this section, we analyze why our rather naive approach for hypernym discovery works and what it lacks.

6.1. Collecting vs ranking candidates

We start by comparing the impact of the quality of collecting and ranking candidate hypernyms on the overall quality. For this purpose, we estimate MAP with an oracle ranker on the dev set and get 81.6% (vs 47%) on nouns and 70.7% (vs 25%) on verbs. In more intuitive terms, this corresponds to 79% and 70% recall for nouns and verbs, respectively. It might mean that poor ranking is more responsible for the low score than poor candidate collection because with perfect ranking the gap between our solution and ground truth decreases by more than half. Our hypothesis is reinforced by the fact that our competitors' system that got the highest score on nouns was using features from numerous data sources for ranking.

6.2. Error analysis

Table 2: Relative frequency of model errors

Error type	Nouns	Verbs
No errors	.44	.12
Domain heuristics	.20	.40
Too general predictions	.08	.14
Homonymy	.05	.17
Abstract concept	.11	.01
Compositionality	.04	.09
Domain knowledge	.08	.00
Inversion of valence	—	.04
Labeling error	.00	.03

To further understand the upsides and downsides of our system, we manually inspect 200 samples from the dev set and label the errors of our system on them. The frequencies of these errors are given in [Table 2](#). If our system made several errors on a sample, we assume their equal contribution.

The major causes of errors include:

- *domain heuristics*: extracted neighbors are semantically related to the query (from the same domain), but reflect a different concept. For example, some of the close neighbors for **ЗАРЯЖАНИЕ** (loading or charging) are **ПРИЦЕЛИВАНИЕ** (aiming) and **ЛАФЕТ** (gun carriage), because they often occur together in the context of guns.
- *homonymy*: the term conveys multiple meanings, and golden homonyms are provided for one meaning, but neighbors — for another. For example, for **ВЫГОРАНИЕ** (fading, burnout) the golden hypernym is **ОБЕСЦВЕТИТЬСЯ** (to lose color), but our system provided **ГОРЕТЬ** (to burn).
- *too general predictions*: predicted hypernyms are more abstract than needed. For example, for **ПЕРЕСТАВЛЯТЬ** (to change places) the golden hypernyms include **СТАВИТЬ** (to set place), but our system predicted a more general **ПЕРЕМЕСТИТЬ** (to move).
- *abstract concepts*: the queries and the golden hypernyms are quite abstract (this is especially true for properties or processes), but our system interprets only one specific context of their usage, and does it wrong. For example, the word **ПОРТАТИВНОСТЬ** (portability) is out-of-vocabulary, so our system makes an inference about it from the word **ПОРТАТИВНЫЙ** (portable), which is distributionally close to gadgets, although semantically it is more general. As a result, the model predicts false hypernyms, such as **ЭЛЕКТРОННОЕ ОБОРУДОВАНИЕ** (electronic equipment).
- *compositionality*: mean word embeddings poorly reflect the meaning of a multi-word term, because they are not aware of the syntax. For example, for the term **ПРОГРЕВАНИЕ БОЛЬНОГО МЕСТА** (warming up a sore spot) some of the predicted candidates are **БОЛЬНОЙ ЧЕЛОВЕК** (ill person) and **МЕСТО В ПРОСТРАНСТВЕ** (place).
- *domain knowledge*: predicting a correct hypernym requires specific knowledge about the world which is difficult to extract from distributional semantics. For example, one of the hypernyms for **ИРЛАНДСКАЯ СТОЛИЦА** (Irish capital) is **ГРАФСТВО** (county), but our model seems to be unaware of it.
- *inversion of valence*: the model mixes up the verbs describing the same process for the subject and the object. For example, for **ЗАТЮКАТЬ** (to harass) the model wrongly predicts a hypernym **ПЕРЕЖИТЬ** (to experience).
- *labeling errors*: the golden hypernyms do not correspond to the generally accepted meanings of the query term. For example, for **НАБИВАТЬСЯ** (to be stuffed, or to foist) there is a golden hypernym **НАДОЕСТЬ** (to pester), which is semantically related to the second sense of the query but seems to be its sibling, not a hypernym.

6.3. Areas of improvement

The analysis above indicates that the limitations of our method are mostly due to the limitations of static word embeddings themselves: they do not utilize morphology and syntax of the queries and provide only a narrow way of understanding their semantics. A better model would take into account:

- word morphology — to better extrapolate meaning between related words;
- phrase structure — to correctly integrate the meanings of the head and the dependent words into the phrase embedding, and to resolve homonymy;
- definitions of terms from external sources — to reason more correctly about the meaning of the rare words, and to use domain-specific relations;
- the structure of the taxonomy itself — to filter out too general or too specific hypernyms.

7. Conclusion

In this paper, we introduce a simple baseline for hypernym prediction, based solely on fixed word embeddings and their similarity. Its distinctive features are a large number of retrieved neighbors, nonlinear distance-based candidate scoring, and heuristics for obtaining phrase embeddings. Despite its simplicity, our system got the best score for the verbs track, which may indicate that nobody knows a smart way of predicting hypernyms for Russian verbs. Further research could integrate our system with other techniques of hypernymy prediction, which is necessary to overcome the limitations of static word embeddings.

References

1. *Berend, G. et al.*: 300-sparsans at SemEval-2018 task 9: Hypernymy as interaction of sparse attributes. In: Proceedings of the 12th international workshop on semantic evaluation. pp. 928–934 Association for Computational Linguistics, New Orleans, Louisiana (2018).
2. *Bernier-Colborne, G., Barrière, C.*: CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In: Proceedings of the 12th international workshop on semantic evaluation. pp. 725–731 Association for Computational Linguistics, New Orleans, Louisiana (2018).
3. *Biemann, C.*: Ontology learning from text: A survey of methods. In: LDV forum. pp. 75–93 (2005).
4. *Biemann, C. et al.*: A framework for enriching lexical semantic resources with distributional semantics. *Natural Language Engineering*. 24, 265–312 (2017).
5. *Bojanowski, P. et al.*: Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*. 5, 135–146 (2017).
6. *Camacho-Collados, J. et al.*: SemEval-2018 task 9: Hypernym discovery. In: Proceedings of the 12th international workshop on semantic evaluation. pp. 712–724 Association for Computational Linguistics, New Orleans, Louisiana (2018).

7. *Espinosa-Anke, L. et al.*: Supervised distributional hypernym discovery via domain adaptation. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 424–435 Association for Computational Linguistics, Austin, Texas (2016).
8. *Hearst, M. A.*: Automatic acquisition of hyponyms from large text corpora. In: COLING 1992 volume 2: The 15th International Conference on Computational Linguistics. (1992).
9. *Jurgens, D., Pilehvar, M. T.*: SemEval-2016 task 14: Semantic taxonomy enrichment. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 1092–1102 Association for Computational Linguistics, San Diego, California (2016).
10. *Kutuzov, A., Kuzmenko, E.*: WebVectors: A toolkit for building web interfaces for vector semantic models. In: Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers. pp. 155–161 Springer International Publishing, Cham (2017).
11. *Loukachevitch, N.*: Thesauri in problems of information retrieval [tezaurusy v zadachah informatsionnogo poiska], (2010).
12. *Loukachevitch, N. V. et al.*: Creating Russian WordNet by Conversion. In: Proceedings of Conference on Computational Linguistics and Intellectual Technologies Dialog-2016. pp. 405–415 (2016).
13. *Mikolov, T. et al.*: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013).
14. *Miller, G. A.*: WordNet: An electronic lexical database. MIT press (1998).
15. *Nikishina, I. et al.*: RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian Language. In: Computational linguistics and intellectual technologies: Papers from the annual conference “dialogue”. (2020).
16. *Schlichtkrull, M., Martínez Alonso, H.*: MSejrKu at SemEval-2016 task 14: Taxonomy enrichment by evidence ranking. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 1337–1341 Association for Computational Linguistics, San Diego, California (2016).
17. *Shavrina, T., Shapovalova, O.*: TO the methodology of corpus construction for machine learning: “TAIGA” syntax tree corpus and parser. In: Proceedings of the international conference Corpus linguistics–2017. St. Petersburg. pp. 78–84 (2017).

RENERSANs: RELATION EXTRACTION AND NAMED ENTITY RECOGNITION AS SEQUENCE ANNOTATION

Davletov A. A. (davletov-aa@ranepa.ru)^{1,3},

Gordeev D. I. (gordeev-di@ranepa.ru)¹,

Rey A. I. (rey-ai@ranepa.ru)¹,

Arefyev N. V. (nick.arefyev@gmail.com)^{2,3,4}

¹RANEPA, Moscow, Russia

²Samsung R&D Institute Russia, Moscow, Russia

³Lomonosov Moscow State University, Moscow, Russia

⁴National Research University Higher School of Economics
Moscow, Russian Federation

In this work we present our system for RuREBus shared task held together with Dialog 2020 conference. The task consisted of 3 subtasks: named entity recognition, relation extraction with provided named entity tags and end-to-end relation extraction. Our system took the first and the second place in the first and the second subtasks respectively. For the third subtask we submitted our solution only in the post-evaluation phase, however, it was among the top 2 best performing systems. The systems for all tasks are based on Transformer models. Relation extraction was solved as a sequence labelling problem. We also used joint task named entity and relation extraction learning.¹

Key words: Relation Extraction, Named Entity Recognition, Transformer, BERT

DOI: 10.28995/2075-7182-2020-19-187-197

¹ <https://github.com/AdisDavletov/DeftEval2020/tree/dev>

РЕНЕРСАНС: ИЗВЛЕЧЕНИЕ ОТНОШЕНИЙ И АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ КАК РАЗМЕТКА ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Давлетов А. А. (davletov-aa@ranepa.ru)^{1,3},
Гордеев Д. И. (gordeev-di@ranepa.ru)¹,
Рей А. И. (rey-ai@ranepa.ru)¹,
Арефьев Н. В. (nick.arefyev@gmail.com)^{2,3,4}

¹РАНХиГС, Москва, Россия

²Московский Исследовательский Центр Самсунг

³Московский Государственный

Университет им. М. В. Ломоносова

⁴Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

В данной работе мы представляем нашу систему для соревнования RuREBus, проводящегося совместно с конференцией Dialog 2020. Задача состояла из 3 дорожек: распознавание именованных сущностей, классификация отношений между заранее аннотированными именованными сущностями и извлечение отношений из неаннотированного текста. Наша система заняла первое место на первой дорожке и второе место на второй. Для третьей задачи мы не успели своевременно представить решение, но оно бы оказалось в числе лучших систем. Системы для всех задач основаны на моделях Transformer. Извлечение отношений мы рассматривали как задачу разметки последовательностей. Также мы использовали совместное обучение для задач распознавания именованных сущностей и извлечения отношений.

Ключевые слова: извлечение отношений, распознавание именованных сущностей, Transformer, BERT

1. Introduction

This work is devoted to our solution for RuREBus [4] shared task held together with the conference Dialog 2020. RuREBus shared task was devoted to the problem of relation extraction and named entity recognition (NER) in a specialized business domain. It consisted of three subtasks: named entity recognition, relation extraction with provided named entity labels and end-to-end relation extraction. Our first subtask solution was a BERT-based [2] sequence labelling model. For the second one we applied joint named entity and relation extraction learning. We went with a similar approach for the third subtask. However, due to having no labelled named entities, they were inferred using the model trained for the first subtask.

Our NER model with the 0.561 F1-score at the test dataset took the first place in the shared task. Our second subtask model took the second place with the F1-score equal to 0.394.

Our work shows that the sequence labelling approach is viable for relation extraction. It also demonstrates that correct named entity labels are vital for relation extraction due to the difference in scores between the second and the third subtask models.

2. Related work

There are many ways to extract information from text. This task is often solved by extracting named entities and classifying relations between them. One of the most popular datasets for this task is TACRED [10] where semantic relations are understood as relations between two pairs of entities.

Nowadays, state-of-the-art results for this dataset are achieved with Transformer-based models [8]. The most advanced models (according to paperswithcode²) use extra training data or additional knowledge bases. For example, in the state-of-the-art system the authors use Wikipedia data [1]. However, such data is useless for domain-specific relations.

Among the systems that do not use encyclopedias or other labeled data, the best results were achieved by Joshi et al. [5]. They pre-trained a BERT-like system, but instead of predicting individual masked tokens they trained the model to infer contiguous random spans. The model was also trained to predict each token in the masked span using output representations of only span boundary tokens. This significantly improved results of their model in comparison with the vanilla BERT. As in both described works we also incorporated information about named entity spans.

However, it is difficult to compare results for relation extraction systems for languages besides English (including Russian) because such annotated datasets are scarce for them. Some researchers have tried to solve this problem using unsupervised language-agnostic approaches and relying on knowledge databases such as Wikidata and various online encyclopedias such as Wikipedia [3]. Models trained this way tend to be not specialized because the original database does not contain relations from the required domain. The results are good only for the most popular relation types such as geographical or professional ones, which frequently appear in Wikipedia.

3. Shared task overview

The organizers of the shared task have provided approximately 300 annotated texts in total. All texts were provided by the Ministry of Economic Development of the Russian Federation. The corpus consists of various regional and strategic plan reports. There are in total 8 named entity classes and 11 semantic relation classes (see **Tables 1** and **2**). The organizers have also provided a large unannotated dataset for

² <https://paperswithcode.com/sota/relation-extraction-on-tacred>

language model fine-tuning. However, we did not use it. A named entity can consist of several words. All entities and relations do not span across sentences. There may be many-to-many, many-to-one and other types of relations (see Fig. 1).

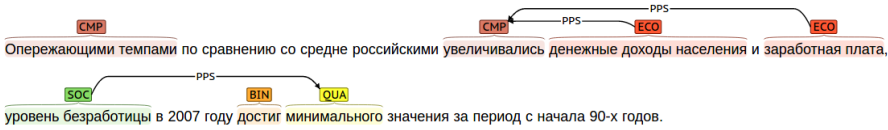


Figure 1: RuREBus annotation example

Named entity groups could contain rather broad types of entities, for example “SOC” entities contained social groups as well as various social attributes—phrases like ‘blue collar workers’ and ‘housing accessibility’ corresponded to this group.

Table 1: Named entity types

Type	Description	Examples
MET	Some quantitative metric	доля сельского населения (rural population ratio); положение в округе (ranking in the neighbourhood)
ECO	An economy entity or facility	обрабатывающим сектором промышленности (processing industry); экономического кризиса (economic crisis)
BIN	A binary attribute	входит в состав (is part of)
CMP	Comparative attribute	рост (growth); увеличился (increased); в наибольшей степени (to the greatest extent)
QUA	Qualitative attribute	лидирующее (leading)
ACT	Activity, actions, implemented policies	восстановление экономики региона region economy reconstruction
INST	Institutions and organizations	Алтайского края (Altai region); Сибири (Siberia)
SOC	Social groups and characteristics	населения края (region population); здравоохранение (health care)

Table 2: Semantic relation types

Group	Type	Description
Current state of affairs	NNG	now negative
Current state of affairs	NNT	now neutral
Current state of affairs	NPS	now positive
Results	PNG	past negative
Results	PNT	past neutral
Results	PNS	past positive
Forecasts	FNG	future negative

Group	Type	Description
Forecasts	FNT	future neutral
Forecasts	FNS	future positive
Goals	GOL	some abstract goals
Tasks	TSK	tasks and actions performed to achieve goals

The organizers first held tracks 1 and 3 and after that track 2 was also run. We describe our solutions in the same order (first tracks 1 and 3, then track 2).

4. Named entity recognition and relation extraction as sequence labelling

The data for the shared task was presented in brat format [7] where texts were given as plain text files and annotations were provided in another file with mixed labels for named entities and relations between them. Thus, we first had to separate the labels and transform the data into special formats used by our models.

We used Razdel library to split plain texts into sentences and tokens.³ It is a rule-based system that along with splitting sentences can also provide sentence and token offsets in the source text. Offset ranges provided by Razdel were used during preprocessing and postprocessing to map tags and relations to text spans which are required by the brat format (see Table 3). We had some conversion problems and the number of NER tags in the brat format did not correspond to the number of tags after processing, for the training dataset the difference was minor, but for the test dataset almost 1% of named entity tags were lost during the preprocessing stage (see Table 3).

Table 3: Named entity types

Dataset	Number of			
	Sentences	Tokens	processed NER tags	Original NER tags (brat)
train	10,460	336,023	54,377	54,388
test	20,483	643,668	89,006	89,879

4.1. Subtask 1: Named Entity Recognition

The first task was to annotate named entities. First we transformed the data word-wise into the BIO-format (beginning, inside, outside). We randomly split the data into training and validation datasets in 0.7 to 0.3 ratio. The split was performed text-wise. After hyperparameter tuning we did not retrain the model using both training and validation data. We used a BERT-based system [2] with PyTorch model code and pretrained weights provided by Hugging Face [9]. Due to the shared task datasets being in Russian, we used the multilingual uncased base BERT model.

³ <https://github.com/natasha/razdel>

BERT is a Transformer based model [8]. On top of BERT outputs we added a linear layer and dropout regularization. The cross entropy loss function was used to train the model. BERT outputs an embedding for each token, i.e. a word may include several BPE-tokens. As we had one label for a word we needed to decide how to aggregate predictions from word tokens. We went with the easiest approach and for each word in the sentence we took a BERT embedding only from its first BPE-token and fed it to the dropout layer followed by the linear layer. All non entity tokens were ignored (i.e. padding tokens and tokens with O tag).

Table 4: Subtask 1: Results on test and development sets on single system

Model	Learning rate	Weight Decay	Dropout	test
bert	1e-5	0.1	0.2	0.548
bert	1e-5	1.0	0.2	0.547
bert	5e-5	0.01	0.1	0.552
bert	5e-5	0.01	0.2	0.555
bert	5e-5	0.1	0.1	0.555
bert	5e-5	0.1	0.2	0.561
bert	5e-5	1.0	0.1	0.554
bert	5e-5	1.0	0.2	0.56

Our system with 0.561 micro F1-score on the public leaderboard outperformed solutions presented by other contestants. In [Table 4](#) we provide results for the test dataset that was provided by the organizers after the shared task.

4.2. Subtask 3: End-to-end Relation Extraction

The second and the third subtasks were relation classification. In the second subtask the organizers provided named entity tags while in the third they did not. For both tracks we used the equivalent approach.

Akin to BERT-multitask learning, in this shared task we wanted to experiment with simultaneous finetuning for separate tracks. RuREBus shared task provided an excellent framework for this idea because we had separate tracks with different target values but the same input data. Thus, we tried a multitask architecture to jointly predict tags and relations. To do so, we consider relation extraction as a sequence labeling problem (similar to how named entity recognition is usually solved). In each example we have one marked main entity and we predict all named entity tags and all relations between the main token and all other tokens in the sentence (see [Fig. 2](#)). We put an empty relation label ("0") if a token does not have relation to the marked entity and the relation tag otherwise. Special tokens marking the beginning and the end of the main entity are added to input to tell the system which entity it should predict relations with. Thus, for each sentence we had to make n predictions where n is the number of named entities in the sentence. We did not relabel previously inferred named entity tags with new predictions.

Sequence labelling might be a preferable solution if we are interested in processing the whole sentence in a smaller number of batches and our algorithm runtime does not depend on the sequence length (unlike recurrent neural networks).

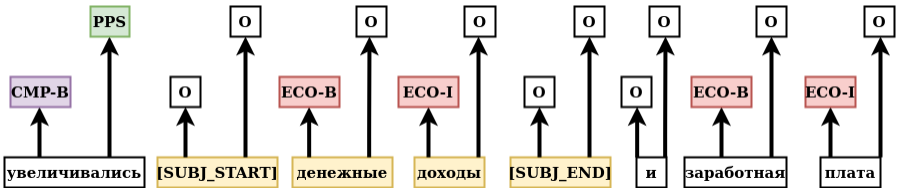


Figure 2: Joint relation extraction and named entity recognition training

Table 5: Subtask 3: F1 scores on test set. Models marked with * were trained after the competition

model	Learning rate	Weight decay	Dropout	NER loss weight	Relation loss weight	test
bert	1e-5	0.1	0.1	0	1.0	0.121
bert	1e-5	0.1	0.1	0.1	1.0	0.127
*xlm-r	1e-5	0.1	0.1	0	1.0	0.194
*xlm-r	1e-5	0.1	0.1	0.01	1.0	0.187
*xlm-r	1e-5	0.1	0.1	0.05	1.0	0.193
*xlm-r	1e-5	0.1	0.1	0.1	1.0	0.182
*xlm-r	1e-5	0.1	0.1	0.2	1.0	0.157
*xlm-r	1e-5	0.1	0.1	1.0	0	0.002
*xlm-r	1e-5	0.1	0.1	1.0	1.0	0.128
*xlm-r	1e-5	0.1	0.2	0.1	1.0	0.169
*xlm-r	1e-5	0.2	0.1	0.1	1.0	0.173
*xlm-r	2e-5	0.1	0.1	0	1.0	0.188
*xlm-r	5e-5	0.1	0.2	0	1.0	0.171
*xlm-r	5e-5	0.2	0.2	0	1.0	0.113

For end-to-end relation extraction we went with a two-stage approach. At first we used the model from the first track to label named entities. After that using the provided named entity predictions we trained our model to infer semantic relations.

In this task we used the same multilingual uncased BERT model as in subtask 1. However, to get simultaneous relation and named entity predictions on top of the model we added another dropout layer followed by tag and relation linear layers. We use weighted sum of cross entropy losses for tag and relation labeling as our final loss for optimization. Padding tokens do not contribute to our loss calculation.

The system showed 0.132 micro F1-score using public test data and it would have taken the first place among the provided systems, if we had managed to submit our solution before the deadline. Joint task learning has slightly improved our results. After the shared task end we also tried RoBERTa base model instead of BERT. It improved our results but the model did not benefit from joint task learning (see [Table 5](#)). Our local evaluation results using the test dataset are slightly worse than the results at the public leaderboard.

4.3. Subtask 2: Relation Extraction for given Named Entities

The model for this track is equivalent to the system used for end-to-end relation extraction. This track was very similar to end-to-end relation extraction. However, instead of using named entity labels predicted by our model, we could use the manual annotation provided by the organizers of the shared task.

For subtask 2 we also tried a base XLM-RoBERTa [6] model also provided by Hugging Face. RoBERTa is BERT inspired model which optimized many hyper-parameter choices in the underlying model. RoBERTa authors have replaced static masking with random masking during language training. They also removed additional sentence prediction loss, increased the batch size, trained on longer sequences and enhanced the original Wikipedia dataset with various Common Crawl datasets. All these adjustments helped RoBERTa to outperform BERT in many benchmarks such as GLUE or SQuAD 2.0.

Table 6: Subtask 2: Results on test and development sets for relation extraction and additional named entity subtask

Model	Learning rate	Weight decay	Dropout	NER loss weight	Relation loss weight	Relation extraction F1		NER F1	
						test	dev	test	dev
bert	1e-5	0.1	0.1	0	1.0	0.25	0.784	0.002	0.001
bert	1e-5	0.1	0.1	0.1	1.0	0.263	0.757	0.189	0.172
xlm-r	1e-5	0.1	0.1	0	1.0	0.391	0.678	0.040	0.049
xlm-r	1e-5	0.1	0.1	0.01	1.0	0.381	0.677	0.330	0.294
xlm-r	1e-5	0.1	0.1	0.05	1.0	0.39	0.685	0.482	0.440
xlm-r	1e-5	0.1	0.1	0.1	1.0	0.379	0.667	0.503	0.468
xlm-r	1e-5	0.1	0.1	0.2	1.0	0.34	0.662	0.501	0.492
xlm-r	1e-5	0.1	0.1	1.0	0	0.006	0.022	0.465	0.463
xlm-r	1e-5	0.1	0.1	1.0	1.0	0.271	0.679	0.489	0.456
xlm-r	1e-5	0.1	0.2	0.1	1.0	0.355	0.668	0.497	0.465
xlm-r	1e-5	0.2	0.1	0.1	1.0	0.357	0.668	0.004	0.001
xlm-r	2e-5	0.1	0.1	0	1.0	0.394	0.675	0.004	0.001
xlm-r	5e-5	0.1	0.2	0	1.0	0.369	0.65	0.059	0.067
xlm-r	5e-5	0.2	0.2	0	1.0	0.272	0.599	0.010	0.0187

In this task our best model with the F1 score equal to 0.394 took the second place.

RoBERTa-based models outperformed BERT-based ones. As we did not include named entity type information in the input, but only spans, we also attempted at using the multi-task learning procedure described in the previous section. However, unlike the previous case the quality deteriorated when the model was trained to predict named entity tags. Thus, the loss coefficient for named entity recognition was set to zero in the final model. Learning rate, weight decay and other hyperparameters you may see in [Table 6](#).

5. Results

All in all, our named entity recognition model with micro F1-score equal to 0.561 took the first place in the shared task. However, the results are lower than for other

named entity recognition datasets (e.g. for the Ontonotes dataset Transformer-based models usually get > 0.85 in F1-score⁴). It can be attributed to the small number of training examples and complexity of the domain. While training relation extraction models we also did not use information about named entity types. The authors of SpanBERT [5] claim that it may improve model scores. It can be further investigated in future work.

Our end-to-end relation extraction model despite being one of the best solutions at the shared task was much worse than the model trained with manual annotations provided by the organizers. In future we will try to use approaches similar to pseudo labelling where we include only those named entity predictions that have high logit scores instead of all predictions. The difference in results also demonstrates that correct named entity labels are vital for relation extraction. The results may also benefit from a larger dataset.

Multi-task learning improved our results only in end-to-end relation extraction for BERT-based models. RoBERTa was better than BERT in this shared task in both relation extraction subtasks.

Some typical errors may be found in Table 7. All our models tend to generate a lot of false positives, e.g. for a single text we predicted 165 relations instead of 88 and 421 named entities instead of 354. This might be tweaked in future. Sometimes these mislabellings might be due to errors in the dataset. For relation extraction such mistakes tend to snowball. Another typical NER mistake is span mismatch, usually it is due to the model not including the whole phrase group into the prediction.

Table 7: Error examples

Task	Prediction	Test	Comment
NER	администрации Костромской области	—	Might be alabelling mistake
NER	производства бумаги	производства бумаги и бумажных изделий, прочих готовых изделий, прочих транспортных средств и оборудования	Only partial span overlap
NER	—	действующих производств	Not found
End-to-end RelEx	динамика региональной экономики -> NPS -> Положительная	динамика региональной экономики -> NPS -> Положительная	Correct
End-to-end RelEx	динамика региональной экономики -> NPS -> Экспорт	—	Relation does not exist

⁴ see <http://docs.deeppavlov.ai/en/master/features/models/ner.html>

6. Conclusion

In this work we present our system for RuREBus shared task held together with Dialog 2020 conference. The task consisted of 3 tracks: named entity recognition, relation extraction with provided named entity tags and end-to-end relation extraction. All tracks were considered as sequence labelling problems. We show that sequence labelling might be a decent approach for the relation extraction problem. We also attempted to use joint-task learning for relation extraction and named entity recognition. However, it only slightly improved our results for end-to-end relation extraction and was outperformed by single task learning in most cases. Yet it should be noted that in other unrelated domains it enhanced results of our models so joint tasks should be carefully chosen and it requires further investigation. The system took the first place in the named entity recognition track and the second place in the third track. For the second task we failed to submit the solution till the deadline but it was among the best systems. The systems for all tasks are based on Transformer models.

7. Acknowledgments

We would like to thank the organizers of the shared task. We believe that their work will be very helpful for the development of natural language processing for the Russian language. The contribution of Nikolay Arefyev to the paper was partially done within the framework of the HSE University Basic Research Program funded by the Russian Academic Excellence Project ‘5-100’.

References

1. *Baldini Soares, L. et al.*: Matching the Blanks: Distributional Similarity for Relation Learning. In: Arxiv.org. pp. 2895–2905 (2019).
2. *Devlin, J. et al.*: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
3. *Heist, N., Paulheim, H.*: Language-agnostic relation extraction from wikipedia abstracts. In: Lect. Notes comput. Sci. (Including subser. Lect. Notes artif. Intell. Lect. Notes bioinformatics). pp. 383–399 (2017).
4. *Ivanin V., Artemova E., Batura T., Ivanov V., Sarkisyan V., Tutubalina E., Smurov I.*: RuREBus-2020 Shared Task: Russian Relation Extraction for Business. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’iuternaiia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog”], Moscow, Russia (2020).
5. *Joshi, M. et al.*: SpanBERT: Improving Pre-training by Representing and Predicting Spans.
6. *Liu, Y. et al.*: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arxiv.org. (2019).
7. *Stenetorp, P. et al.*: brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proc. Demonstr. Sess. EACL 2012. Association for Computational Linguistics, Avignon, France (2012).

8. *Vaswani, A. et al.*: Attention is all you need. In: *Adv. Neural inf. Process. Syst.* pp. 5999–6009 (2017).
9. *Wolf, T. et al.*: HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv. abs/1910.0*, (2019).
10. *Zhang, Y. et al.*: Position-aware attention and supervised data improve slot filling. In: *EMNLP 2017—conf. Empir. Methods nat. Lang. Process. Proc.* pp. 35–45 (2017).

STABILITY OF TOPIC MODELING VIA MODALITY REGULARIZATION

Derbanosov R. (derbanosov@gmail.com)

National Research University Higher School of Economics,
Moscow, Russia

Bakhanova M. (marybakhanova@gmail.com)

Skolkovo Institute of Science and Technology;
National Research University Higher School of Economics,
Moscow, Russia

Probabilistic topic modeling is a tool for statistical text analysis that can give us information about the inner structure of a large corpus of documents. The most popular models—Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation—produce topics in a form of discrete distributions over the set of all words of the corpus. They build topics using an iterative algorithm that starts from some random initialization and optimizes a loss function. One of the main problems of topic modeling is sensitivity to random initialization that means producing significantly different solutions from different initial points.

Several studies showed that side information about documents may improve the overall quality of a topic model. In this paper, we consider the use of additional information in the context of the stability problem. We represent auxiliary information as an additional modality and use BigARTM library in order to perform experiments on several text collections. We show that using side information as an additional modality improves topics stability without significant quality loss of the model.

Key words: topic modeling, topic modeling stability, artm, topic models regularization

DOI: 10.28995/2075-7182-2020-19-198-210

ПОВЫШЕНИЕ СТАБИЛЬНОСТИ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ С ПОМОЩЬЮ ДОПОЛНИТЕЛЬНОЙ МОДАЛЬНОСТИ

Дербаносов Р. (derbanosov@gmail.com)

Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Баханова М. (marybakhanova@gmail.com)

Сколковский институт науки и технологий;
Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

1. Introduction

Topic modeling is a statistical method for analyzing a corpus of documents. The result of the modeling is a set of topics. Each topic is usually represented as a discrete distribution over the set of all words in the corpus. Some applications of topic modeling are information search [30], [31], [13], analysis of text documents [3], [25], [28], [29], images and video data [8], [11], [20], audio data [32], problems of bioinformatics [23], [24].

The most popular algorithms for topic modeling solve the task of stochastic matrix factorization i. e. approximate representation of a stochastic matrix F as a product of two stochastic matrices $F \approx \Phi\Theta$. Matrix F is obtained from the collection of texts by assigning $F[i, j]$ to the number of occurrences of i -th word in j -th document and column normalization. Matrix $F \in \mathbb{R}^{|W| \times |D|}$ is usually called word-document matrix, where $|W|$ is a number of words and $|D|$ is a number of documents in the corpus. Matrices $\Phi \in \mathbb{R}^{|W| \times |T|}$ and $\Theta \in \mathbb{R}^{|T| \times |D|}$ are called word-topic matrix and topic-document matrix, where $|T|$ is a number of topics that is usually fixed before run of the algorithm. If we fix some stochastic matrix factorization $F \approx \Phi\Theta$ we may interpret distributions in columns of the matrix Φ as topics.

Two most popular approaches to the topic modeling are *Probabilistic Latent Semantic Analysis* (PLSA) [10] and *Latent Dirichlet Allocation* (LDA) [3]. The basic hypothesis of the PLSA model is the conditional independence hypothesis: the probability of a word occurrence in a document is conditionally independent of the document given a topic. LDA is a Bayesian version of PLSA. The main assumption of the LDA model is that ϕ_{wt} and θ_{td} are generated from the Dirichlet distribution. *Additive Regularization of Topic Models* (ARTM) [25], [28], [14] extends the formulation of PLSA by adding different regularizers to the loss function. Some of them are described in 2.2.

Usually algorithms use random initialization and then converge to some local optimum. One of the main problems of topic modeling is instability i. e. convergence

to different solutions from different initializations. Mathematical origins of this issue were studied in [6], [9], [18], [5] where authors research the problem of uniqueness of Nonnegative Matrix Factorization (NMF). Another approach to the problem is customization of basic algorithms to achieve better stability.

In the paper [2] the authors proposed ensemble methods and compared their performance with standard LDA and NMF approaches. The idea of their K-Fold method is to train several base topic models, transform them into the intermediate representation and build the final topic model on the top of this representation. According to experiments performed on annotated text corpora, K-Fold ensemble strategy can produce more stable and accurate topic models.

The authors of [15] proposed a modification of the standard latent Dirichlet allocation (LDA) model called granulated LDA (GLDA). The method is based on local density regularization that assigns the same topic with high probability to the words that meet together in the context. As for evaluation of model stability, the authors used Jaccard similarity and the number of stable topics based on Kullback-Leibler distance. The study shows that GLDA seems to reduce instability while yielding the same topic quality as classical topic models.

There are several studies [12], [33], [21] that show positive influence of additional information about documents in the collection on topic modeling performance. In this paper, we propose a method of increasing the stability that uses multimodal topic modeling. We use words as a first modality and different types of tags as additional modalities. We show that even using a partially labeled corpus (5% or 20% of the whole corpus) may increase the stability of PLSA model without significant loss of model quality.

2. Background

2.1. PLSA

Let D be a collection of documents, and let W be its vocabulary. The idea of probabilistic topic modeling is to describe how a collection of documents D is generated by a finite set of topics T . According to PLSA [10], the term distribution in each document $d \in D$ can be decomposed as a mixture of term probabilities for topics and topic probabilities for documents:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d), d \in D, w \in W, \quad (1)$$

where $\phi_{wt} = p(w|t)$ is the distribution of words in topics and $\theta_{td} = p(t|d)$ is the distribution of topics in documents. The parameters ϕ_{wt} and θ_{td} form stochastic matrices Φ and Θ . The problem of finding these matrices can be considered as an approximate matrix factorization task $F \approx \Phi\Theta$, $F = (\hat{p}_{wd})_{|W| \times |D|}$, where $\hat{p}_{wd} = n_{wd}/n_d$ is a frequency estimate of the conditional probability $p(w|d)$, n_d is the length of the document d , n_{wd} is the number of occurrences of the word w in the document d .

Parameters of the PLSA model are estimated via maximizing log-likelihood function with linear constraints:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{wd} \log \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta} \quad (2)$$

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \quad (3)$$

The process of solving this problem consists of random initialization of the matrix Φ and application of EM algorithm.

Most Bayesian approaches, such as LDA, use a prior Dirichlet distribution as the main regularizer, thus complicate the combination with other regularizers. ARTM is a modern extension of PLSA model proposed in [25] that is free from excess probabilistic assumptions. It does not require parameters to be generated from Dirichlet distribution and allows to use different regularizers that may have no probabilistic interpretation at all. Suppose $R_i(\Phi, \Theta)$, $i = 1, 2, \dots, n$ are n regularizers that we want to maximize along with the likelihood $L(\Phi, \Theta)$. In ARTM, we solve multi-criteria problem via maximization of the linear combination of L and R_i with some nonnegative regularization coefficients τ_i :

$$R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta), \quad (4)$$

$$L(\Phi, \Theta) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}.$$

Matrices Φ and Θ are estimated using EM algorithm, which can be described by two iteratively repeated steps.

At the E-step, we estimate the condition probability $p(t|d, w)$ for all words in documents (d, w) using Bayes formula:

$$p(t|d, w) = \frac{\phi_{wt} \theta_{td}}{\sum_{s \in T} \phi_{ws} \theta_{sd}}. \quad (5)$$

These probabilities are used to calculate parameters n_{wt} —the number of occurrences of the word w in the collection D with relation to the topic t and n_{td} —the number of words in the document d with relation to the topic t .

$$n_{wt} = \sum_{d \in D} n_{wd} p(t|d, w), \quad n_{td} = \sum_{w \in W} n_{wd} p(t|d, w). \quad (6)$$

At the M-step, we calculate parameters ϕ_{wt} and θ_{td} as frequency estimates of the corresponding conditional probabilities:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R(\Phi, \Theta)}{\partial \phi_{wt}} \right)_+, \quad \theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R(\Phi, \Theta)}{\partial \theta_{td}} \right)_+, \quad (7)$$

where the sign \propto means that the distribution on the left is obtained after the normalization of the right expression, and $(x)_+ = \max\{x, 0\}$. Thus, we can add different regularizers to set necessary constrains to the topic model. In this work, we will use the following regularizers: smoothing, sparsing, decorrelation and modality.

2.2. Additive regularization of topic models

Smoothing regularizer. If we want ϕ_{wt} and θ_{td} to be close to some discrete distributions β_w and α_d in terms of Kullback–Leibler divergence we can use a smoothing regularizer:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \rightarrow \max_{\Phi, \Theta}, \quad (8)$$

where β_0 and α_0 are regularization coefficients. Hence, the M-th step of the algorithm gives equations:

$$\phi_{wt} \propto (n_{wt} + \beta_0 \beta_w)_+, \quad \theta_{td} \propto (n_{td} + \alpha_0 \alpha_t)_+. \quad (9)$$

It is recommended to use a prior Dirichlet distributions or Bayesian inference for distributions β_w and α_t . The effect of this regularizer is an increase in small values of ϕ_{wt} and θ_{td} due to a slight decrease in their large values. As a result, generated topics may include general vocabulary words, stop words and rare words that are usually excluded from topics.

Sparsing regularizer. Usually we assume that each word and each document relate to a small number of topics. It means that matrices Φ and Θ should be sparse. We can achieve it using a sparsing regularizer. One can notice that sparsing is an inverse procedure to smoothing. Therefore, sparsing and smoothing differ only in the sign of parameters β_w and α_t .

Decorrelation regularizer. Decorrelation regularizer formalises the requirement that topics have to differ from each other. It can be satisfied via minimizing the sum of covariances between distributions ϕ_{wt} and ϕ_{ws} for all pairs of topics t, s :

$$R(\Phi, \Theta) = -\tau \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max_{\Phi, \Theta}, \quad (10)$$

where τ is a regularization coefficient. In this case, the formula for the regularized M-step takes the form:

$$\phi_{wt} \propto (n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws})_+. \quad (11)$$

2.3. Multimodal topic modeling

Usually documents can be described not only by words but also by terms of other modalities [27]. For example, textual modalities are tags, n-grams, named entities and natural language words. The last one is what we used to deal with in topic modeling. Pictures and web-sites are non-textual modalities. We can consider documents as a set of tokens taken from different modalities. The diverse meta-data represented by modalities can be helpful for determining topics, and, vice-versa, topics may be used to predict missing meta-data.

Multimodal topic modeling occurred to be an effective approach for solving different problems. For example, for a given parallel collection of text translation we can model topics and then use them for the cross-language search. In this case, each language is considered as a modality. Experiments showed that the combination of parallel documents and bilingual dictionaries improves the quality of cross-language

search in comparison with models using only bilingual dictionaries [7]. Also, multimodal topic model can be applied for constructing recommendations [27]. In this study, the authors focused on the article recommendation in the online-platform, they used different modalities, such as words from text, user’s feedback, tags, authors and user-specified categories. According to the results, the combination of modalities reasonably improves recommendation ranking.

Multimodal topic model and the regularized EM-algorithm for this case were firstly introduced in [27].

Let M be a set of modalities, and let W_m , $m \in M$ be a vocabulary of modality m . These vocabularies do not intersect and can be united into the set $W = \sqcup_{m \in M} W_m$ containing terms of all modalities. A model of $p(w|d)$ is introduced for each modality W_m , $m \in M$:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d), \quad d \in D, \quad w \in W_m. \quad (12)$$

The main concept of such modeling is that topics $p(t|d)$ are the same for all modalities. As for the distribution of words in topics, the matrices $\Phi_m = (\phi_{wt})_{|W_m| \times |T|}$ are normalized separately and stacked vertically into the matrix $\Phi = (\phi_{wt})_{|W| \times |T|}$.

If we consider the log-likelihood of each modality as a regularizer with coefficient τ_m , then the optimization problem has the following form:

$$\sum_{m \in M} \sum_{d \in D} \sum_{w \in W_m} \tau_m n_{wd} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}, \quad (13)$$

$$\sum_{w \in W_m} \phi_{wt} = 1, \quad \phi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \quad (14)$$

3. Metrics

3.1. Quality of topic modeling

There are several metrics for measuring quality of topic model. Most previous works have exclusively focused on perplexity measure that describes the speed and level of convergence of the model. Perplexity can be represented as an inverse function of the likelihood of model parameters. One of the drawbacks of this metric is that it depends on the data size, therefore, it is hard to compare results of this measure obtained from models trained on different datasets.

Some recent studies [15], [1] evaluate their models by pairwise information based metric called *coherence* [22]. The practical meaning of coherence follows a simple idea: if we describe the topic as a set of words then these words are likely to meet together in the context. In addition, coherence seems to reflect well the interpretability of topics [1]. Let k be an adjustable parameter meaning the number of top words in the topic $t \in T$, and let $W^t = \{w_1, \dots, w_k\}$ be the corresponding set of top words. Then coherence formula for topic t is defined as follows:

$$C_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^k \left(\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \right)_+, \quad (15)$$

where probabilities can be estimated by frequencies:

$$p(u, v) = \frac{n(u, v)}{n}, \quad p(u) = \frac{n(u)}{n}, \quad (16)$$

$$n(u) = \sum_{w \in W} n(u, w), \quad n = \sum_{w \in W} n(w). \quad (17)$$

There are different types of calculating co-occurencies $n(u, v)$. In this paper, we calculate in how many documents the pair (u, v) occurred at least once:

$$n(u, v) = \sum_{d=1}^{|D|} [\exists i, j : w_{id} = u, w_{jd} = v]. \quad (18)$$

We have described above how to compute coherency only for one topic. To obtain the coherence score for the topic model we simply average coherencies for all topics in the model. The higher coherency is, the better.

3.2. Stability of topic modeling

Let's denote by $\{M_1, M_2, \dots, M_r\}$ the set of topic models generated as a result of r runs of the algorithm on the same data. Assume that these models are similar if their topics are similar. To measure similarity between two topics represented by t top words, we propose to calculate the measure that we call *Stable Words* (SW), and describe it by the following formula:

$$SW(R_i, R_j) = \frac{|R_i \cap R_j|}{t}, \quad (19)$$

where R_i is a list of top t tokens of topic i . SW takes values in $[0, 1]$, and the value 1 corresponds to the identical top words. SW can be interpreted as a modified Jaccard Index, these two metrics differ only by the denominator: in Jaccard Index we divide by the set union size $|R_i \cup R_j|$. We consider SW as more interpretable measure in terms of topic stability because it is simply the share of stable words.

We should find topic correspondence between two sets of topics in order to compute similarity of these two sets. The best topic matching between two models with $|T|$ topics can be found using SW defined in Eq. 20. We construct a matrix S , where the element s_{ij} represents similarity between the i -th topic of the first model and the j -th topic of the second model. Then we find the optimal matching P by solving the minimal weight bipartite matching problem applying the Hungarian algorithm [16].

To obtain the score of similarity between the set of r models we compute *Average Stable Words* (ASW):

$$ASW = \frac{2}{r(r-1)} \sum_{i \leq j, i \neq j} \frac{1}{|T|} \sum_{s=1}^{|T|} SW(R_{is}, R_{j\pi(s)}), \quad (20)$$

where $\pi(s)$ is a topic of the model j matched to the topic s of the model i .

4. Experiments

We performed experiments on five texts collections: 20NewsGroups, Reuters52, Cade, WebKB and Habr. 20NewsGroups [17] is a set of documents classified in 20 news-groups. Reuters52 [19] is a collection of articles of 1987 year from Reuters that was manually classified by Reuters Ltd. The documents in WebKB¹ are webpages collected by the World Wide Knowledge Base project of the CMU text learning group. Cade is a subset of web pages extracted from the CADÉ Web Directory which points to Brazilian webpages labeled by human experts [4]. Habr is a dataset of articles from IT blogging platform <http://habrahabr.ru> [27] with 5 modalities: text of the blogpost, author, users that leave comments at the blogpost, hub that is a site section, tags that are generated by the author. We used preprocessing from [4] for 20NewsGroups, Reuters52, Cade and WebKB datasets. All datasets were splited by train and test sets in a ratio of 60 to 40. Coherence was measured on hold-out test dataset.

We take text labels as an additional modality for 20NewsGroups, Reuters52, Cade and WebKB datasets. Each document contains one token of such a modality. We use different percentage of labeled documents: 5%, 20%, 50% and 100% in order to simulate partially labeled collection. We tested tags, habs modalities and the combination of four modalities: authors, tags, hubs and users on Habr dataset.

All experiments were performed using an open source library for topic modeling BigARTM [26]. Models were trained until convergence on the train part of a dataset.

Each topic can be described as a set of the most frequent words of this topic. Several descriptions in terms of top words for 20NewsGroups are presented in **Table 1**.

Table 1: Example of 5 topics on 20NewsGroups with number of topics $|T|=10$

% of labels	Top 10 words
0	topic 1: game, team, plai, player, win, season, hockei, last, score, leagu topic 2: space, nasa, research, univers, gov, orbit, launch, program, center, system topic 3: car, price, sale, bui, want, mail, sell, speed, apr, engin topic 4: gun, state, israel, law, isra, govern, weapon, american, right, arab topic 5: kei, encrypt, chip, govern, secur, clipper, system, presid, public, work
50	topic 1: game, team, plai, player, win, season, hockei, last, leagu, score topic 2: space, nasa, scsi, system, control, orbit, work, card, launch, data topic 3: car, wire, ground, engin, power, work, water, back, want, light topic 4: gun, state, israel, isra, bike, weapon, kill, apr, law, arab topic 5: kei, govern, encrypt, system, secur, chip, presid, clipper, public, program

¹ <http://www.cs.cmu.edu/~webkb/>

% of labels	Top 10 words
100	topic 1: window, game, team, plai, win, hockey, file, player, season, nhl topic 2: space, nasa, work, presid, govern, orbit, state, launch, system, program topic 3: car, engen, want, come, back, work, speed, price, start, auto topic 4: gun, weapon, law, state, firearm, fire, govern, crime, control, arm topic 5: kei, armenian, encrypt, chip, govern, israel, secur, isra, system, turkish

To measure stability of the set of models generated over $r = 100$ runs, we used ASW (Eq. 21) with top $t = 10$ tokens for each topic. The estimation of the quality of the models was conducted using coherence score (Eq. 16) based on top $t = 10$ terms for each topic. We tried several set of hyperparameters for each model and present results with the best coherence. We performed several experiments with usual decorrelation, sparcing and smoothing regularizations.

Table 2: Topic stability and quality on 20NewsGroups with number of topics $|T|=10$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.86 ± 0.01	0.16 ± 0.01
Words, 50% of labels	Labels modality	0.86 ± 0.01	0.18 ± 0.01
Words, 20% of labels	Labels modality	0.83 ± 0.01	0.21 ± 0.01
Words, 5% of labels	Labels modality	0.79 ± 0.01	0.24 ± 0.01
Words	—	0.78 ± 0.01	0.26 ± 0.02
Words	Decorrelation Φ	0.77 ± 0.01	0.26 ± 0.02
Words	Sparcing Θ	0.75 ± 0.01	0.28 ± 0.02
Words	Smoothing Φ	0.53 ± 0.05	0.90 ± 0.14
Words	Sparcing Φ	0.53 ± 0.01	0.40 ± 0.02

Table 3: Topic stability and quality on 20NewsGroups with number of topics $|T|=60$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.76 ± 0.00	0.35 ± 0.01
Words, 50% of labels	Labels modality	0.70 ± 0.00	0.46 ± 0.01
Words, 20% of labels	Labels modality	0.63 ± 0.01	0.55 ± 0.01
Words, 5% of labels	Labels modality	0.60 ± 0.01	0.62 ± 0.01
Words	—	0.61 ± 0.01	0.62 ± 0.02
Words	Decorrelation Φ	0.60 ± 0.01	0.62 ± 0.02
Words	Sparcing Θ	0.12 ± 0.00	0.10 ± 0.02
Words	Smoothing Φ	0.43 ± 0.08	0.68 ± 0.58
Words	Sparcing Φ	0.25 ± 0.00	0.74 ± 0.01

Table 4: Topic stability and quality on Reuters52 with number of topics $|T|=60$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.65±0.01	0.72±0.01
Words, 50% of labels	Labels modality	0.60±0.01	0.81±0.01
Words, 20% of labels	Labels modality	0.56±0.00	0.87±0.01
Words, 5% of labels	Labels modality	0.54±0.01	0.87±0.01
Words	—	0.53±0.01	0.90±0.01
Words	Decorrelation Φ	0.52±0.01	0.91±0.01
Words	Sparcing Θ	0.08±0.00	0.07±0.01
Words	Smoothing Φ	0.68±0.05	0.53±0.19
Words	Sparcing Φ	0.20±0.00	0.83±0.02

Table 5: Topic stability and quality on Cade with number of topics $|T|=10$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.75±0.02	1.30±0.03
Words, 50% of labels	Labels modality	0.72±0.02	1.27±0.03
Words, 20% of labels	Labels modality	0.71±0.02	1.31±0.02
Words, 5% of labels	Labels modality	0.69±0.02	1.32±0.03
Words	—	0.69±0.02	1.33±0.02
Words	Decorrelation Φ	0.69±0.02	1.33±0.02
Words	Sparcing Θ	0.71±0.02	1.37±0.02
Words	Smoothing Φ	0.45±0.02	1.57±0.11
Words	Sparcing Φ	0.50±0.01	1.43±0.04

Table 6: Topic stability and quality on WebKB with number of topics $|T|=10$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.70±0.02	0.37±0.02
Words, 50% of labels	Labels modality	0.65±0.02	0.43±0.02
Words, 20% of labels	Labels modality	0.66±0.02	0.44±0.01
Words, 5% of labels	Labels modality	0.64±0.02	0.47±0.01
Words	—	0.64±0.02	0.49±0.02
Words	Decorrelation Φ	0.64±0.02	0.49±0.02
Words	Sparcing Θ	0.54±0.02	0.46±0.03
Words	Smoothing Φ	0.68±0.04	0.31±0.04
Words	Sparcing Φ	0.40±0.01	0.53±0.02

Table 7: Topic stability and quality on Habr with number of topics $|T|=60$

Modality	Regularizer	ASW	Coherence
Words, authors, users, tags, hubs	Combination of modalities	0.73 ± 0.00	0.40 ± 0.01
Words, tags	Tags modality	0.63 ± 0.00	0.56 ± 0.01
Words, hubs	Hubs modality	0.54 ± 0.01	0.73 ± 0.02
Words	Smoothing Φ	0.51 ± 0.06	0.27 ± 0.09
Words	Decorrelation Φ	0.51 ± 0.01	0.77 ± 0.02
Words	—	0.51 ± 0.01	0.77 ± 0.02

The results of topic modeling on 20NewsGroups, Reuters52, Cade and WebKB datasets (Tables 2–6) indicate that increase in the percentage of labels leads to stability growth. Moreover, models with regularizers, such as sparcing and smoothing, yield very low values of ASW compared to models with labels modality. Even 5% or 20% of labels may be enough to significantly increase model stability. However, we observe a drop in coherence score, especially in the models with high percentage of labels. Note, that models with labels modality trained on Reuters52 produce comparable and even higher coherence than models with other regularizers.

Experiments on Habr dataset show that the model combination of all five modalities outperforms all other models in terms of stability measure (Table 7). We see that the use of one additional modality—hubs or tags—increases ASW score but results in a slight decrease of quality in comparison with the use of other regularizers.

Overall, we conclude that models with different modalities, such as labels and additional meta-data, produce more stable topics. At the same time, the model with labels modality may yield low coherence score if the percentage of labels is high.

5. Conclusion

Modern topic modeling approaches suffer from instability of their results even with fixed dataset and hyperparameters. We have demonstrated that stability of topic modeling algorithm may be improved with the help of side information. Evaluation on several text corpora shows that regularization of the PLSA model with additional modalities leads to less impact of random initialization and thus more stable modeling even if side information was provided only for some subset of documents.

While our experiments were conducted on five significantly different datasets, it is still an open question what combination of additional information is the best choice for improving stability with the smallest degradation of metrics of a model. The topic for further research is to find a combination of various regularizers with the best balance between modeling stability and the quality of topics.

References

1. *Alekseev, V. et al.*: Intra-text coherence as a measure of topic models' interpretability. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference 'Dialogue 2018'*. 100–108 (2018).
2. *Belford, M. et al.*: Stability of topic modeling via matrix factorization. *Expert Syst. Appl.* 91, 159–169 (2017).
3. *Blei, D. et al.*: Latent dirichlet allocation. *Journal of Machine Learning Research.* 3, 993–1022 (2003).
4. *Cardoso-Cachopo, A.*: Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa (2007).
5. *Derbanosov, R., Irkhin, I.*: Issues of stability and uniqueness of stochastic matrix factorization. *Computational Mathematics and Mathematical Physics.* 3, (2020).
6. *Donoho, D., Stodden, V.*: When does non-negative matrix factorization give a correct decomposition into parts? In: *Advances in neural information processing systems* 16. pp. 1141–1148 MIT Press (2004).
7. *Dudarenko, M.*: Regularization of multilingual topic models. *Vychisl. Metody Programm.* 16, 26–38 (2015).
8. *Feng, Y., Lapata, M.*: Topic models for image annotation and text illustration. In: *Human language technologies: The 2010 annual conference of the north American chapter of the association for computational linguistics.* pp. 831–839 Association for Computational Linguistics, Los Angeles, California (2010).
9. *Gillis, N.*: Sparse and unique nonnegative matrix factorization through data pre-processing. *The Journal of Machine Learning Research.* 13, 3349–3386 (2012).
10. *Hofmann, T.*: Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval.* pp. 50–57 ACM, New York, NY, USA (1999).
11. *Hospedales, T. M. et al.*: Video behaviour mining using a dynamic topic model. *International Journal of Computer Vision.* 98, 303–323 (2011).
12. *Hu, C. et al.*: Non-negative matrix factorization for discrete data with hierarchical side-information. (2016).
13. *Ianina, A. et al.*: Multi-objective topic modeling for exploratory search in tech news. In: *Artificial intelligence and natural language.* pp. 181–193 Springer International Publishing, Cham (2018).
14. *Kochedykov, D. et al.*: Fast and modular regularized topic modelling. In: *2017 21st conference of open innovations association (FRUCT).* IEEE (2017).
15. *Koltcov, S. et al.*: Stable topic modeling with local density regularization. In: *Internet science.* pp. 176–188 Springer International Publishing, Cham (2016).
16. *Kuhn, H.*: The hungarian method for the assignment problem. In: *Naval Research Logistics Quaterly.* pp. 83–97 (1955).
17. *Lang, K.*: Newsweeder: Learning to filter netnews. In: *Proceedings of the twelfth international conference on machine learning.* pp. 331–339 (1995).
18. *Laurberg, H. et al.*: Theorems on positive data: On the uniqueness of NMF. *Computational Intelligence and Neuroscience.* 2008, 1–9 (2008).
19. *Lewis, D. D. et al.*: RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5, 361–397 (2004).

20. *Li, X.-x. et al.*: Simultaneous image classification and annotation based on probabilistic model. *The Journal of China Universities of Posts and Telecommunications*. 19, 107–115 (2012).
21. *Mehrotra, R. et al.*: Improving LDA topic models for microblogs via tweet pooling and automatic labeling. Presented at the (2013).
22. *Newman, D. et al.*: Automatic evaluation of topic coherence. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. pp. 100–108 (2010).
23. *Pritchard, J. et al.*: Inference of population structure using multilocus genotype data. *Genetics*. 155, 945–959 (2000).
24. *Shivashankar, S. et al.*: Multi-view methods for protein structure comparison using latent dirichlet allocation. *Bioinformatics*. 27, i61–i68 (2011).
25. *Vorontsov, K.*: Additive regularization for topic models of text collections. *Doklady Mathematics*. 89, 301–304 (2014).
26. *Vorontsov, K. et al.*: BigARTM: Open source library for regularized multimodal topic modeling of large collections. In: *AIST*. pp. 370–384 (2015).
27. *Vorontsov, K. et al.*: Non-bayesian additive regularization for multimodal topic modeling of large collections. In: *TM '15*. (2015).
28. *Vorontsov, K., Potapenko, A.*: Additive regularization of topic models. *Machine Learning*. 101, 303–323 (2014).
29. *Vorontsov, K., Potapenko, A.*: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. 29–46 (2014).
30. *Vulic, I. et al.*: Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*. 16, 331–368 (2012).
31. *Vulic, I. et al.*: Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Inf. Process. Manage.* 51, 111–147 (2015).
32. *Wang, W.*: Instantaneous versus convolutive non-negative matrix factorization. *Machine Audition: Principles, Algorithms and Systems*. 353–370 (2011).
33. *Zhao, H. et al.*: Leveraging external information in topic modelling. *Knowledge and Information Systems*. 61, 661–693 (2018).

ДИФФЕРЕНЦИАЛЬНЫЕ СЕМАНТИЧЕСКИЕ СКЕТЧИ ДЛЯ РУССКОЯЗЫЧНЫХ ИНТЕРНЕТ-КОРПУСОВ

Деткова Ю. (julia.detkova@abbyy.com)

АВВУУ Lab МФТИ

Новицкий В. (valeriy.novitskiy@abbyy.com),

Петрова М. (m.petrova@abbyy.com),

Селегей В. (vladimir.selegey@abbyy.com)

АВВУУ

В статье описывается новый тип агрегированной корпусной выдачи — семантические скетчи, получивший пробную реализацию на одном из подкорпусов ГИКРЯ. Семантические скетчи являются естественным распространением идеи корпусных скетчей на анализ сочетаемости в терминах семантических отношений и семантических классов. Уточняющий атрибут «дифференциальный» означает возможность дополнительной параметризации скетчей метатекстовыми характеристиками. Разумеется, построение таких скетчей требует семантической разметки корпуса, в качестве которой в данной работе использовались частичные семантические разборы Comprepro. В статье приводятся примеры построенных скетчей и оцениваются достоинства и проблемы корпусной статистики такого рода.

Ключевые слова: семантические скетчи; автоматическая семантическая разметка; семантические классы; глубинные позиции; лексическая сочетаемость, ГИКРЯ

DOI: 10.28995/2075-7182-2020-19-211-227

DIFFERENTIAL SEMANTIC SKETCHES FOR RUSSIAN INTERNET-CORPORA

Detkova J. (julia.detkova@abbyy.com)

АВВУУ Lab MIPT

Novitskiy V. (valeriy.novitskiy@abbyy.com),

Petrova M. (m.petrova@abbyy.com),

Selegey V. (vladimir.selegey@abbyy.com)

АВВУУ

The current paper suggests a new representation type of word collocations—the semantic sketches. It was first tested on one of the subcorpora of the General Internet-Corpus of Russian. The semantic sketches continue the idea of word sketches based on grammatical relations between words and expand it by adding the semantic information—word meanings and semantic relations between words. Moreover, the sketches can be additionally provided with metatextual characteristics.

Certainly, building such sketches demands the semantic markup of the corpora. Therefore, we have used partial semantic analysis of the Compreno parser for our purposes. The paper demonstrates the examples of the sketches, provides the quality evaluation of the markup they are based on, and shows the advantages and disadvantages of the given approach.

Keywords: semantic sketches; corpora semantic markup; semantic classes; semantic relations; lexical collocations, GICR, Compreno

1. Введение

Идея представления корпусной выдачи в виде статистики сочетаемости анализируемого слова с синтаксически связанными лексическими единицами была предложена и реализована Адамом Килгарифом в рамках проекта SketchEngine [9].

Информация о сочетаемости слова представляется в виде скетчей — обобщенных лексикографических портретов, где зависимые слова классифицируются по грамматическим отношениям: для дочерних зависимых указывается, являются ли они определением, субъектом, объектом или соответствуют другой синтаксической роли.

Подобные скетчи не только позволяют продемонстрировать в сжатом виде основную сочетаемость слова, но также являются удобным инструментом для сравнения сочетаемости разных слов, что бывает полезно при сопоставлении разных лексических единиц одного лексико-семантического поля.

Получение скетчей может основываться на уже имеющейся [частичной] синтаксической разметке корпусов или использовании так называемых скетч-грамматик, имеющих вид контекстно-свободных правил, которые строят отдельные типы синтаксических связей на основе автоматической морфоразметки корпуса [8].

Идея построения синтаксических скетчей оказалась очень продуктивной, а многоязычные интернет-корпуса, собранные в рамках проекта SketchEngine, стали востребованным инструментом корпусных лингвистических исследований.

Надежность и польза полученных скетчей для исследователя ограничена сегодня следующими факторами:

- состав корпуса (зависимость от состава корпуса очевидна, но редко анализируется [4]);
- качество установленных связей (его влияние на получаемую статистику также на удивление остается без внимания);
- невозможность строить скетчи с учетом лексической омонимии.

Последний пункт особенно важен: результаты агрегации по всем значениям лексической единицы сложно интерпретировать, и эту проблему нельзя решить на уровне синтаксиса.

Чтобы снять лексическую неоднозначность, необходимо дополнить имеющийся инструментарий семантическим анализом. Это позволило бы не только решить (насколько это возможно) проблему омонимии, но и учесть диатезное варьирование при реализации семантических отношений.

Целью данной работы является создание семантических скетчей. В них зависимые классифицируются уже не по грамматическим отношениям, а в терминах семантических ролей, таких как Агент, Экспериенсер, Объект, Локатив и т. п., и для каждого слова, помимо грамматических признаков, учитывается также его семантическое значение.

Кроме того, хотелось бы анализировать лексикографические портреты лексем с учетом жанрово-тематического и социолингвистического варьирования [4].

Для решения этой задачи нужны дифференциальные корпуса¹ и достаточно надежные семантические анализаторы, основанные на разметке, которая устроила бы лингвистов и лексикографов.

В качестве корпуса мы естественным образом выбрали ГИКРЯ, поскольку других дифференциальных корпусов для РЯ пока нет. «Ближайший» корпус со свойством дифференциальности [17] реализует пока только идею сегментной различимости данных для лингвистического анализа.

Сама идея скетчей на основании семантических отношений достаточно очевидна, но ее реализации препятствует отсутствие сегодня систематической корпусной семантической разметки. Создать такую разметку с помощью аналога скетч-грамматик, оперирующих синтаксическими отношениями, вряд ли возможно.

Естественным решением данной проблемы представляется прямое использование семантических парсеров. Остановимся подробнее на возможностях их применения для построения семантических скетчей с указанными характеристиками.

2. Использование семантических парсеров при создании семантических скетчей

В настоящее время исследования в области *frame semantic parsing* ведутся весьма активно, достаточно упомянуть хотя бы популярную для такого анализа DL-платформу SLING [13], основанную на разметке OntoNotes.

Возможность оценки SOTA в этой области (см. [раздел 4](#)) дает проведенная на ACL2019 shared task по семантической разметке [11]. Участникам было предложено пять фреймворков для семантической разметки: DELPH-IN MRS Bi-Lexical Dependencies [6], Prague Semantic Dependencies [7], Elementary Dependency

¹ При дифференциальном подходе [4] всякая корпусная статистика параметризуется метатекстовыми характеристиками: от года создания и социолингвистических характеристик автора до сегментов Интернета, что позволяет обнаруживать статистически значимые различия (смещения) в корпусной выдаче.

Structures [10], Universal Conceptual Cognitive Annotation [1] и Abstract Meaning Representation [3]. Из перечисленных систем разметки только UCCA поддерживает атрибуты ребер, остальные же ограничиваются метками вида “argN”, сравнивать которые нецелесообразно.

Анализ результатов Shared Task on Cross-Framework Meaning Representation Parsing показал, что существующие open-source решения не подходят для семантического анализа русского языка по нескольким причинам.

1. Отсутствие большого разнообразного по составу качественно размеченного семантического корпуса. Это относится ко всем языкам, и к РЯ в частности. Например, на том же треке ACL’2019 для обучения использовались сравнительно небольшие датасеты, объем максимального из которых составляет 56 тысяч предложений. Некоторая семантическая разметка без снятия лексической омонимии реализована в подкорпусе НКРЯ [16], для построения скетчей она, разумеется, не подходит.
2. Отсутствие стандарта семантической разметки: на уже упоминавшемся треке ACL использовались 5 альтернативных систем разметок, и очевидно, акцент был сделан на машинном обучении, а не на сравнительном анализе подходов.
3. Предложенные фреймворки не используют информацию о семантических ролях зависимых, либо используют её очень ограниченно [11].

По указанным выше причинам мы решили использовать для разметки семантическую модель и парсер Comrgeno [2], предоставленный авторам для исследовательских целей по академической ограниченной лицензии².

Основные его характеристики представлены в [2], семантическая часть модели описана в [12], некоторые примеры также представлены в приложении. Отметим две составляющие модели Comrgeno, релевантные для целей настоящего исследования:

- во-первых, это организация лексики: все слова представлены в виде семантической иерархии тезаурусного типа, где каждому слову соответствует свой семантический класс (СК), определяющий значение слова и его место в иерархии. Разделение по семантическим классам позволяет рассматривать отдельно разные значения слова;
- во-вторых, — организация семантических отношений между словами: семантические отношения представлены в виде глубинных позиций (ГП). Если валентности в традиционном понимании покрывают, в основном, только актантные зависимые, то ГП описывают все возможные зависимые, как актантные, как и сирконстантные, что позволяет выводить в скетчах полную сочетаемость слова.

² Данная лицензия позволяет размечать и выводить в открытый доступ корпуса ограниченного объема. Кроме того, АБВУУ как участник проекта ГИКРЯ предоставляет разработчикам корпуса возможность использовать разметку для целей лингвистических исследований в отдельных подкорпусах, включая отдельные SQL-запросы и агрегацию их результатов.

Мы использовали не полную семантико-синтаксическую разметку Compreno, а только подмножество разметки с указанием СК и ГП, актуальное для построения семантических скетчей.

Рассмотрим далее особенности выбранной разметки и корпус, на котором строились семантические скетчи.

3. Особенности корпуса и разметки

В настоящей работе для экспериментов был использован журнальный подкорпус ГИКРЯ.

Объем подкорпуса:

- 74 тысячи документов,
- 24,7 млн предложений,
- 326 млн слов.

Все тексты были размечены с помощью технологии ABBYY Compreno [15]. Результатом разметки является семантическая структура для каждого предложения в виде дерева семантических отношений между семантическими классами входящих в него слов.

Для целей данного исследования было решено ограничиться частичной семантической разметкой, где маркируются только поддеревья глаголов: все глаголы размечаются семантическими классами и глубинными позициями для их непосредственных зависимых; все зависимые, для которых определены ГП, маркируются также и по СК, что позволяет нам получать пары вида "глагол:СК [ГП: зависимая:СК]", например:

Окна выходят в сад.

"выходить:TO_FRONT" [Object: окно "окно:WINDOW_OF_BUILDING"]

"выходить:TO_FRONT" [в Locative_FinalPoint: сад "сад:сад:GARDEN"]

На семантических отношениях этих поддеревьев и строятся скетчи в первой версии.

Вершиной такого поддерева является семантический класс глагола (наиболее вероятный по результатам анализа), его дочерние элементы связаны с ним семантическими отношениями — ГП из некоторого набора Compreno (примеры используемых для построения скетчей ГП и иллюстративный фрагмент иерархии семантических классов приведены в приложении).

Из атрибутов ядра оставлены лексема, семантический класс, морфологические граммы категории GrammaticalType (часть речи) и синтаксические граммы категории SyntacticCase (падеж). На **рис. 1** ниже представлен разбор предложения с указанной информацией при лексемах и семантическими отношениями между ними.

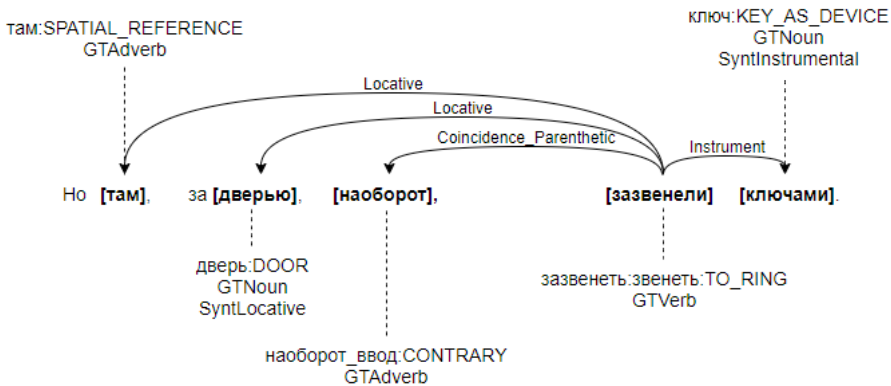


Рис. 1. Разбор предложения с указанными семантическими отношениями, семантическими классами и грамматическими значениями ядер составляющих

Таким образом, на данном этапе мы не размечали СК и ГП для прилагательных, именных групп при существительных и прочих неглагольных зависимых, для синтаксически перемещенных групп, а также зависимые при эллиптированных глаголах (в предложениях типа *Мы правы*. не маркируются ГП и СК при эллиптированном «быть») и сами эллиптированные зависимые (например, в предложении *Составители попросили каждого из поэтов сочинить...* подразумевается эллиптированное ядро «поэта» в составляющей «каждого (поэта) из поэтов»).

Большинство этих ограничений не является принципиальным, поскольку, во-первых, используемый парсер данную информацию предоставляет, а, во-вторых, степень влияния таких случаев на скетчевую выдачу предоставляется несущественной (что, впрочем, планируется проверить на следующих этапах исследования).

Местоимения размечаются классами вида PRONOUN в тех случаях, когда не удастся установить по контексту связь с референтом, и классом референта — когда такая связь устанавливается. В соответствии с этим возможны такие варианты разметки:

```
#— А как зовут "звать:TO_CALL_AND_DESIGNATE" Object: ee
"#pronoun_personal:#pronoun_personal:PRONOUN_BEING"?
```

— референт «ее» не найден;

```
#— А как зовут "звать:TO_CALL_AND_DESIGNATE" Object: ee
"болонка:PET_DOG"?
```

— связь с референтом «ее» установлена по контексту.

В итоге, размеченный корпус содержит:

- 55 тысяч различных семантических классов;

- 91 тысяча различных лексических классов (под лексическим классом понимается потомок семантического класса в конкретном языке);
- 305 различных глубинных позиций;
- 95 млн экземпляров семантических отношений — то есть, размеченных вхождений данных ГП в корпусе.

Разметка получена с помощью API, предоставленного АBBYU для исследований по ограниченной лицензии, и выгружена в локальную базу данных с примерами для сбора статистик. Также написан экспортер разметки в формат xml, необходимый для публикации корпуса-прототипа в open-source, ссылка на который представлена ниже.

4. Оценка качества семантической разметки

Для оценки качества разметки случайным образом было отобрано 200 предложений из журнального зала ГИКРЯ. Автоматическая разметка была проверена и исправлена вручную, и уже с полученной таким образом эталонной разметкой проводилось сравнение.

В качестве метрики была выбрана официальная метрика для ACL Shared Task 2019 — Meaning Representation Parsing [11]. Эта метрика (далее MRP F1) представляет собой невзвешенное среднее между F1-score для различных объектов разметки: узлы дерева разбора, метки узлов (семантические классы), атрибуты узлов (граммемы), привязка к тексту, связи, атрибуты связей (глубинные позиции).

Привязка к тексту сравнивается максимально мягко: не учитываются ошибки на пробельных символах, пунктуации, скобках и кавычках, не учитываются ошибки разбиения аннотации на несколько подряд идущих. MRP F1 разработана как универсальная метрика для сравнения деревьев разбора различных форматов представления семантической структуры и хорошо адаптируется к отсутствию любого из объектов разметки. Так, мы не учитывали качество определения граммем ядер составляющих, так как они не влияют напрямую на качество скетчевой выдачи. Кроме того, ввиду ограничений разметки, взятой для эксперимента, мы имеем дело не с полными деревьями разбора, а с их не обязательно связными подграфами. Выбранная метрика не противоречит этой модификации, но нужно учитывать, что сравнение с результатами MRP Shared Task 2019 не вполне корректно по нескольким причинам: мы оцениваем разметку на другом языке, другом корпусе и с упрощенной структурой.

В результате была получена следующая оценка для корпуса: MRP F1 = 96,78, что является достаточно высоким значением для автоматического парсера. Для сравнения, максимальное значение MRP F1, которого удалось достичь на ACL 2019 Shared Task, составляет 0,84 (оценка для 100 случайных предложений из The Little Prince) [11].

Основное количество ошибок было связано с выбором СК для нужного значения глагола (например, «выйти» в буквальном значении движения в «*статья вышла в свет*» вместо «выйти» в значении «появиться»), с выбором СК для

зависимой (например, "нос:NOSE_AS_FRONT_PART" в «из носу кровь» вместо ожидаемого «нос:NOSE»), а также с выбором ГП — как правило, в случае их возможной омонимии (например, Locative для «[в книге] написано» вместо предполагаемого MetaphoricLocative (разбор с неметафорическим локативом был бы уместен в предложении типа «[в книге] лежала закладка»)).

Более подробно с корпусом, использованным для оценки качества разметки, можно ознакомиться по ссылке: [SemanticSketchCorpora](#).

5. Семантические скетчи

Описанная семантическая разметка позволяет нам строить семантические скетчи с обозначенными выше характеристиками, а именно — с учетом разделения слова на разные значения, с группировкой зависимых по семантическим отношениям и возможностью единообразно представлять одинаковые семантические роли независимо от их разного синтаксического выражения при разных предикатах.

Рассмотрим теперь технические средства построения скетчей данного формата.

5.1. Обозначения

Здесь и далее используются обозначения в заголовках таблиц с примерами:

- **Lexeme** — лексема ядра составляющей
- **SemanticClass** — семантический класс составляющей
- **SlotName** — семантическая роль дочерней составляющей
- **ChildText** — текст заполнителя
- **Count** — поддержка коллокации в корпусе
- **F** — частота встречаемости коллокации, умноженная на 10^6 для удобства отображения
- **MI** — метрика ассоциации Mutual Information
- **Dice** — метрика ассоциации logDice

5.2. Метрики

Ранжирование только по частоте встречаемости сочетания в корпусе не представляет большого интереса, так как не несёт никакой информации о том, насколько то или иное словосочетание характерно для исследуемого слова или семантического класса. Поэтому мы предусмотрели возможность ранжировать скетчевую выдачу по нескольким метрикам:

Частота встречаемости: $f(x, y)$

Mutual Information: $MI(x, y) = \log_2 \left(\frac{f(x, y)}{f(x)f(y)} \right)$

logDice: $Dice(x, y) = 14 + \log_2 \left(\frac{2f(x, y)}{f(x) + f(y)} \right)$

Так как метрики MI и logDice чувствительны к случайным сочетаниям, было решено использовать их вместе с ограничением на частоту встречаемости в корпусе. Примеры, для которых она ниже некоторого устанавливаемого пользователем порога, не попадают в скетчевую выдачу. Преимущество метрики logDice хорошо видно на следующих примерах:

- (1) Топ-5 коллокаций глагола "сообщать:ТО_INFORM" по частоте встречаемости:

SlotName	ChildText	Count	f	MI	Dice
Agent	он	1042	287.53	3.97	6.75
Theme	об этом	694	191.5	8.3	9.65
Addressee	ему	580	160.04	7.36	9.14
Agent	она	394	108.72	3.93	6.65
Addressee	ей	247	68.16	7.18	8.78
Theme	о том	215	59.33	8.46	9.44

- (2) Топ-5 коллокаций глагола "сообщать:ТО_INFORM" по logDice:

SlotName	ChildText	Count	f	MI	Dice
Ch_EvaluationOfHumanTemperAndActivity	доверительно	144	39.73	14.09	12.08
Ch_Information	по секрету	49	13.52	14.56	11.49
Metaphoric_Route	по рации	22	6.07	14.34	10.84
TextStructure	настоящим	11	3.04	15.34	10.79
Ch_Information	под большим секретом	13	3.59	15.67	10.72
Ch_Information	конфиденциально	13	3.59	15.45	10.68

5.3. Механизм построения скетчей

Все данные, необходимые для построения скетчей, выгружаются из корпуса в базу данных SQLServer. Там же хранятся списки ГП и СК. Для ускорения вычисления метрик MI и logDice в базе сохранены предсчитанные частоты встречаемости для лексических и семантических классов ядер составляющих. Непосредственно вычисления оформлены в виде хранимых процедур T-SQL. Для получения данных из базы и визуализации скетчей использованы библиотеки Python: pyodbc, pandas, seaborn.

5.4. Семантические классы

Основным преимуществом нового инструмента является взаимодействие с семантической разметкой предложений. Это позволяет снять омонимию в большинстве контекстов: на этапе разметки для каждого слова при помощи парсера

определен наиболее вероятный класс, к которому слово может принадлежать в данном контексте. Как следствие, мы можем строить скетчи для различных значений слова. Можно продемонстрировать разницу на примере глагола «выходить»:

(3) Примеры для различных семантических классов глагола «выходить»:

Класс	Пример употребления
TO_FRONT	<i>Окна выходят в сад.</i>
TO_WALK	<i>Мальчик вышел из комнаты.</i>
TO_TAKE_PLACE	<i>Из этой затеи ничего не вышло.</i>
TO_TREAT_AND_CURE	<i>Выходить больного щенка.</i>

Приведем семантические скетчи для некоторых из данных значений (для удобства сравнения представим скетчи в виде таблиц без частотных характеристик, но с разделением по ГП, аналогично тому, как в Sketchengine представлено разделение по синтаксическому выражению). Здесь и далее примеры приведены в порядке убывания метрики logDice:

(4) Семантический скетч для глагола "выходить:TO_TAKE_PLACE":

Ch_Relation_Coincidence	Modality	Object_Situation	Ch_Evaluation	Time	Locative
наоборот вышло наоборот	само_собой вышло само собой	заминка вышла заминка	складно вышло складно	четверть выходила за четверть	сумма вышло в сумме
такой вышло так	так вышло так	скандалить вышел скандал	нехороший вышло нехорошо	как-то вышло как-то	рассказ выходит в рассказах гулко
иной вышло по-иному	на_деле вышло на деле	размолвка вышла размолвка	скверный вышло скверно	восьмой вышла восьмого марта	практик выходит на практике
похожий выходило очень похоже	неправдоподобный вышло крайне неправдоподобно	казус вышел казус	красивый вышло красиво	иной_раз выходит иной раз	конкурентка выходит у конкурентки
другой вышло по-другому	криво вышло криво	неприятность вышли неприятности	ничего вышло ничего	подчас выходило так подчас	Земский выходит у него
этакый вышло этак	кривовато выходило кривовато	ссориться вышла ссора	паскудный вышло паскудно	ноябрь выходило на прошлогодний ноябрь	Вяльцев выходит у вяльцева

(5) Семантический скетч для глагола "выходить: TO_FRONT":

Locative_FinalPoint	Object	Locative_PartAsOrientation	Locative_Orientation_FinalPoint	Time	Locative
двор выходили во двор	окно выходили окна	окно выходила окнами	запад выходили на запад	поныне выходит и поныне	Соня выходит у нас
сад выходили в сад	балкон выходили балкон	дверь выходили дверями	север выходит на север	частенький выходил частенько	Ярус выходили на втором ярусе
дворик выходило во внутренний дворик	фасад выходил его главный фасад	фасад выходили фасадом на твердую	восток выходит на восток	параллельный выходила параллельно	n_этажка выходит в каждой пятиэтажке
улица выходили на улицу	окошко выходило окошко	конец выходит одним концом	юг выходит на юг	как_раз выходило как раз	Питер выходит в питере
проспект выходили на проспект	веранда выходила веранда	стена выходила северной стеной	сторона выходит в сторону	иной_раз выходила иной раз	крыша выходили поверх низких крыш
север выходили на север	подъезд выходили на подъезды		напротив выходили напротив мох	осень выходил под осень	кафе выходили у него

Для сравнения приведем также скетч для «выйти», получаемый на sketch-engine.eu:

WORD SKETCH Russian Web 2011 (ruTenTen11)

выйти as verb 2,861,777 ×

subject	post_prep	pp_на	pp_из	pp_в	adv_modifier
книга вышла книга	из вышел из	улица вышел на улицу	строй вышел из строя	финал вышли в финал	замуж замуж вышла
версия Вышла новая версия	на вышел на	сцена вышел на сцену	мода вышли из моды	полуфинал вышла в полуфинал	скоро скоро выйдет
постановление вышло постановление	за вышла за	крыльцо вышел на крыльцо	комната вышел из комнаты	отставка вышел в отставку	недавно недавно вышла
фильм ошибочка вышла	около вышел около	пенсия вышел на пенсию	кабинет вышел из кабинета	эфир выйдет в эфир	впервые впервые вышел
альбом альбом вышел	во вышел во двор	балкон вышел на балкон	тюрьма вышел из тюрьмы	свет вышла в свет	вперед вперед вышел
указ вышел указ	через вышел через	экран вышел на экраны	ванная вышел из ванной	коридор вышел в коридор	вскоре вскоре
издание издание вышло в	к вышли к	ринг выйдет на ринг	употребление вышли из употребления	прокат выйдет в прокат	поспешно поспешно вышел
девушка девушка вышла	в вышел в	старт вышли на старт	подъезд вышел из подъезда	четвертьфинал вышла в четвертьфинал	навстречу навстречу вышел
	ко ...	орбита ...	печать ...	издательство ...	давно давно вышла

Как видно, выдача содержит разные значения данного глагола: «выйти на крыльцо, на балкон» соответствует СК TO_WALK выше, — «выйти в свет, в прокат, вышел альбом» — значению «появиться», «ошибочка вышла» — «произойти» (TO_TAKE_PLACE) и так далее.

Конечно, в ряде случаев омонимия разрешается неправильно и в нашей модели. Для примера приведем скетч "выйти:TO_WALK":

Locative_FinalPoint	Locative_InitialPoint	Time	Agent	Agent_Metaphoric	OrderInTimeAndSpace
улица вышел на улицу	дом вышел из дома	утро вышел утром	человек выходили люди	книга вышла книга	наконец вышел наконец
двор вышел во двор	комната вышел из комнаты	только_что вышел только что	женщина вышла женщина	издание вышло второе издание	потом вышел потом
коридор вышел в коридор	из_дому вышел из дому	минута вышел через минуту	мужчина вышел мужчина	срок вышел срок	наконец-то вышел наконец-то
сцена выходит на сцену	кабинет вышел из кабинета	вечер вышел вечером	девушка вышла девушка	сборник вышел сборник	затем вышел затем
крыльцо вышел на крыльцо	машина вышел из машины	ранний вышел рано	старик вышел старик	что_Interrog вышло чего	снова вышел снова
свет вышел в свет	подъезд вышел из подъезда	час вышел через полчаса	жена вышла жена	роман вышел роман	опять вышел опять

Как видно, примеры, попавшие в колонку с ГП Agent_Metaphoric, не на данное значение. Данная ошибка — результат неправильного семантического анализа, на который, в данном случае, влияет как модель, так и статистика. Это позволяет надеяться, что при дальнейшем использовании и обучении на текстах большего объема число подобных ошибок будет уменьшаться.

5.5. Глубинные позиции

Семантическая разметка позволяет оперировать не поверхностными, синтаксическими позициями, а глубинными, семантическими. Это, с одной стороны, позволяет различать разную семантику в отношениях, имеющих одинаковое синтаксическое выражение, как в примерах (6) ниже:

- (6) Примеры, в которых зависимые слова имеют одну поверхностную позицию Object_Dative, но разные глубинные:

Глубинная позиция	Пример
Addressee	Рассказать сказку [детям].
Experiencer	Позволить [гостю] войти.
Possessor	Дать [другу] книгу.

С другой стороны, дает возможность сопоставить с ГП релевантные семантические отношения даже в случае их разного синтаксического выражения. Сравним, к примеру, семантические скетчи двух лексических классов из СК TO_INFORM — «сообщать» (7) и «извещать» (8):

- (7) Семантический скетч для глагола "сообщать:TO_INFORM":

Agent	Addressee	Theme	Object_Situation	Time	Object
газета	читатель	это	адрес	мимоходом	новость
сообщали газеты	сообщает читателю	сообщил об этом	сообщил адрес	сообщает мимоходом	сообщил новость
радио	родитель	смерть	подробность	немедленный	сведения
сообщило радио	сообщил родителям	сообщила о его смерти	сообщить какие-либо подробности	сообщить немедленно	сообщил сведения шпионского характера
автор	жена	то	номер	между_делом	весть
сообщает автор	сообщил жене	сообщает о том	сообщили выигравший номер	сообщил между делом	сообщил радостную весть
агентство	мать	результат	собираться	незамедлительный	известие
сообщает агентство	сообщил матери	сообщить о результатах	сообщила и собирается бросить университет	сообщает незамедлительно	сообщить пренебрежительное известие
СМИ	друг	событие	следовать	вскользь	факт
сообщают сми	сообщил другу	сообщает об этом событии	сообщить куда следует	сообщается вскользь	сообщать факты
источник	отец	приехать	вещь	срочный	информация
сообщает источник	сообщил отцу	сообщить о своем приезде	сообщил важные вещи	сообщить срочно	сообщить информацию

- (8) Семантический скетч для глагола "извещать:TO_INFORM":

Addressee	Theme	Agent	Time	Object_Situation	Agent_Metaphoric
полиция	прибыть	Григоров	современный	пожечь	объявление
известить полицию	известно о прибытии	известил григоров	известно современно	известил и пожег	известно объявление
граф	решать	доброжелатель	непосредственно	передать суду	табличка
известил граф николай герард фонке	известил о мере решения	известил редакционные доброжелатели	известил непосредственно	известил и предан суду	известил табличка на забора
родственник	кончина	генерал-губернатор	заранее	принимать	бейдж
известить родственников	известили о кончине владыки	известил генерал-губернатор граф ростокин	известить заранее	известил что статья принята	известил бейдж на директорском пиджаке
родные	приехать	Измайлов	предварительный	причаливать	пометочка
известить родных	известить о вашем приезде	известил измайлов со вздохом облегчения	известить предварительно	известил что причаливает	известил пометочка другой рукой
общественность	рожать	газета	нескоро	миновать	Известия
известить и общественность	известить о рождении наследника престола	известили газеты	известить нескоро	известил что опасность благополучно миновала	известить известия
Шлейер	визит	яхта	заблаговременный	подлежать	мэйл
известил шлейера	известить о визите бога	известил отца наша	известили заблаговременно	известить и тоже подлечь назначено	известил второй мэйл

Как видно, ГП Addressee имеет разное синтаксическое выражение при данных глаголах и соответствует прямому объекту при «извещать» и дативу — при «сообщать». Однако это не мешает продемонстрировать общность семантической модели данных глаголов в отношении валентности адресата.

5.6. Семантические скетчи для семантических классов

Выше мы рассматривали примеры скетчей для отдельных слов. Тем не менее, интересным может быть также создание скетчей для целых семантических классов, что позволило бы сравнивать разные лексемы одного семантического поля (или, в терминах нашего формализма, разные лексические классы одного СК).

В качестве примера приведем фрагмент скетча для класса TO_COMMIT — лексической функции, где многие глаголы имеют весьма ограниченную сочетаемость:

Object_Situation	Object	Agent	Time	Agent_Metaphoric	Modality
роль сыграли свою роль	это сделать это	человек сделал он	год делали в прошлом году	это произвело это	правильный сделал правильно
впечатление произвести впечатление	вопрос задавать вопросы	автор делает автор	жить сделал в жизни	что_Interrog сделал я	так сделал так
участвовать приняли участие	решение принято решение	власти делает власть	сейчас делать сейчас	книга производит книга	действительно сделал действительно
шагать сделал шаг	вывод делать выводы	поэт делает поэт	тогда делать тогда	оно сделал он	конечно сделал конечно
дело сделано дело	что_Interrog делать что	отец делал он	время сделал в свое время	жить дала жизнь	возможно сыграло возможно
должное отдать должное	оно взять их	писатель делает писатель	теперь делать теперь	рука делают руки	может_быть сделал может быть

Наиболее показательна сочетаемость с объектными позициями, где проявляется лексикализованность сочетаемости рассматриваемых глаголов (*принять участие vs задать вопрос vs сыграть роль* и под.).

5.7. Другие возможности

Помимо перечисленного, в предлагаемом инструменте реализованы следующие возможности:

- ограничение части речи и падежа ядра зависимой составляющей;
- ограничение части речи и падежа ядра родительской составляющей;
- ограничение на количество упоминаний сочетания в корпусе — при желании можно включать в скетчи только зависимые, встретившиеся больше n раз.

Кроме того, при создании семантических скетчей в качестве зависимых модель позволяет выводить как фрагменты размечаемых предложений (либо равные составляющим, либо ядра составляющих с необходимыми грамматическими характеристиками — в релевантных предложно-падежных формах, например), так и ядра семантических классов, лексических классов или лексем.

6. Анализ результатов

Новизной настоящего исследования является эксперимент по построению семантических скетчей на значительном корпусном материале.

Семантические скетчи позволяют получать сравнительные портреты слов из одного семантического поля независимо от разницы в их синтаксических моделях, а также получать данные о сочетаемости целых семантических классов и учитывать конкретные значения рассматриваемых слов.

Реализация на основе разметки Comreno показала вполне приемлемые результаты, хотя некоторой проблемой является большое количество типов глубинных позиций — более 300. Эту проблему еще предстоит решить при включении механизма семантических скетчей в функционал ГИКРЯ (мы надеемся на фидбэк от пользователей корпуса).

7. Заключение и планы на будущее

В настоящий момент проводится обучение SLING-подобных технологий на корпусе RuSemSketches. Стоит задача попробовать применить обученную нейросеть для семантической разметки всего корпуса ГИКРЯ (при этом скетчевая разметка все же проще, чем полная семантическая разметка, что дает надежду получить приемлемые для пользователей результаты). Мы рассчитываем, что на основе RuSemSketches удастся провести shared task по семантической разметке на следующем Диалоге.

В ближайшее время планируется включить семантические скетчи на основе разметки Comreno в функциональность новой версии ГИКРЯ. Существенной особенностью скетчей в ГИКРЯ является их дифференциальность: возможность сравнения результатов с учетом всех доступных в ГИКРЯ метатекстовых признаков. Это позволит в ближайшей перспективе проводить исследования по дифференциальной лексической семантике.

8. Благодарности

Данная работа выполнена совместно исследовательской лабораторией АВВУ Lab МФТИ (проект ГИКРЯ) и отделом Advanced Research Development компании АВВУ. Мы благодарим всех коллег, и в особенности Константина Дружкина и Евгения Инденбома за полезную критику и помощь в проведении исследования.

Литература

1. *Abend O., Rappoport A.* (2013), UCCA. A semantics-based grammatical annotation scheme, Proceedings of the 10th International Conference on Computational Semantics, Potsdam, Germany, pp. 1–12.
2. *Anisimovich K., Druzhkin K., Minlos F., Petrova M., Selegey V., and Zuev K.* (2012), Syntactic and semantic parser based on АBBY Comreno linguistic technologies. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", vol. 11, pp. 91–103.

3. *Banarescu L., Bonial C., Cai S., Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn P., Palmer M., Schneider N.* (2013), Abstract Meaning Representation for sembanking, Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria, pp. 178–186.
4. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.*, (2013), [Big and diverse is beautiful: A large corpus of Russian to study linguistic variation](#), Web as Corpus Workshop (WAC-8).
5. *Belikov V., Sharoff S., Kopylov N. et al.* (2015), Corpus with automatically removed morphological ambiguity: methodology of linguistic research [Korpus s avtomaticheski snyatoi morfologicheskoi neodnoznachnostju: K metodike lingvisticheskikh issledovanij], Computational Linguistics and Intellectual Technologies [Komp'iuternaia lingvistika i intellektualnye tekhnologii], Vol. 14, № 1. pp. 84–95.
6. *Hajic J., Hajicova E., Panevova J., Sgall P., Bojar O., Cinkova S., Fucikova E., Mikulova M., Pajas P., Popelka J., Semecky J., Sindlerova J., Stepanek J., Toman J., Uresova Z., Zabokrtsky Z.* (2012), Announcing Prague Czech-English Dependency Treebank 2.0., Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, pp. 3153–3160.
7. *Ivanova A., Oepen S., Øvrelid L., Flickinger D.* (2012), Who did what to whom? A contrastive study of syntacto-semantic dependencies, Proceedings of the 6th Linguistic Annotation Workshop, Jeju, Republic of Korea, pp. 2–11.
8. *Kilgarriff A., Rychlý P., Smrž P., Tugwell D.* (2004), The sketch engine, Information Technology.
9. *Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V.* (2014), The Sketch Engine: ten years on, Lexicography, 1: 7–36.
10. *Oepen S., Lønning J. T.* (2006), Discriminant-based MRS banking, Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, pp. 1250–1255.
11. *Oepen S. et al.* (2019), MRP 2019: Cross-Framework Meaning Representation Parsing, CoNLL 2019.
12. *Petrova M. A.* (2014), The Compreno Semantic Model: The Universality Problem, International Journal of Lexicography, Vol. 27, Issue 2, pp. 105–129.
13. *Ringgaard M., Gupta R., Pereira F. C. N.* (2017), SLING: A framework for frame semantic parsing.
14. *Rychlý P.* (2008), A Lexicographer-Friendly Association Score, Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN, Brno, pp. 6–9.

Интернет-ресурсы

15. <https://www.abbyy.com/ru-ru/science/technologies/compreno/>
16. <http://www.ruscorpora.ru/new/corpora-sem.html>
17. https://tatianashavrina.github.io/taiga_site/
18. <http://www.webcorpora.ru/>

Приложение

Семантические классы

Семантический класс — это универсальная (разделяемая всеми языками системы) единица лексического описания. Включает базовое лексическое значение конкретного языка со всеми его продуктивными грамматическими и семантическими дериватами.

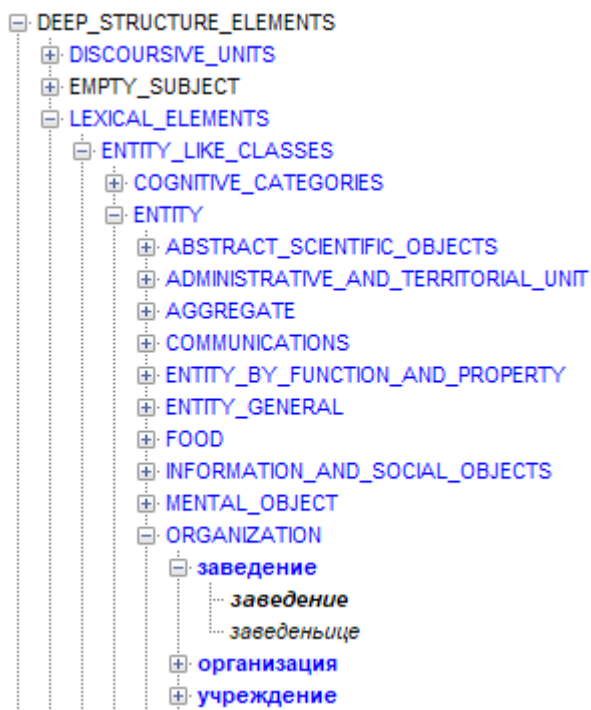


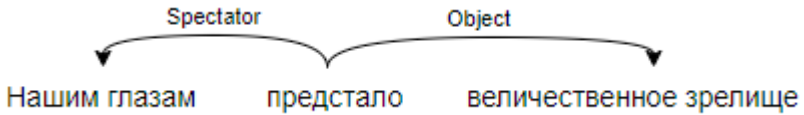
Рис. 2. Общий вид семантической иерархии

Каждому классу соответствует набор ограничений, накладываемых на семантические отношения, которые представители класса могут заполнять.

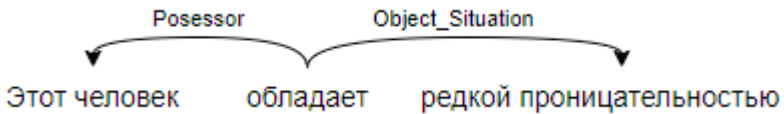
Семантические отношения

Ниже приведены краткие описания некоторых глубинных позиций и примеры размеченных фрагментов текста.

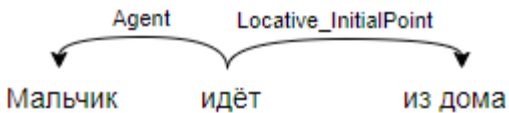
- **Object** — объект действия в широком смысле.
- **Spectator** — Свидетель, зритель — лицо, при котором происходит какое-то действие, но которое не вовлечено в это действие.



- **Object_Situation** — ситуационный объект.
- **Possessor** — посессор, обладатель.



- **Agent** — агенс (в широком смысле).
- **Locative_InitialPoint** — исходный пункт, начальная точка.



ANNOTATING AND EXPLORING CODE-SWITCHING IN FOUR CORPORA OF MINORITY LANGUAGES OF RUSSIA¹

Dyachkov V. V. (hyppocentaurus@mail.ru)

Institute of Linguistics, Russian Academy of Sciences,
Moscow, Russia

Khomchenkova I. A. (irina.khomchenkova@yandex.ru)

Russian Language Institute, Russian Academy of Sciences &
Lomonosov Moscow State University, Moscow, Russia

Pleshak P. S. (polinapleshak@yandex.ru)

University of Maryland, College Park, USA

Stoynova N. M. (stoynova@yandex.ru)

Russian Language Institute, Russian Academy of Sciences &
NRU Higher School of Economics, Moscow, Russia

This paper describes code-switching with Russian in four spoken corpora of minority languages of Russia: two Uralic ones (Hill Mari and Moksha) and two Tungusic ones (Nanai and Ulch). All narrators are bilinguals, fluent both in the indigenous language (IL) and in Russian; all the corpora are comparable in size and genres (small field collections of spontaneous oral texts, produced under the instruction to speak IL); the languages are comparable in structural (dis)similarity with Russian. The only difference concerns language dominance and the degree of language shift across the communities. The aim of the paper is to capture how the degree of language shift influences the strategy of code-switching attested in each of the corpora using a minimal additional annotation of code-switching. We added to each corpus a uniform annotation of code-switching of two types: first, a simple semi-automatic word-by-word language annotation (IL vs. Russian), second, a manual annotation of structural code-switching types (for smaller sub-corpora). We compared several macro-parameters of code-switching by applying some existing simple measures of code-switching to the data of annotation 1. Then we compared the rates of different structural types of code-switching, basing on annotation 2. The results of the study, on the one hand, verify and enhance the existing generalizations on how language shift influences code-switching strategies, on the other hand, they show that even a very simple annotation of code-switching integrated to an existing field records collection appears to be very informative in code-switching studies.

Keywords: corpus linguistics, quantitative linguistic studies, language shift, code switching, code-switching metrics, Uralic languages, Tungusic languages

DOI: 10.28995/2075-7182-2020-19-228-240

¹ Supported by RFBR grant N° 18-312-00155.

РАЗМЕТКА И ИССЛЕДОВАНИЕ ПЕРЕКЛЮЧЕНИЯ КОДОВ В ЧЕТЫРЕХ КОРПУСАХ МАЛЫХ ЯЗЫКОВ РОССИИ²

Дьячков В. В. (hyppocentaurus@mail.ru)

ИЯз РАН, Москва, Россия

Хомченкова И. А. (irina.khomchenkova@yandex.ru)

ИРЯ РАН & МГУ им. М. В. Ломоносова, Москва, Россия

Плешак П. С. (polinapleshak@yandex.ru)

Университет Мэриленда, Колледж Парк, США

Стойнова Н. М. (stoynova@yandex.ru)

ИРЯ РАН & НИУ ВШЭ, Москва, Россия

1. Introduction

The aim of this paper is to show how frequency and structural types of code-switching correlate with sociolinguistic parameters, using quantitative data of the spoken corpora of several minority languages of Russia.

The use of two or more languages within one conversation or even within one utterance is known as the phenomenon of code-switching (CS). In this project, we understand it quite broadly (cf. [Treffers-Daller 1991]; [Myers-Scotton 1992]). We take into consideration all types of inter-clausal switching (1), as well as intra-clausal switching of constituents of different length (*i* ‘and’, *ixnjuju familiju* ‘their last name’ in (2)) and nonce borrowings from Russian, which often bear morphological affixes of the main language of the text (*veləs’ipeca* [bike.R.IN] in (1), *sestra-ni* [sister.R-3SG] in (2)).

(1) *veləs’ipeca ar-n’ə-s’-t’.*

bike.R.IN run-IPFV-PST.3-PL

oj, togda velosiped-ov ne bylo

oh.R then.R bike-PL.GEN.R NEG.R be.PST.3SG.N.R

‘We were riding the bicycles. Oh, there were no bicycles that time’. (baa, Moksha)

(2) *i ti sestra-ni ti aldač-i bi-či-n*

and.R that sister.R-3SG so tell-PRS be-PST-3SG

ixnjuju familiju

their.SG.F.ACC.R last.name.SG.ACC.R

‘And his sister mentioned their last name.’ (aid, Ulch)

Four languages of Russia were chosen for the research: Moksha, Hill Mari (Uralic); Nanai, and Ulch (Tungusic). They belong to different language families, but their

² При поддержке гранта РФФИ № 18-312-00155.

structural (dis)similarity with Russian, which is relevant for our study, is more or less equal. For this study, we used the corpora that had been created by larger teams (including the authors) in the field during documentation projects on the corresponding languages, and the aim of the narrator was to tell a story in the indigenous language (IL). The text collections are comparable in sizes and genres (see below). For all the narrators, IL is the first language, or it was acquired simultaneously with Russian. All the speakers are highly proficient in Russian. They are also proficient in IL enough to tell a spontaneous story. So, although the degree of speakers' proficiency differs a lot on the level of the whole speech communities (see below), it is comparable for our text samples.

The only crucial difference among our datasets is the current use of the languages, which can be interpreted in terms of language dominance and language shift. The Hill Mari speakers use IL in their everyday communication at least as frequently as Russian. The Ulch narrators use IL much more restrictively than Russian or do not use it at all. In the Moksha and Nanai samples the situation varies across speakers. On the community level, across the speakers of Hill Mari a stable balanced bilingualism takes place, while the Moksha speakers, the Nanais, and especially the Ulchas are undergoing a progressing language shift to the dominant Russian language [Koryakov & Kholodilova 2018]; [Kalinina & Oskolskaya 2016]; [Gerasimova 2002]; [Sumbatova & Gusev 2016]. This can be represented as a hierarchy of language shift, cf. (3); see also **Table 1**:

(3) (language shift) Ulch > Nanai > Moksha >> Hill Mari (no language shift)

Table 1. Sociolinguistic information

language	N of speakers [Census 2010]	% of the ethnic group
Ulch	154	6%
Nanai	1,347	11%
Moksha	2,025	43%
Hill Mari	23,062	98%

Basing mostly on generalizations made in [Benthalia & Davies 1992], [Backus 1996], and [Muysken 2000: 227–228; 247–248], we have the following expectations on possible correlations between language dominance and inter-generation shift³, on the one hand, and structural types of CS, on the other hand:

- inter-clausal switches are more frequent in balanced bilinguals, than in the situation of dominance asymmetry;
- word-internal switches are more frequent in the situation of dominance asymmetry, than in balanced bilinguals;

³ Note, however, that they discuss mostly inter-generation shift within local communities, while language shift (i.e. the loss of language within the whole language community) is much less studied from this point of view.

- in the situation of language shift, the number of more syntactically integrated constituents (insertions in terms of [Muysken 2000]) increases, while the number of less integrated constituents (alternations in terms of [Muysken 2000]) decreases;
- in the situation of language shift the number of non-constituent switches (which also belong to alternations in terms of [Muysken 2000]) decreases.

Our general hypothesis was that, according to hierarchy (3), the main difference in CS strategies would be between the Hill Mari and Ulch corpora, as they represent the opposite sociolinguistic situations, with Nanai and Moksha corpora in between, having an intermediate stage of the language shift. To reveal specific properties of CS that vary across the text collections under discussion, we annotated each word for the language (IL vs. Russian) and added a specific annotation of structural types of CS (Section 2). Then, for each corpus we conducted calculations, based on language annotation, using some existing metrics developed for corpus-based studies on CS (Section 3). After that, for a smaller part of the text collection, we conducted more precise calculations, based on our annotation of structural types of CS (Section 4). Finally, we compared the results of the calculations for our four corpora, checked how they match to the language shift hierarchy, and related the correlations, attested in our data, to the existing observations (Section 5).

2. Annotation of code-switching types

All the text collections were annotated in ELAN using the same set of tiers and labels. The annotation tiers are the following: LANG, which indicates the language of each word (token), CS_TYPE, which indicates the syntactic type of the switched fragment (all tags are aligned to words). The annotated collections are available at: http://web-corpora.net/tsakorpus_russian_nonst/CS.html.

The LANG tier contains two tags: IL—indigenous language and Rus—Russian (including Russian words with IL-affixes). This tier was created semi-automatically, based on the script: the main transcription in our corpora is in Latin, while the majority of Russian fragments are in Cyrillic. So, the tag Rus was first assigned automatically to all words⁴ transcribed in Cyrillic, then some tags were changed or added manually.

The CS_TYPE tier contains syntactic tags, which were assigned manually to a smaller part of the text collection (see Table 2 for general information on the corpora). They are listed in Table 3.

⁴ The word was recognized simply as an item separated by space-bars. Intermediate cases, such as clitics, were treated according to the writing system, adopted in each particular corpus. This might create some discrepancies in our data, but they are rather minor.

Table 2. Corpora: sizes and types of annotation

	provided with LANG tags, texts (tokens)	provided with CS_TYPE tags, texts (tokens)	other features of the corpus
Nanai	167 (47,411)	52 (16,368)	synchronized with audio
Ulch	179 (47,509)	50 (11,334)	synchronized with audio
Moksha	53 (17,578)	53 (17,578)	glossed
Hill Mari	17 (15,895)	17 (15,895)	glossed

Table 3. CS-type tags

tag	description
adj(+) ⁵	adjectival phrase
adv(+)	adverbial phrase
conj(+)	conjunction phrase
dep	dependent clause
disc(+)	discourse marker
ideoph(+)	ideophone
interj(+)	interjection
morph	Russian stem with IL-affixes
morph_p	Russian multi-word phrase marked with IL-affixes
np(+)	noun phrase
nump(+)	numeral phrase
pp(+)	prepositional phrase
pred(+)	predicative word
s	sentence
v_rus ⁶	clause with Russian verb
voc(+)	vocative forms
vp(+)	verb phrase
other	other constituent types

Russian fragments that do not form any syntactic constituent are marked with corresponding tags separately (*conj*, *conj*, *pp* and *adj*) in (4).

⁵ “+” stands for multi-word constituents. In this case, the tag is assigned to the head.

⁶ The texts under consideration are positioned by narrators as texts in the corresponding indigenous language (IL) and not in Russian, and the total amount of Russian fragments is much smaller than those in the IL. However, a potential possibility to reveal the main language (“matrix”, ML) and the secondary one (“embedded”, EL) for each particular clause with intrasentential CS is a matter of theoretical discussion (cf. [Myers-Scotton 1993: 46–74]; [2002: 15–16; 58–69]; [Muysken 2000: 1–34]). Our technical solution is to mark Russian fragments as switched (i.e. to consider the IL as ML) in all mixed clauses, except for those with Russian finite verbs (which are much less numerous in the sample). The latter are treated separately and take a tag *v_rus* with no further annotation.

- (4) [no] [i] [do vojny] [molodaja] bi-či-ni=goa
 but.R and.R before.R war.GEN.R young.SG.F.R be-PST-3SG=PTCL
 ‘But before the war she also was young.’ (itg, Nanai)

3. Metrics based on the word-by-word language annotation

The general information on CS, which allows us to compare the four text collections, was obtained from the word-by-word language annotation, available for the whole corpora (cf. Section 2). To characterize CS patterns, we used the existing metrics, proposed for corpus-based studies on CS and summarized in [Guzmán et al. 2016], [2017a,b]. Some of them are based on the ratio of L1-words and L2-words, some others are based rather on the ratio of L1-spans and L2-spans (where a L1-span is a word sequence in L1 bounded between L2-words). The general information on these metrics is given in Table 4.

Table 4. CS metrics based on word-by-word language annotation

Metric	Description	Formula	[from...; to...]	Reference
Multi-lingual index	measures how “bilingual” the text is: the (in)equality of the distribution between L1 and L2	$M-I = \frac{1 - \sum p_i^2}{\sum p_i^2}$	[0; 1] [all words in L1; L1 and L2 in equal proportions]	[Barnett et al. 2000], [Gardner-Chloros et al. 2007], [Guzmán et al. 2016] [2017a]
Integration Index	measures how “bilingual” the text is: the probability of L1 vs. L2 within the text	$I-I = \frac{1}{n-1} \sum_{1 \leq i < j \leq n} S(i, j)$	[0; 1] [L1, L1, L1...; L2, L1, L2,...]	[Gambäck & Das 2014, 2016]; [Guzmán et al. 2016], [2017a]
Burstiness	measures how (non)-random switches are: the regularity of switching spans	$Burstiness = \frac{(\sigma_s - m_s)}{(\sigma_s + m_s)}$	[-1; 1] [regular heart-beat-like switching; irregular switching]	[Goh & Barabási 2008]; [Guzmán et al. 2017a]
Language Span Entropy	measures how predictable language spans are: how many bits of information are needed to describe the distribution of language spans	$Span\ Entropy = - \sum_{i=1}^M p_i \log_2 p_i$	[0; log2(M), M = the number of possible span states] [all spans are of equal length; spans are of a different length]	[Guzmán et al. 2017b]
Memory	measures how (non)-random switches are: whether the length of L1-span correlates with the length of the preceding L2-span	$Memory = \frac{1}{n_r - 1} \sum_{i=1}^{n_r-1} \frac{(r_i - m_1)(r_{i+1} - m_2)}{\sigma_1 \sigma_2}$	[-1; 1] [short L1 spans are preceded by long L2 spans; short L1 spans are preceded by short L2 spans]	[Goh & Barabási 2008]; [Guzmán et al. 2017a]

Figure 1 contains the values of these measures for all our corpora. While calculating, initial and final Russian fragments were omitted.

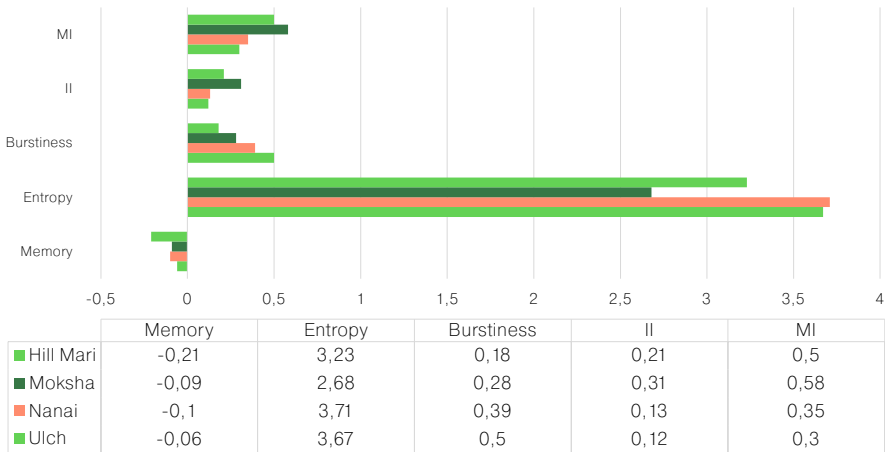


Fig. 1. Measures of CS: the data of the four corpora

The Multilingual Index (MI) is a word-count-based measure that quantifies the inequality of the distribution of language tags in a corpus. According to it, the Ulch corpus is the most monolingual, while the Moksha corpus is the most bilingual (Ulch < Nanai < Hill Mari < Moksha). Integration Index (II) is a metric that describes the probability of switching within a text. Languages with the same MI can have different number of switches (compare [IL, IL, Rus, Rus]₁ and [IL, Rus, IL, Rus]₂ with both MI=1, but II₁=0,(3) and II₂=1). The IIs of our four corpora correspond to their MIs and form the same hierarchy.

We also calculated metrics reflecting the distribution of CS across the corpus using language spans—the distance between switch points, i.e. the length of monolingual discourse. The Burstiness measures whether CS has a periodic character or occurs in bursts, i.e. how predictable switches are. All our corpora have unpredictable patterns of switching with the following hierarchy: Ulch > Nanai > Moksha > Hill Mari. The switching patterns in Ulch and Nanai are more unpredictable, while that in Hill Mari is the most predictable. In order to take into account the time ordering of the language spans, we calculated the Memory Index, which shows to which extent the length of language spans influences the length of following spans. The hierarchy of corpora is exactly the same as for the Burstiness. All language spans are rather unpredictable, but Hill Mari language spans are more negatively correlated, while Ulch language spans are more positively correlated. The Span entropy returns how many bits of information are needed to describe the distribution of the language spans. The hierarchy is a bit different: Nanai ≥ Ulch > Hill Mari > Moksha, so it does not correlate with the Burstiness and Memory Index, but rather similar to those for MI and II (although with the opposite direction).

Thus, two out of five measures, i.e. those operating with spans, give the results, more or less correlating with the language shift hierarchy in (3), while three others show different results, which, however, are all similar to each other.

4. Structural types of code-switching

For the manually annotated sub-corpora, we compared frequencies of different structural types of switched fragments and frequencies of fragments of different sizes. The frequency distribution for switched fragments' sizes (morpheme vs. one-word vs. multi-word vs. clause⁷) is given in **Figure 2**.

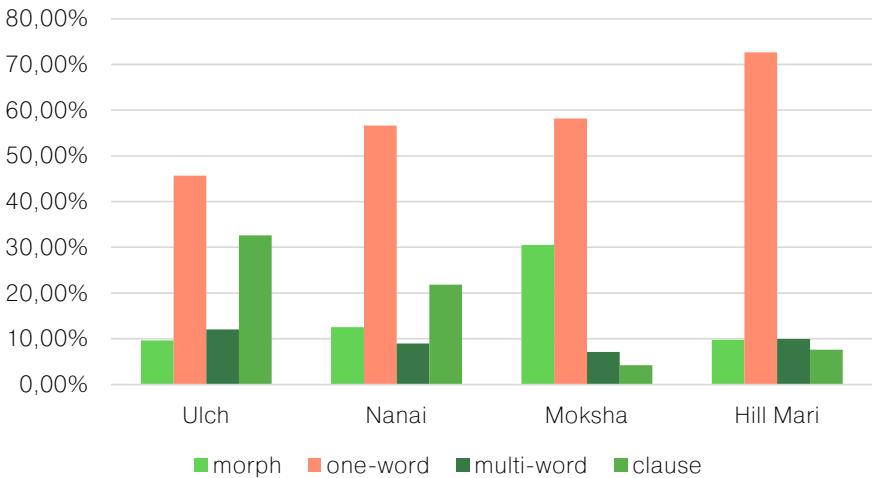


Fig. 2. Switched fragments' sizes

The rates of multi-word switches are comparable in all the collections and relatively low. The rates of other types of switches vary across the collections. Word-internal switches are much more frequent in Moksha than in all other languages. The percentage of clausal and one-word switches correlates with the language shift hierarchy (3). One-word switches form the most frequent type in all the samples, and their rate is lower in languages more affected by language shift (although the difference is quite modest). In contrast, clausal switches are much more frequent in languages more affected by language shift.

For switched constituents (excluding Russian stems with IL-affixes and Russian sentences), we calculated the frequency distribution of different syntactic types, see Figure 3. Only frequent types (> 1%) were included.

The most frequent types of switched constituents in all the corpora are *np(+)* (noun phrases), *disc(+)* (discourse markers), *conj(+)* (conjunctions) and *adj(+)* (adjectives). We expect the ratios to reflect the language shift hierarchy, introduced in (3). Across the frequent types, the correlation is attested for NPs and for discourse markers. Switched NPs are especially frequent in Ulch and the least frequent in Hill Mari; in opposite, discourse markers are the most frequent in Hill Mari and the least frequent in Ulch.

⁷ Multi-clause switched fragments were not treated separately. Each of them was counted as several independent switched clauses. The same is true for multi-word switched fragments that do not form a syntactic constituent: they were counted as several independent constituents.

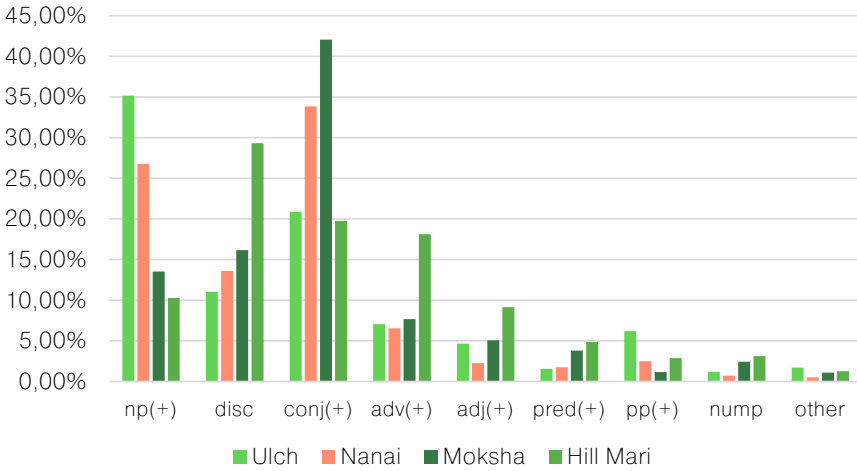


Fig. 3. Syntactic types of intra-sentential CS

This effect seems to be connected with the degree of syntactic integration. Discourse markers are elements which are not integrated into the syntactic structure, are uninflected forms and do not bear any overt markers of syntactic dependency, unlike NPs that are highly integrated into the clausal structure. We calculated the total ratio for non-integrated and integrated elements where we treated constituents of types *np(+)*, *pp(+)*, *num p(+)* as integrated and *disc(+)*, *pred(+)*, *interj*, *adv(+)*, *voc(+)*, and *adj(+)* as non-integrated⁸. For this total ratio, the correlation with the sociolinguistic hierarchy appears to be even clearer, see **Figure 4**.

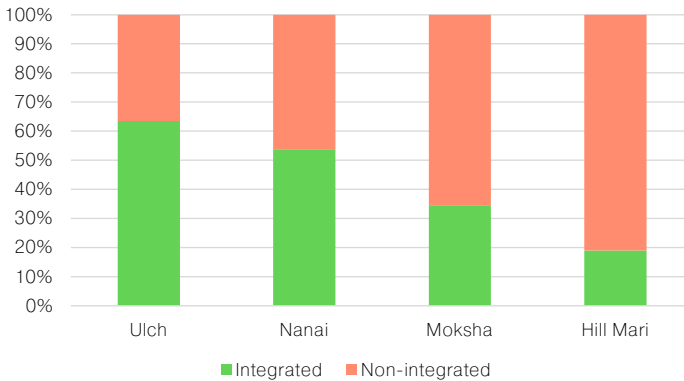


Fig. 4. Ratio of integrated and non-integrated elements

Integrated switched elements are typical of the language shift situation (Ulch), while non-integrated ones are typical of the situation of stable balanced bilingualism (Hill Mari).

⁸ Conjunctions, which do not form part either of the two types of elements, were excluded.

5. Results and discussion

The crucial difference between the text collections under discussion concerns language dominance, i.e. the degree of language shift attested in the community. The hypothesis was that structural differences in CS would follow a hierarchy of languages that reflects the difference in their sociolinguistic status:

(5) (=3) Ulch > Nanai > Moksha >> Hill Mari

We applied to our data several simple measures of CS, based on the distribution of Russian words (Multilingual Index, Integration Index) and word-sequences (spans) across the text (Burstiness, Entropy, and Memory). Then, we checked whether they are interpretable in terms of language shift hierarchy. The measures based on switched spans appear to correlate with this hierarchy better, than those based on switched words.

We also checked some more fine-grained parameters of CS: the rate of clausal switches, the rate of word-internal switches, and the rate of different syntactic types of switched constituents. Not all of the attested asymmetries between the corpora exactly correlate with the language shift hierarchy in (5), but they still can be explained by sociolinguistic factors. Interestingly, our data do not confirm the previous observations on CS types and language shift.

1) *Inter-clausal switches*: Ulch > Nanai > Moksha > Hill Mari. The rate of clausal switches in the corpora correlates well with the language shift hierarchy: the more progressed language shift is, the more frequently clausal switches occur. At the same time, [Bentahila and Davies 1992] report the opposite tendency for code-switching between Moroccan Arabic and French. This contradiction can result from the deliberate specific of our texts. Being instructed to speak IL, speakers with dominant Russian (the Ulchas) try to speak IL, but insert Russian clauses in cases where they have difficulties with IL. Therefore, inter-clausal switches have to be rather frequent in their speech. However, in spontaneous communication, the same speakers would speak mostly Russian and include only short IL-fragments in their Russian speech, i.e. in fact use more intra-clausal switches. In contrast, balanced bilinguals (the Hill Mari speakers) do not need to use Russian sentences more often than sentences in IL, since they are equally prominent in both languages.

2) *Word-internal switches*: Moksha >> Nanai \approx Ulch \approx Hill Mari. The same apparent contradiction takes place for word-internal switches: their distribution corresponds neither to the hierarchy in (5), nor to the previous observations. Word-internal switches are mostly connected to cultural vocabulary (including “soviet realities”). According to [Bentahila and Davies 1992], they are frequent among the speakers with the dominant Arabic using French cultural words. In contrast, in our data, both speakers with dominant Russian (Ulch, Nanai) and balanced bilinguals (Hill Mari) use word-internal (Russian) switches with comparable frequency. In our corpora (in contrast to that of Bentahila and Davies), morphologically integrated cultural words come from the dominant language (Russian). For balanced bilinguals (Hill Mari) cultural words seem to be the main source of word-internal switches. In situations of progressed language shift (Nanai and Ulch), basic words are involved in CS as well as cultural lexicon, so the expected number of word-internal switches might be higher than it is. However, in the situation of progressed language shift, speakers are not very creative in IL-morphology and prefer to use non-integrated Russian constituents instead of morphologically-integrated Russian

stems, so the rate of word-internal shifts is as low as for balanced bilinguals. In contrast, on the intermediate stage of language shift the demand for Russian lexemes is equally high, but speakers feel free in integrating them into IL. This is the case of Moksha.

3) *Syntactic integration (constituent type)*: Ulch > Nanai > Moksha > Hill Mari. For switched intra-clausal constituents, the degree of syntactic integration correlates with the language shift hierarchy: in the situation of language shift syntactically integrated constituents (e.g. NPs) tend to be switched, while balanced bilinguals more frequently switch non-integrated constituents (e.g. discourse markers). These two strategies of CS correlate with [Muysken's 2000] *insertion* and *alternation* respectively. According to Muysken, insertions are single constituents, content rather than functional words and complements rather than adjuncts. This is exactly what opposes NPs and PPs (counted as “integrated”) to discourse particles, adverbs, interjections etc. (counted as “non-integrated”). Alternation, on the contrary, requires less integration into syntax, and is mostly represented by discourse particles, adverbs and other items, counted in our study as “non-integrated”. Therefore, in Muysken's terms, languages more affected by language shift prefer insertion, while less affected ones prefer alternation.

Muysken himself [Muysken 2000: 227–228; 247–248] makes the general prediction, that in the process of language shift the rate of insertions would become higher and the rate of alternations lower (see also [Backus 1996] for the same claim). There is a nuance that has to be clarified. Making his prediction, [Muysken 2000] considers as insertions not only syntactically-integrated constituents, but also word-internal switches. However, we have already shown that word-internal switches are typical not for the progressed language shift situation (in contrast to syntactically integrated switches) and not for balanced bilinguals (as follows from Muysken's generalization), but for the early stage of language shift. The extensive use of word-internal and syntactically integrated switches is caused by the same reason, i.e. the need for Russian nouns. Ulch and Nanai speakers (progressed language shift) use switched NPs largely, since they are restricted in morphological adaptation of Russian nouns and syntactic integration is the only option in this case (see above). This leads to the high rate of syntactically integrated switched constituents. Moksha speakers (the early stage of language shift) mark Russian nouns with IL-affixes. This leads to the high rate of word-internal switches. In contrast, Hill Mari balanced bilinguals widely use IL nouns instead of Russian ones, so they show both the lowest rate of syntactically integrated and word-internal switches. Thus, in our data, the correlation between the degree of language shift and the rate of insertions, observed by Muysken, works differently, and an additional parameter of morphological vs. syntactic integration within the insertion should be considered.

Summing up, we can say that a very simple annotation, that contains only tags for languages and constituent types, can indeed shed light on correlation between sociolinguistic situation in the community and CS types. The metrics provide numeric data which can be projected on a hierarchy of language shift. Strong oppositions between situations with language shift and without it hold, however, they can work in the opposite direction as well. The explanation of such a variation is an object for a future research. A possible parameter that has to be considered is (non)-equivalence of the main language of the clause/text and the dominant language.

6. Abbreviations

3—3rd person, ACC—accusative, F—feminine, GEN—genitive, IN—inessive, IPFV—imperfective, N—neuter, NEG—negative, NPST—non-past, PL—plural, PRS—present, PST—past, PTCL—particle, R—Russian, SG—singular.

References

1. *Backus A.* (1996), *Two in One: Bilingual Speech of Turkish Immigrants in the Netherlands*, Tilburg University Press, Tilburg.
2. *Barnett R., Codo E., Eppler E., Forcadell M., Gardner-Chloros P., van Hout R., Moyer M., Torras M. C., Turell M. T., Sebba M., Starren M., Wensing S.* (2000), *The LIDES Coding Manual: A document for preparing and analyzing language interaction data*, Version 1.1–July, 1999, *International Journal of Bilingualism*, 4(2), pp. 131–132.
3. *Bentahila A., Davies E. D.* (1992), *Code-switching and language dominance*, Harris, R. J. (Ed.) *Cognitive processing in bilinguals*, Elsevier, Amsterdam, pp. 443–458.
4. *Gambäck B., Das A.* (2014), *On measuring the complexity of code-mixing*, *Proceedings of the 11th International Conference on Natural Language Processing*, Goa, India, pp. 1–7.
5. *Gambäck B., Das A.* (2016), *Comparing the level of code-switching in corpora*, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1850–1855.
6. *Gardner-Chloros P., Moyer M., Sebba M.* (2007), *Coding and Analysing Multilingual Data: The LIDES Project, Creating and Digitizing Language Corpora*, pp. 91–120.
7. *Gerasimova A. N.* (2002), *Nanai and Ulch in Russia: a comparative characteristics of the sociolinguistic situation [Nanajskij i ulčskij jazyki v Rossii: sravnitel'naja harakteristika sociolingvističeskoj situacii]*, *Jazyki Korennyh narodov Sibiri [Languages of Indigenous Peoples of Siberia]*, 12, pp. 246–257.
8. *Guzmán G. A., Serigos J., Bullock B., Toribio A. J.* (2016), *Simple tools for exploring variations in code-switching for linguists*, *EMNLP-2016: Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pp. 12–20.
9. *Guzmán G. A., Serigos J., Bullock B., Toribio A. J.* (2017a), *Moving code-switching research toward more empirically grounded methods*, *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017)*, Bloomington, IN, USA, pp. 1–9.
10. *Guzmán G. A., Serigos J., Bullock B., Toribio A. J.* (2017b), *Metrics for modeling code-switching across corpora*, *Proc. Interspeech 2017*, pp. 67–71.
11. *Goh K. I., Barabási A. L.* (2008), *Burstiness and memory in complex systems*, *EPL (Europhysics Letters)*, 81(4): 48002.
12. *Kalinina E. Ju., Oskolskaya S. A.* (2016), *Nanai [Nanajskij jazyk]*, *Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ėnciklopedija]*, Azbukovnik, Moscow, pp. 293–296.

13. *Koryakov Ju. B., Kholodilova M. A.* (2018), General information about the Moksha language and the idiom [Obsčie svedenija o mokshanskom jazyke i issleduemom govore], Toldova S. Ju., Kholodilova M. A. (Eds.), *Elementy mokshanskogo jazyka v tipologicheskom osvesčenii*, Buki Vedi, Moscow, pp. 6–18.
14. *Muysken P.* (2000), *Bilingual speech: A typology of code-mixing*, Cambridge University Press, Cambridge/New York.
15. *Myers-Scotton C.* (1992), Comparing codeswitching and borrowing, *Journal of Multilingual & Multicultural Development*, 13(1–2), pp. 19–39.
16. *Myers-Scotton C.* (1993), *Duelling languages: Grammatical structure in codeswitching*, Oxford University Press, Oxford/New York.
17. *Myers-Scotton C.* (2002), *Contact linguistics: Bilingual encounters and grammatical outcomes*, Oxford University Press, Oxford/New York.
18. *Sumbatova N. R., Gusev V. Ju.* (2016), Ulch [Ul'čskij jazyk], *Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ėnciklopedija]*, Azbukovnik, Moscow, pp. 513–515.
19. *Treffers-Daller, J.* (1991), Towards a uniform approach to codeswitching and borrowing. Papers for the workshop on constraints, conditions and models, *European Science Foundation, Strasbourg*, pp. 259–279.

РУССКИЙ КОНСТРУКТИКОН: НОВЫЙ ЛИНГВИСТИЧЕСКИЙ РЕСУРС, ЕГО УСТРОЙСТВО И СПЕЦИФИКА

Эндресен А. А. (anna.endresen@uit.no)

Университет Тромсё — Норвежский арктический университет

Жукова В. А. (valentina.zh96@gmail.com)

Национальный исследовательский университет
«Высшая Школы Экономики»

Мордашова Д. Д. (mordashova.d@yandex.ru)

Институт языкознания РАН, МГУ им. М. В. Ломоносова

Рахилина Е. В. (rakhilina@gmail.com),

Ляшевская О. Н. (olesar@yandex.ru)

Национальный исследовательский университет
«Высшая Школы Экономики»;

Институт русского языка им. В. В. Виноградова РАН

DOI: 10.28995/2075-7182-2020-19-241-255

THE RUSSIAN CONSTRUCTICON: A NEW LINGUISTIC RESOURCE, ITS DESIGN AND KEY CHARACTERISTICS

Endresen A. A. (anna.endresen@uit.no)

UiT The Arctic University of Norway

Zhukova V. A. (valentina.zh96@gmail.com)

National Research University Higher School of Economics

Mordashova D. D. (mordashova.d@yandex.ru)

Institute of Linguistics (Russian Academy of Sciences),
Lomonosov Moscow State University

Rakhilina E. V. (rakhilina@gmail.com),

Lyashevskaya O. N. (olesar@yandex.ru)

National Research University Higher School of Economics;
Vinogradov Institute for Russian language (Russian Academy
of Sciences)

We present a new open-access electronic resource named the Russian Construction that offers a searchable database of Russian constructions accompanied by descriptions of their properties and illustrated with corpus examples. The project was carried out over the period 2016–2020 and at present contains an inventory of over 2200 multi-word constructions of Contemporary Standard Russian. We prioritize “partially schematic” constructions that lie between the two extremes of fully compositional syntactic sequences on the one hand and fully idiomatic (phraseological) expressions on the other hand. Constructions of this type are difficult to account for in terms of either lexicon or grammar alone, and are often underrepresented in reference works of Russian. A typical construction in our database contains a fixed part (anchor words) and an open slot that can be filled with a restricted set of lexemes. In this paper we first focus on key characteristics of this resource that make it different from existing constructions of other languages. Second, we describe how the new interface will be designed and how it will serve the needs of both linguists and L2 learners of Russian. In particular, we discuss various search possibilities relevant for different users and those parameters that are available for specifying the retrieval output. An example of an entry is given to show how the information about each construction is structured and presented. Third, we provide an overview of our multi-level semantic classification of constructions. We argue that our system of semantic and syntactic tags subdivides our items into meaningful classes and smaller groups and eventually facilitates the identification of constructional families and clusters. This methodology works well in turning the initial list of constructions as unrelated units into a structured network and makes it possible to refine and expand the collected inventory of constructions in a systematic way.

Keywords: Russian, constructions, semantics, syntax, language technology, electronic database, language pedagogy, typology

1. Введение

Русский Конструктикон представляет собой новый электронный лингвистический ресурс по конструкциям современного русского языка — базу данных, которая на сегодняшний день содержит более 2200 единиц. Примерами таких конструкций являются **NP-Nom так и не VP** (*Он так и не женился*)¹, **NP-Nom возьми и VP-Imper** (*А он возьми и купи новую машину!*)², **NP-Gen бы** (*Чаю бы!*)

¹ В Русском Конструктиконе принята система обозначений, с помощью которых записываются конструкции в базе: **NP-Nom** = именная группа с существительным в именительном падеже.

² Данная конструкция тесно связана с конструкцией **взять и VP** (*он взял и увез меня в деревню*) и рассматривается как ее подтип в статье [Stojnova 2007]. В Русском Конструктиконе это две разные конструкции. Такой подход позволяет объяснить различие употреблений типа *А ты возьми и купи машину!* (оба императива в императивном значении, семантика побуждения к действию, согласование предиката с подлежащим, конструкция **взять и VP**) и употреблений типа *А мы возьми и купи машину!* (псевдоимперативные формы, императив в неимперативном значении, семантика неожиданно совершенного действия, отсутствие согласования подлежащего и предиката по лицу и числу, конструкция **NP-Nom возьми и VP-Imper**).

и др. Каждая конструкция сопровождается толкованием, семантическими, синтаксическими и стилистическими пометами, а также примерами из Национального корпуса русского языка (НКРЯ; подробнее о структуре словарной статьи см. [Janda et al. 2018]). Ресурс является бесплатным и находится в открытом доступе³.

Настоящая статья посвящена лингвистическим характеристикам ресурса. В **разделе 2** мы рассмотрим отличительные особенности Русского Конструктикона, сопоставив его с существующими конструктиконами других языков. В **разделе 3** мы кратко опишем его новый интерфейс, над которым идет работа, после чего в **разделе 4** представим принципы семантической разметки конструкций. В заключении будет сформулировано значение проекта как для русистики, так и для *конструктикографии* в целом.

2. С NP-Ins не сравнить: Русский Конструктикон в ряду других конструктиконов

В рамках проекта *конструкция* понимается, в соответствии с Грамматикой Конструкций [Fillmore et al. 1988]; [Croft 2001]; [Goldberg 2006]; [Rakhilina 2010], как центральная единица языковой структуры. В последние годы стал активно употребляться термин *конструктикон*, обозначающий, с одной стороны, систему конструкций определенного языка, а с другой стороны, детальное описание этой системы.

Это второе, прикладное, понимание конструктикона было впервые введено Ч. Филлмором [Fillmore 2008]; [Fillmore et al. 2012], который инициировал разработку конструктикона для английского языка⁴ в дополнение к ресурсу FrameNet. Идея Филлмора довольно быстро была подхвачена и нашла применение в отношении других языков. В настоящий момент идет работа по созданию

³ Новый сайт Русского Конструктикона находится в разработке, его адрес будет указан на интернет-странице проекта (<https://site.uit.no/russian-constructicon/>). Более ранняя версия интерфейса доступна по адресу <https://spraakbanken.gu.se/karp/#?mode=konstruktikon-rus> и устроена по образцу шведского конструктикона. Работа над созданием Русского Конструктикона ведется более четырех лет (2016–2020) в рамках совместного проекта сотрудничества Университета Тромсе — Норвежского Арктического Университета (UiT) и Национального исследовательского университета «Высшая Школа Экономики» (НИУ ВШЭ) в Москве. В разработке ресурса участвуют Л. А. Янда, Т. Нессет, А. А. Эндресен (группа CLEAR, UiT); Е. В. Рахилина, О. Н. Ляшевская, В. А. Жукова (Школа лингвистики НИУ ВШЭ); Д. Д. Мордашова (МГУ им. М. В. Ломоносова, Институт языкознания РАН); Ф. Тайерс (Университет Индианы, США). Программное обеспечение и создание нового интерфейса осуществляет Р. Баст (Отдел цифровых исследований, UiT), ранее работа велась А. Г. Клезович (Школа лингвистики НИУ ВШЭ) и Ф. О. Сизовым (РГГУ). В проекте также принимают участие студенты НИУ ВШЭ и UiT.

⁴ The Berkeley FrameNet Constructicon:
<http://www1.icsi.berkeley.edu/~hsato/cxn00/21colorTag/index.html>.

конструктиконов для шведского⁵, японского⁶, бразильского португальского⁷, немецкого⁸ и русского. Разработка этих ресурсов вылилась в появление нового направления, получившего название *конструктикография* и объединяющего теоретические основы грамматики конструкций и методы лексикографии [Lyngfelt et al. 2018a]. Выявление типологически сходных и лингвоспецифичных конструкций в разных языках входит в задачи многоязычной конструктикографии [Voas et al. 2019].

2.1. Объем

Объем Русского Конструктикона пока значительно превышает размеры других аналогичных ресурсов. На сегодняшний день, база данных Русского Конструктикона включает более 2200 конструкций, из которых первые 600 предварительно описаны. Ведется активная работа по пополнению базы новыми конструкциями. Для сравнения, шведский конструктикон, самый крупный из имеющихся, содержит описание около 400 конструкций, а конструктикон Беркли для английского языка включает 73 конструкции. Впрочем, такое расхождение в количественных оценках в определенной степени зависит от уровня обобщения (абстракции) в записи конструкций, который частично определяется целевой аудиторией (см. [раздел 2.2](#)), отбором материала (см. [раздел 2.3](#)) и связью конструктиконов с платформой FrameNet (см. [раздел 2.4](#)).

2.2. Полифункциональность

Целевой аудиторией Русского Конструктикона являются исследователи языка (русисты, типологи), студенты и преподаватели русского языка как иностранного (РКИ), а также специалисты по автоматической обработке текста.

Целям лингвистического описания русских конструкций служит разветвленная система семантических тегов, ориентированная на терминологию, принятую в типологической литературе («универсальный грамматический набор» значений, см. [Plungian 2011: 65]). Студентам и преподавателям РКИ адресован перевод толкований на другие языки, а также пометы сложности конструкций, соответствующие уровню владения языком. Разметка синтаксических структур и якорных элементов конструкции ориентирована на задачи в том числе автоматического выявления конструкций в текстах.

Среди других проектов только шведский конструктикон [Lyngfelt et al. 2018b] рассматривает в качестве своей целевой аудитории не только профессиональных лингвистов, но и студентов, изучающих шведский как иностранный.

⁵ The Swedish Constructicon: <https://spraakbanken.gu.se/konstruktikon>.

⁶ The Japanese FrameNet Constructicon, доступ через японский FrameNet: <http://jfn.st.hc.keio.ac.jp>.

⁷ The Brazilian Portuguese Constructicon, или the FN-Br Constructicon: <http://webtool.framenetbr.ufjf.br/index.php/webtool/report/cxn/main>.

⁸ The German Constructicon (GCon): <https://gsw.phil.hhu.de/constructicon/>.

Остальные ресурсы на данный момент преследуют скорее общетеоретические цели: описание конструкций и ограничений на их элементы, объединение конструкций в семьи. Заметим, что эти задачи решаются и в Русском Конструктиконе.

2.3. Отбор материала

При отборе материала предпочтение отдается неоднословным синтаксическим конструкциям, которые лежат на пересечении лексики и грамматики и являются «частично идиоматичными» (*partially schematic*, см. [Ehrlemaek et al. 2016], ср. также принятое в русскоязычной традиции понятие «конструкции малого синтаксиса» [Apresjan, Iomdin 2010]). С одной стороны, такие конструкции слишком некомпозициональны, чтобы быть понятными носителям русского языка. С другой стороны, они не настолько идиоматичны, чтобы считаться фразеологизмами. Как правило, такие конструкции имеют неизменяемую (якорную) часть и слоты, которые могут заполняться различными лексемами. Примером может служить конструкция (не) время VP-Inf (Самое время *вершить большие дела*), где якорной частью является слово *время*, допускающее перед собой отрицание, а слот заполняется глагольной группой в инфинитиве.

Наряду с лексикализованными конструкциями, конструктиконе Беркли для английского языка содержит описание конструкций более высокого уровня абстракции в плане синтаксиса, такие как: *Absolute_clause* (конструкция с нефинитной клаузой), *Coordination* (сочинительная конструкция), *Infinitival_relative_modal* (инфинитивная конструкция, модифицирующая именную группу) и др. Конструктикону Беркли близки аналогичные ресурсы для японского, немецкого и бразильского португальского языков.

Шведский конструктиконе так же, как и русский, фокусируется в основном на частично идиоматичных конструкциях, однако в перспективе предполагает покрытие всего лексико-грамматического континуума конструкций, которые будут иерархически организованы по принципу наследования признаков, от более абстрактных синтаксических до более частных лексических. Это связано с общим представлением языка как разветвленной системы конструкций [Lyngfelt 2018: 6], разделяемым создателями всех конструктиконов. В Русском Конструктиконе идея системного представления конструкций реализуется в выявлении *семей* конструкций и объединении их в более крупные группы — *гнезда* [Endresen et al. 2019].

2.4. Связь с другими семантическими ресурсами

Конструктиконе Филлмора и его аналоги «выросли» из платформы FrameNet, представляющей классификацию семантических фреймов в связи с их лексическим и синтаксическим выражением в том или ином языке [Fillmore et al. 2003]. FrameNet имеет жесткую структуру, устанавливая однозначное соответствие между формой и значением у каждого из аргументов предиката, а фреймовая семантика определяет роль отдельных элементов

в общем фрейме. В этих рамках разрабатывать описание менее прозрачных неоднозначных выражений крайне затруднительно. Конструктикон же позволяет описывать единицы, имеющие сложную внутреннюю и внешнюю структуру в отношении как семантики, так и синтаксиса, и видится как дополнение к FrameNet [Ohara 2018: 143].

Многие конструктиконы дают отсылку к связанному с конструкцией фрейму (*evoked frame*), если таковой имеется (ср. классификацию конструкций в соответствии с тем, провоцируют ли они возникновение того или иного фрейма в сознании говорящего, предложенную в [Ohara 2018]). Так, например, конструкция *be_recip* (e. g. *She is good friends with her mother*) в конструктиконе Беркли адресует к фрейму *Reciprocity* (Отношение взаимности), см. [Lee-Goldman, Petruck 2018: 32]. В конечном счете, уникальная метка FrameNet ID позволяет связать узлы Конструктикона с другими семантическими ресурсами и онтологиями (WordNet, PropBank, SemLink+ и др. [Palmer et al. 2014]), что открывает перспективы для его использования в IT-технологиях.

Заметим, что связывание через фреймы не является единственно возможным рецептом: FrameNet не раз критиковали за неполноту фреймовой сети, отсутствие единого принципа выделения фреймов в разных частях ресурса (см. [Baker, Ruppenhofer 2002]; [Shi, Mihalcea 2005]). Так, шведский Конструктикон ориентируется не на Swedish FrameNet+, а на SALDO — систему, представляющую лексикографические ресурсы как иерархически организованную сеть.

Для Русского Конструктикона связывание с другими ресурсами — задача на перспективу. Аналогом FrameNet для русского является русский ФреймБанк⁹ [Lyashevskaya, Kashkin 2015a], [Lyashevskaya, Kashkin 2015b]. Однако, последний покрывает меньше лексических зон (фокус прежде всего на глаголах, а не на именных и адвербиальных предикатах), а уровень представления семантики менее абстрактен: так, например, в сеть связываются конструкции индивидуальных глаголов, а не ассоциированные с ними обобщенные фреймы. Вместе с тем, материал ФреймБанка и Конструктикона пересекается. С одной стороны, во ФреймБанке документируются все конструкции предиката, засвидетельствованные в текстах, в том числе и такие идиоматизированные, как «стоило NP-Dat V-Inf, как/и Cl». С другой стороны, в ограниченном объеме Русский Конструктикон все же включает конструкции «нестандартного» глагольного управления, и в этом отношении ФреймБанк является одним из источников для пополнения списков конструкций Конструктикона. Наконец, Конструктикон и ФреймБанк объединяют общие принципы выделения и классификации семантических ролей (меток, приписываемых элементам слотам конструкции).

Несмотря на то, что прямой адресации между этими ресурсами сейчас нет, в будущем предполагается интегрировать информацию о тех фреймах, которые описаны во ФреймБанке, в документацию конструкций в Русском Конструктиконе. В более широком ключе, представляется перспективным

⁹ Доступ к ресурсу ФреймБанк: <https://github.com/olesar/framebank>.

ориентироваться, в рамках идеологии linked data, на инициативы по созданию мультязычных прикладных *Net-, *Bank- и *Link- ресурсов, а также исследовательских типологических баз данных по лексике, фразеологии и конструкциям.

С учетом названных отличительных особенностей Русского Конструктикона, прежде всего его полифункциональности, объема и содержания, разработан новый интерфейс, который ориентирован на решение и общетеоретических, и прикладных задач и подойдет пользователям разного типа. О нем речь пойдет в следующем **разделе 3**.

3. NP-Nom как на ладони: устройство ресурса и возможности поиска

На новом сайте Русского Конструктикона пользователям будут предложены различные возможности поиска конструкций. Для каждой конструкции на сайте будет отдельная словарная статья, включающая в себя несколько полей¹⁰, рассчитанных как на студентов РКИ, так и на лингвистов. Структура словарной статьи показана в **таблице 1**.

Поле **Имя конструкции/Name** содержит обобщенную запись конструкции, включающую как якорные элементы, так и слоты с ограничениями на их лексическое заполнение. Поле **Иллюстрация/Illustration** дает короткий репрезентативный пример употребления конструкции. Затем следует поле **Толкование/Definition**, в котором раскрывается семантика конструкции и указываются её отличия от синонимических вариантов¹¹. В поле **Перевод толкования/Translation of Definition** предлагается перевод толкования на английский и норвежский. Далее представлены: **Структура конструкции/Structure** в терминах универсальных зависимостей (UD)¹², **Якорные слова/Anchor words** (фиксированная часть конструкции), **Частотные заполнители слотов/Common fillers**, а также **Примеры/Examples** использования конструкции из НКРЯ. Поле **Уровень/CEFR Level** соотносит конструкцию с уровнем владения русским языком, которому она соответствует. Отдельное поле содержит **Семантические теги/Semantic tags**, которые

¹⁰ Названия всех полей на сайте даны на английском, тогда как содержание полей дается на русском, за исключением принятых сокращений (типа **NP** для именной группы, **Gen** для генитива и пр.) и переводов толкований. Для удобства читателя мы приводим русские эквиваленты английских обозначений.

¹¹ Толкования конструкций составляются таким образом, чтобы они были содержательны, но компактны и лаконичны. Мы по возможности избегаем узких лингвистических терминов, так чтобы и студенты РКИ без специальной лингвистической подготовки тоже могли эффективно пользоваться этим ресурсом. При составлении толкования мы опираемся на данные НКРЯ, сопоставляем конструкции внутри *семей* и *гнезд*, а также стараемся учитывать опубликованные результаты исследований конструкций русского языка, прежде всего работы Л. Л. Июдина, Д. О. Добровольского, А. Н. Баранова, С. С. Сая, В. Ю. Апресян, Н. Р. Добрушиной, А. Б. Летучего, Е. Л. Вилинбаховой, Н. М. Стойновой и др. Ссылки на учтенные работы планируется приводить в поле Комментария, а также привести на сайте проекта по адресу <https://site.uit.no/russian-constructicon/sources/>.

¹² Подробнее о принципах анализа в терминах UD см. <https://universaldependencies.org>.

обобщенно отражают семантику конструкции (подробнее см. [раздел 4](#)). Затем следуют поля, где указывается **Синтаксическая функция/Syntactic function** конструкции целиком и **Синтаксическая структура якорной части/Syntactic structure of the anchor**. В поле **Стилистический тег/Usage label** дается информация о регистре, в котором конструкция уместна. Наконец, поле **Комментарий/Comment** содержит дополнительную информацию о конструкции: ссылки на синонимичные и антонимичные конструкции, представленные в конструктиконе, как опция могут приводиться ссылки на научные исследования данной конструкции.

На стартовой странице нового сайта (**Browse**) пользователь сможет посмотреть полный список конструкций, по которым можно осуществлять поиск. Вводя в поисковую строку якорные слова или словоформы, можно будет искать конструкции по точным элементам. Например, если ввести в строку слово *ещё*, пользователю будет предложен список конструкций, включающих эту лексему.

Сайт будет содержать несколько вкладок.

Вкладка **Дневная порция/Daily Dose** будет выдавать случайную подборку из пяти конструкций, соответствующих уровню владения языком, задаваемому в меню. Такая обозримая порция конструкций может использоваться студентами РКИ для самостоятельной работы.

Вкладка **Продвинутый поиск/Advanced Search** будет предоставлять разнообразные возможности поиска, предусмотренные для лингвистов. Здесь можно будет задать грамматические и синтаксические параметры, по которым будет осуществляться запрос. В разделе **Морфология/Morphology** пользователь сможет задать поиск конструкции по таким грамматическим категориям, как **Падеж/Case**, **Род/Gender**, **Число/Number**, **Вид/Aspect**, **Наклонение/Mood**, **Время/Tense** и др.

Например, можно будет найти все конструкции с именными группами в дальнем падеже. Пользователю можно будет выбрать как один пункт из списка, так и комбинацию грамматических признаков. В разделе **Синтаксис/Syntax** параметрами поиска будут теги, характеризующие синтаксическую функцию конструкции в целом и состав якорных элементов конструкции. Здесь пользователь сможет найти, к примеру, все биклаузальные конструкции или все конструкции с производными предлогами. Кроме того, поиск будет возможен и по таким структурным особенностям, как инверсионный порядок слов, наличие отрицания и редупликация лексических элементов конструкции. В разделе **Семантика конструкции/Semantics** можно будет задать поиск конструкций по семантическим тегам и их подтипам. Так, можно будет найти все конструкции, обозначающие причину события (тег Cause) или условие его совершения (тег Condition), и проч. В разделе **Семантические роли/Semantic roles** можно будет задать поиск по тегам Агенс, Пациенс, Адресат и др., которыми размечены толкования конструкций и примеры употребления (см. [таблицу 1](#)). В разделе **Уровень/CEFR** можно будет ограничить выдачу конструкций только тем уровнем, который интересует пользователя. Уровни будут варьироваться от A1 до C2 по общеевропейской шкале оценки уровня владения языком. Пользователь сможет выбрать один или несколько уровней из списка. Кроме того, при оформлении запроса все указанные параметры можно будет комбинировать.

Таблица 1. Структура словарной статьи в Русском Конструктиконе

Имя конструкции/ Name	какой/какое (там) VP-Inf.Imp! ¹³
Иллюстрация/ Illustration	<i>Какое там гулять! У нас нет времени!</i>
Толкование/ Definition	Конструкция обозначает, что говорящий отклоняет [действие] _{Action} , названное в предыдущем контексте, потому что считает это действие неуместным или несвоевременным. Обычно конструкция используется в ситуации диалога и выражает реакцию говорящего на предшествующую реплику собеседника. Конструкция также встречается при передаче внутренней речи говорящего, которая конструируется как диалог говорящего с самим собой.
Перевод толкования/ Translation of Definition	This construction means that the speaker rejects the suggested action named in the preceding context, because he/she finds this action inappropriate or poorly timed. Usually, this construction is used in a dialogue and expresses the speaker's reaction to the preceding remark of his/her conversation partner. The construction can also occur when presenting the inner monologue of the speaker that is constructed as a conversation in which the speaker is talking to him- or herself.
Структура конструкции/ Structure	[[advmod какой [advmod там]] VP-Inf.Imp!] [[advmod какое [advmod там]] VP-Inf.Imp!]
Якорные слова/ Anchor words	какой, какое, там
Частотные заполнители слотов /Common fillers	спать, работать, гулять, отдыхать
Примеры/Examples	— <i>Ладно, пойду [спать]_{Action}!</i> — <i>Какой [спать]_{Action}, Марина Михайловна? — возмутилась Ира. — Посидите с нами, хоть чайку попейте.</i> — <i>Вам суп [наливать]_{Action}?</i> <i>Я посмотрела на суп. Какое там [наливать]_{Action}! Серый, непрозрачный, он не вызывал у меня аппетита. — Нет, благодарю, ответила я.</i> <i>И вот народ галдит. Я там присел, попытался [помедитировать]_{Action} как-то. Какое там [медитировать]_{Action}! Ничего подобного там и близко невозможно — весь этот галдеж и тусовка.</i>

¹³ Через косую черту указываются равноправные варианты элементов конструкции, а в скобках даются факультативные элементы.

Имя конструкции/ Name	какой/какое (там) VP-Inf.Imp! ¹³
Уровень/CEFR Level	C1
Семантические теги/Semantic tags	Reaction to the previous discourse: Disagreement
Синтаксическая функция/Syntactic function	Clause
Синтаксическая структура якорной части/Syntactic structure of the anchor	Adj (APRO); Adv (PART)
Стилистический тег/Usage label	Colloquial
Комментарий/ Comment	см. конструкции <i>какой/какое (там) VP-Past! Какой там поел!</i> и <i>какой/какое (там) Adv/Adj! Какое там быстро!</i>

Также на сайте будет вкладка с инструкцией для пользователей (**Instructions**) и вкладка с информацией о ресурсе и разработчиках (**About**).

4. NP-Ном что надо: семантическая разметка конструкций

Для семантической разметки конструкций в Русском Конструктиконе используется двухуровневая система семантических помет (тегов), которая включает 53 общих тега и 87 уточняющих тегов-подтипов. Так, например, общий тег Фаза действия имеет подтипы Инхоатив, Континуатив, Терминатив и Кунктатив согласно [Plungian 1999]. Общий тег Сравнение имеет подтипы Сходство, Идентичность, Несходство, Контраст и Имитация, в соответствии с [Treis 2018]. В **таблице 2** приведен список из 25 «крупных» общих тегов, каждым из которых в Русском Конструктиконе размечено более 40 конструкций. В правом столбце таблицы приведено количество конструкций для каждого тега.

Таблица 2. Общие семантические теги, описывающие более 40 конструкций в базе Русского Конструктикона

Семантический тег	Кол-во
Интенсификация/Intensity	280
Оценка/Assessment	223
Отношение/Attitude	209
Сравнение/Comparison	167
Реакция на предшествующий дискурс/Reaction to the previous discourse	149
Темпоральность/Time	139
Категоризация объектов/Categorization	139
Образ действия/Manner	134
Измерение количества/Measure	120
Временное состояние/Temporary state	109
Фаза действия/Phase of action	105
Причина/Cause	96
Пространство/Spatial expressions	87
Условие/Condition	83
Побуждение к действию/Causation	82
Степень уверенности/Degree of certainty	77
Уступка/Concession	69
Временная референция/Timeline	59
Миративность/Mirative ¹⁴	57
Степень точности/Degree of accuracy	52
Аддитив/Additive	50
Прохибитив/Prohibitive ¹⁵	49
Цель/Purpose	49
Глагольная множественность/Pluractionality	47
Эмфаза/Emphasis	41

¹⁴ Тег Миративность/Mirative широко употребляется в типологической литературе для обозначения лингвистической категории, выражающей удивление говорящего неожиданно установленным фактом [DeLancey 1997]; [Aikhenvald 2012]. Примерами миративных конструкций являются NP-Ном (как) давай VP-Inf (он как давай петь), XR-таки, (а Cl)! (Пришёл-таки, а мы уже и не ждали!) и др.

¹⁵ Термин Прохибитив/Prohibitive описывает лингвистические средства, выражающие запрет на совершение действия: нечего VP-Inf (Нечего баловать детей). Следуя [Rakhilina 2013], в класс прохибитивных конструкций мы также включаем континуативные прохибитивы, выражающие побуждение прервать некоторую ситуацию: хорош VP-Inf! (Эй, наверху! Хорош прыгать!), полно Pron-Dat VP-Inf (полно тебе плакать!), по-VP-Past и хватит (погуляли и хватит) и др.

Те общие семантические теги, которые не вошли в **таблицу 2**, описывают меньшие группы конструкций: Опасение/Apprehension, Угроза/Threat, Просьба/Request, Результат/Result, Дистрибутив/Distributive, Каритив/Caritive; Конатив/Conative, Кванторы/Quantification, Реализованная возможность/Actuality, Calculation/Счет, Подмножество/Subset, Инструмент/Instrument и др.

Многие конструкции имеют более одного общего семантического тега. Так, 933 конструкции (42%) имеют по два общих тега, а 144 конструкции имеют по 3 общих тега. Например, конструкция **(уж/ну) Pron-Pron, a Cl** (*уж кому-кому, а ему можно доверять*) имеет теги Реакция на предшествующий дискурс, Степень уверенности и Исключительность, а конструкция **Noun-Nom на Noun-Loc** (*ошибка на ошибке*) имеет набор тегов Измерение количества, Отношение и Интенсификация.

Система семантических тегов используется для выявления, анализа и системного пополнения семей и гнезд конструкций (см. [Endresen et al. 2019]). По семантическим тегам можно осуществлять поиск конструкций и их групп и сравнивать этот материал с конструкциями других языков. Кроме того, в процессе работы исследуется, насколько широко конструкции русского языка покрывают набор значений, которые в других языках могут выражаться грамматически.

5. Во всех отношениях AP: вместо заключения

Конструкции представляют собой то звено, где происходит пересечение лексики и грамматики. При этом уже ведется работа по пополнению базы морфологическими конструкциями, и таким образом учитывается роль словообразовательных средств: ср. конструкцию **по-XP-ee(ей)** (*смотреть на жизнь повеселее*) для выражения семантики аттенуативного сравнения [Boguslavsky, Iomdin 2009], **пере-V все NP-Acc.Plur** (*перемыть все тарелки в доме*) с семантикой дистрибутива [Tatevosov 2013], и пр.

В целом, обширный материал по конструкциям русского языка, собранный и описанный в Конструктиконе, дает возможность уточнить наше понимание устройства русской грамматической системы. В ходе проекта выверена детальная и многоуровневая семантико-синтаксическая разметка конструкций, на основе которой разработана эффективная методика выявления семей и гнезд конструкций (позволившая, в частности, значительно и системно расширить инвентарь конструкций в базе).

Ближайшей задачей Русского Конструктикона является дальнейшая систематизация семантико-синтаксической иерархии, которая позволит структурировать толкования конструкций, а в дальнейшем добиться большей адекватности в выборе их переводных эквивалентов — для английского, норвежского и других языков. Опыт работы над Русским Конструктиконе может помочь и в создании аналогичных ресурсов для других языков.

References

1. *Aikhenvald, A.* (2012), The essence of mirativity, *Linguistic Typology*, Vol. 16, pp. 435–485.
2. *Apresjan, Ju. D., Iomdin, L. L.* (2010), Small syntax constructions [Konstruktivii malogo sintaksisa]. In Apresjan Ju. D., Boguslavsky I. M., Iomdin L. L., Sannikov V. Z. (eds.). *Theoretical problems of the Russian syntax: Interaction of grammar and lexicon [Teoretičeskie problemy russkogo sintaksisa: Vzaimodejstvie grammatiki i slovary]*, Moscow, pp. 59–280.
3. *Baker, C., and Ruppenhofer, J.* (2002), Framenet's frames versus Levin's verb classes, *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, Berkeley (CA): BLS, pp. 27–38.
4. *Boas, H. C., Lyngfelt, B., Torrent, T. T.* (2019), Framing constructicography, *Lexicographica*, Vol. 35(1), pp. 41–85.
5. *Boguslavsky, I., Iomdin, L.* (2009), Semantics of Attenuated Comparatives in Russian, *Proceedings of the 4th International Conference on Meaning-Text Theory. Observatoire de linguistique Sens-Texte (OLST)*, pp. 65–77.
6. *Croft, W.* (2001), *Radical Construction Grammar*, Oxford University Press, Oxford.
7. *DeLancey, S.* (1997), Mirativity: The grammatical marking of unexpected information. *Linguistic Typology*, Vol. 1, pp. 33–52.
8. *Ehrlemark, A., Johansson, R., Lyngfelt, B.* (2016), Retrieving Occurrences of Grammatical Constructions, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pp. 815–824.
9. *Endresen, A., Klezovich, A., Lyashevskaya, O., Mordashova, D., Nordrum, M., Rakhilina, E., Tyers, F., Zhukova, V.* (2019), Building a Constructicon for Russian: How to identify families of constructions, A talk presented at the 15th International Cognitive Linguistics Conference (ICLC-15), Kwansei Gakuin University, Nishinomiya, Japan, August 2019.
10. *Fillmore, C. J.* (2008), Border Conflicts: FrameNet Meets Construction Grammar, Bernal, E., DeCesaris, J. (Eds.), *Proceedings of the XIII EURALEX International Congress*, Barcelona, Universitat Pompeu Fabra, pp. 49–68.
11. *Fillmore, Ch. J., Kay, P., O'Connor, M. C.* (1988), Regularity and idiomatity in grammatical constructions: The case of *let alone*, *Language*, Vol. 64(3), pp. 501–538.
12. *Fillmore, C. J., Johnson, C. R., Petruck, M. R.* (2003), Background to FrameNet, *International Journal of Lexicography*, Vol. 16.3, pp. 235–250.
13. *Fillmore, C. J., Lee-Goldman, R., Rhodes, R.,* (2012), The FrameNet constructicon, Boas, H. C., Sag, I. A. (eds.), *Sign-based construction grammar*, CSLI Publications, Stanford, CA, pp. 309–372.
14. *Goldberg, A. E.* (2006), *Constructions at Work: The Nature of Generalizations in Language*, Oxford University Press, Oxford.
15. *Janda, L. A., Lyashevskaya, O., Nessel, T., Rakhilina, E., Tyers, F. M.* (2018), A constructicon for Russian: Filling in the Gaps, Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. T. (eds.), *Constructicography: Constructicon development across languages*, John Benjamins, Amsterdam, pp. 165–181.

16. *Lee-Goldman R., Petruck, M.* (2018), The FrameNet constructicon in action. Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. T. (eds.), *Constructicography: Constructicon development across languages*, John Benjamins, Amsterdam, pp. 19–40.
17. *Lyashevskaya, O., Kashkin, E.* (2015a), FrameBank: a database of Russian lexical constructions, Khachay, M. Yu., Konstantinova, N., Panchenko, A., Ignatov, D. I., Labunets, G. V. (eds.), *Analysis of Images, Social Networks and Texts. Fourth International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers. Communications in Computer and Information Science*, Vol. 542, Springer, pp. 337–348.
18. *Lyashevskaya, O., Kashkin, E.* (2015b), Annotation of lexical constructions in Russian FrameBank [Tipy informacii o leksičeskix konstrukcijax v sisteme FrameBank], *Trudy Instituta russkogo jazyka imeni V. V. Vinogradova [Proceedings of Vinogradov Institute of the Russian Language]*, Vol. 6, pp. 464–555.
19. *Lyngfelt, B.* (2018), Introduction: Constructicons and constructicography, Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. T. (eds.), *Constructicography: Constructicon development across languages*, John Benjamins, Amsterdam, pp. 1–18.
20. *Lyngfelt, B., Bäckström, L., Borin, L., Ehrlemark, A., Rydstedt, R.* (2018b), Constructicography at work: Theory meets practice in the Swedish constructicon. Lyngfelt et al. 2018a, pp. 41–106.
21. *Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. T.* (eds.) (2018a), *Constructicography: Constructicon development across languages*, John Benjamins, Amsterdam.
22. *Ohara, K.* (2018), Relations between frames and constructions: A proposal from the Japanese FrameNet constructicon, Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. T. (eds.), *Constructicography: Constructicon development across languages*, John Benjamins, Amsterdam, pp. 141–164.
23. *Palmer, M., Bonial, C., McCarthy, D.* (2014), Semlink+: FrameNet, VerbNet and event ontologies, *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014)*, pp. 13–17.
24. *Plungian, V. A.* (1999), A typology of phasal meanings, Abraham, W., Kulikov, L. (eds.) *Tense-aspect, transitivity, and causativity: Essays in honor of Vladimir Nedjalkov*, John Benjamins, Amsterdam.
25. *Plungian, V. A.* (2011), An introduction to grammatical semantics: Grammatical meanings and grammatical systems in the languages of the world [Vvedenie v grammatičeskiju semantiku: Grammatičeskie značeniya i grammatičeskie sistemy jazykov mira]. Russian State University for the Humanities, Moscow.
26. *Rakhilina, E. V.* (ed.) (2010), *Linguistics of constructions [Lingvistika konstrukcij]*, Azbukovnik, Moscow.
27. *Rakhilina, E. V.* (2013), Conductor, press the brakes... [Konduktor, nažmi na tormoza...], *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” [Komp’juternaja lingvistika i intellektual’nye tehnologii: po materialam meždunarodnoj konferencii «Dialogue»]*, Issue 12 (19), Russian State University for the Humanities, Moscow, pp. 665–673.

28. Shi, L., Mihalcea, R. (2005), Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing, International conference on intelligent text processing and computational linguistics, Springer, Berlin, Heidelberg, pp. 100–111.
29. Stojnova, N. M. (2007), The construction *vzjat' i sdelat'* 'take and do' in Russian [Konstrukcija *vzjat' i sdelat'* v ruskom jazyke], Dudčuk, F. I., Ivlieva, N. V., Podobrjaev, A. V. (eds.), Structures and interpretations: Studies in theoretical and applied linguistics by young researchers [Struktury i interpretacii: raboty molodyx issledovatelej po teoretičeskoj i prikladnoj lingvistike], Moscow State University, Moscow, pp. 158–189.
30. Tatevosov, S. G. (2013), Destructive notes on the distributive Aktionsart [Destrukтивnye zametki o raspredelitel'nom sposobe dejstvija], Russkij jazyk v naučnom osveščanii. Vol. 2(26), pp. 30–51.
31. Treis, Y. (2018), Comparative Constructions: An Introduction, Linguistic Discovery, Vol. 16(1). pp. i–xxvi.

QUANTILE-BASED APPROACH TO ESTIMATING COGNITIVE TEXT COMPLEXITY

Eremeev M. A. (maks5507@yandex.ru)

Lomonosov Moscow State University (MSU), Moscow, Russia

Vorontsov K. V. (vokov@forecsys.ru)

Moscow Institute of Physics and Technology (MIPT),
Moscow, Russia

This paper introduces an approach to measuring the cognitive complexity of texts on various language levels. While standard readability indices are based on the linear combination of primary statistics, our general approach allows us to estimate complexity on morphological, lexical, syntactic, and discursive levels. Each model is defined by the tokens for the specific language level and the complexity function of a single token. We then use the reference collection of moderately complex texts and the quantile-based approach to spot the abnormally rare tokens. The proposed supervised ensemble, based on the ElasticNet model, incorporates models from all language levels. Having collected a labeled dataset through crowdsourcing, consisting of pairs of articles from the Russian Wikipedia, we consider several models and ensembles and compare them to common baselines. Suggested models are flexible due to the freedom in choosing the reference collection. The described experiments confirm the competitiveness of the proposed approach, as the ensembles demonstrate the best target metric value.

Key words: cognitive complexity, language levels, ElasticNet, supervised learning, exploratory search

DOI: 10.28995/2075-7182-2020-19-256-269

КВАНТИЛЬНЫЙ ПОДХОД К ОЦЕНИВАНИЮ КОГНИТИВНОЙ СЛОЖНОСТИ ТЕКСТА

Еремеев М. А. (maks5507@yandex.ru)

Московский Государственный Университет
им. М. В. Ломоносова (МГУ), Москва, Россия

Воронцов К. В. (vokov@forecsys.ru)

Московский Физико-Технический
Институт (МФТИ), Москва, Россия

В данной статье описан подход к оцениванию когнитивной сложности текста на разных уровнях языка. В отличие от индексов удобочитаемости, которые основаны на линейной комбинации текстовых статистик, мы предлагаем обобщенный подход, позволяющий оценивать сложность на морфологическом, лексическом, синтаксическом и дискурсивном уровнях языка. Мы используем референтный корпус текстов и квантильный подход для определения токенов с аномальной частотой. Собрав выборку размеченных пар документов русской Википедии, мы также обучаем и исследуем линейную комбинацию моделей со всех уровней языка. Приведенные в статье результаты экспериментов показывают конкурентоспособность предложенного подхода.

Ключевые слова: когнитивная сложность, уровни языка, обучение с учителем, разведочный поиск

1. Introduction

Automated text complexity measurement tools have been proposed in order to help teachers to select textbooks that correspond to the students' comprehension level and publishers to explore whether their articles are readable. Thus, plenty of readability indexes (RIs) were developed. Readability indexes focus on estimating complexity by evaluating aggregated syntactic and lexical features of the whole texts. There are many well-known RIs, such as *Automated Readability Index* [13], *Flesch-Kincaid readability tests* [7], *Gunning fog* [8] and fairly modern ones like *Linsear Write Formula* [11]. They all use statistics like the total number of words, mean number of words per sentence, or the number of syllables to evaluate how complex given text is. By combining these statistics, RIs assign the given document a *complexity score*. For instance, an Automated Readability Index (ARI) has the following form for the document d :

$$ARI(d) = 4.71 \times \frac{c}{w} + 0.5 \times \frac{w}{s} - 21.43, \quad (1)$$

where c refers to the total number of letters in document d , w is the total number of words, and s denotes the total number of sentences in d .

RIs are interpretable and easy to implement. However, due to the significant amount of constants, they are language-dependent and, most of the time, tailored to the US grade level system. That restrains the number of possible applications a lot.

As for research on complexity estimation of the Russian text, it is worth highlighting works of I. Osborneva [12], where she derives new version of *Flesch Readability Ease* (FRE) [10], customized for the Russian language.

$$FRE(d) = 206.836 - (1.52 \times ASL) - (65.14 - ASW), \quad (2)$$

where ASL stands for the mean number of words per sentence, ASW —for mean syllables per word. In 2018 V. Solovyev [14] obtains new readability formula created explicitly for Russian documents. Text complexities are valuable in different areas, e.g., [4] describes complexity formulas for legal documents in Russian.

In 2007 [1] introduced psychophysiological (cognitive) methods of measuring text complexity, highlighting the following assumptions:

1. Any text can be considered as a sequence of tokens (codes)—parts of the finite alphabet—letters, syllables, sentences, words, etc.
2. When reading the text, our nervous system decodes the tokens, progressively on the following language levels: morphological, lexical, syntactic, discursive, and semantic.
3. Decoding processes occur in different nervous system zones (e.g., part of the cortex). Each zone is responsible for the specific token on a specific language level. When the zone finishes the decoding process, it moves into the state of refractoriness and needs time to recover. During the recovery, the zone cannot execute decoding and forces another zone to take the load. Such a redistribution of nervous system resources diminishes effectiveness of the nervous system as a whole, and the person starts perceiving the document with more effort.
4. Thus, if the token's distance to the previous occurrence exceeds some threshold, the nervous system must allocate additional resources to decode it. Such terms are considered complex. Hence, the complexity of the document is a combination of abnormally frequent (complex) tokens.

In [1], authors propose to count the mentioned threshold as a quantile of the empirical distribution, calculated over the large set of simple texts (*reference collection*). They explore the morphological level, considering letters as a token. [2] introduces a lexical level model, assuming the word complexity is determined only by its length. [19] features the model on the discursive level, counting the number of connector words and phrases in each sentence.

Based on the assumptions above, in this paper we elaborate our research presented in [6], offering models on the morphological, lexical, and syntactic levels, and then training the linear model, obtaining the all-levels complexity model. Experiments were performed on two datasets. We compare models with readability indexes and cognitive models proposed in [1], [2], [19].

2. General Model

Let d be the arbitrary document, consisting of tokens x_1, \dots, x_n from a fixed token alphabet A_h . Here, h refers to the language level, i.e., morphological, lexical, syntactic, or discursive. So, the tokens may be letters, syllables, sentences, words, etc. We denote c_i to be the *cognitive complexity score* of token x_i , and w_i —its weight. The *document complexity score* then is a sum of weights over tokens having abnormal complexities.

To measure the token complexity, we use a reference collection—a set of moderately complex texts—to calculate empirical distributions of complexity scores for each token. Thus, the token's complexity is abnormal when it is greater than a γ -quantile of the counted distribution (**figure 1**) (assumption 4). In our experiments we use *Russian Wikipedia* and *Noosphere* (noosphere.ru) open corpora as reference collections.

Former comprises more domain-specific documents (1.5M in total), while the latter incorporates various types of texts, including fiction and poems (200K in total).

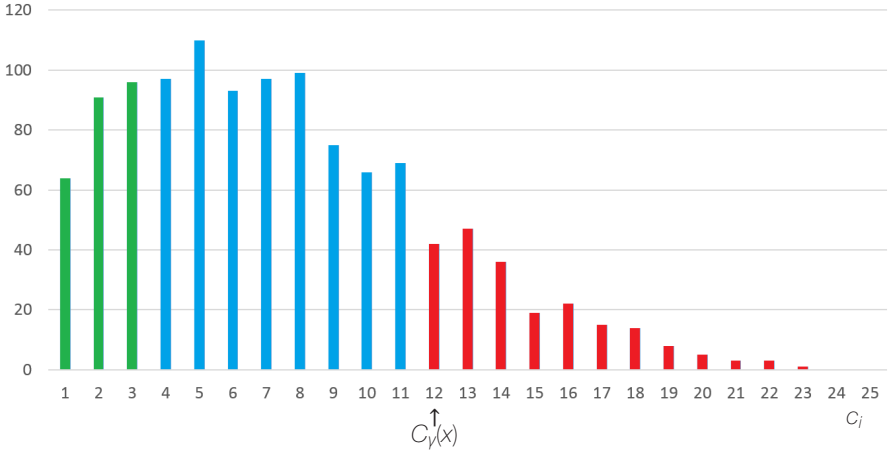


Figure 1: Sample distribution of complexity scores and its γ -quantile

Finally, document d complexity $W(d)$ is calculated by aggregating complexity scores c_i of complex tokens in d .

$$W(d) = \sum_{i=1}^n w_i [c_i > C_\gamma(x_i)], \tag{3}$$

where $[\]$ refers to the Iverson notation (i.e. $[true] = 1, [false] = 0$), n is the number of tokens from A_h in document d . Some examples of interpretable weights w_i are presented in [table 1](#).

Raising the weight to the p power, we obtain a nonlinear sum of weights:

$$W(d) = \sum_{i=1}^n w_i^p [c_i > C_\gamma(x_i)], \tag{4}$$

where $p > 0$ is an integer power.

If the token x_i does not appear in the reference collection, we set $C_\gamma(x_i)$ equal to $-\infty$, therefore always counting it as abnormally complex.

Thus, to set up the model, we need to specify the reference collection D , the alphabet A_h , token complexity function c , weights w , and power p .

Table 1: Weights w_i examples

w_i	Meaning of w_i
1	number of complex tokens
$1/n \times 100\%$	complex tokens percentage
c_i	total complexity
c_i/n	mean complexity
$c_i - C_\gamma(x_i)$	excessive complexity
$(c_i - C_\gamma(x_i))/n$	mean excessive complexity

3. Token complexity functions

Firstly, we indicate two approaches to estimating the complexity of a single token.

3.1. Distance-based complexity function

According to assumptions 3–5, let r_i be a distance from previous token occurrence x_i to its current occurrence in the text:

$$\dots \boxed{x_{i-r_i} = a} \underbrace{x_{i-r_i+1} \ x_{i-r_i+2} \ \dots \ x_{i-2} \ x_{i-1} \ \boxed{x_i = a}}_{r_i} \dots$$

Equally,

$$r_i = \min_{1 \leq j < i} \{i - j \mid x_i = x_j\}. \tag{5}$$

If i is the first occurrence of term t_i in document d , there is no previous occurrence, so r_i is undefined. To solve this issue, we redefine r_i so that sum of r_i over all tokens $x_i = a$ is equal to n .

For example, if A_h consists of the letters:

Table 2: r_i and redefined r_i examples for letter-based model

token	t	h	e	g	r	e	a	t	g	a	t	s	b	y
r_i	—	—	—	—	—	3	—	7	5	3	3	—	—	—
redefined r_i	4	15	11	9	14	3	11	7	5	3	3	14	14	14

Then, we define token complexity function as some decreasing function f of r_i :

$$c_i = f(r_i) \tag{6}$$

The f should be decreasing according to the assumption 4, as only the most frequent terms put pressure on the nervous system. Example of f :

$$c_i = -r_i, \tag{7}$$

Hence, we build an empirical distribution of complexities $\{f(r_i) \mid x_i = a\}$ for all tokens $a \in A_h$, count corresponding quantiles $C_\gamma(x_i)$ and, finally, calculate the complexity score, according to formula (4).

3.2. Counter-based complexity functions

In the counter-based approach, we assume every term has fixed complexity score (not depending on position in the text), so alphabet A_h includes the only token: $A_h = \{a\}$. In other words, the token’s complexity is defined only by its linguistic properties (e.g., length of the word or sentence).

Taking that into account, we construct single empirical distribution over all tokens. Therefore, the quantile is one for all tokens $C_\gamma(x_i) = C_\gamma$ and model (4) takes the following form:

$$W(d) = \sum_{i=1}^n w_i^p [c_i > C_\gamma] \tag{8}$$

4. Considered models

Trying different combinations of tokens and complexity functions, we want to share models on four language levels.

4.1. Morphological complexity models

At the morphological level, tokens are letters, morphemes, syllables, or, in general case, n -grams. Also, we can sort the letters in n -gram, therefore lessening the vocabulary size to acquire more reliable distributions. Indeed, our brain easily handles local letter permutations, so they do not affect the complexity much.

In our experiments, we use a distance-based model with complexity function (7) for both *letters*, *sorted and unsorted syllables*.

The examples of empirical distributions for letter-based models over the Russian Wikipedia and Noosphere reference collections are introduced in **figure 2**. Comparison of the distributions for syllables-based and sorted-syllables-based models are presented in **figure 3**.

4.2. Lexical complexity models

Here we use separate words as tokens. However, in such a case, the vocabulary turns out to be vast and makes the distributions less precise. To shrink it, we eliminate all short words (less than a length of 3) and too rare words (that appears only once on the whole reference collection).

4.2.1. Distance-based model

The distance-based complexity model uses complexity function (7) as it calculates the distributions of the score for every word (*lexical distance model*). The example of the distribution is shown in **figure 4**.

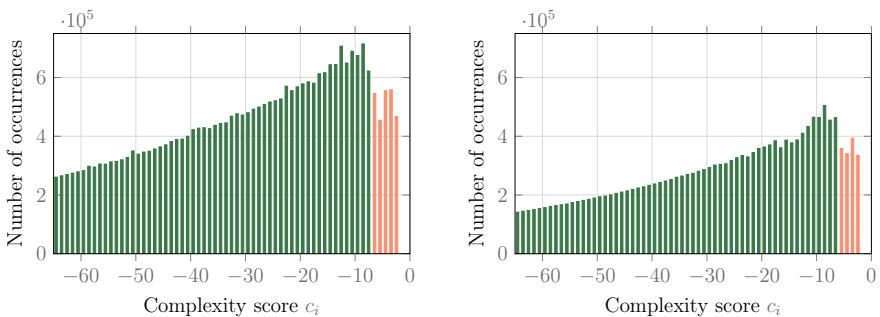


Figure 2: Distribution of c_i for the letter «Y», calculated over the Russian Wikipedia and Noosphere collections. The orange part of the distribution correspond to $c_i > C_\gamma(x)$, $\gamma = 0.95$

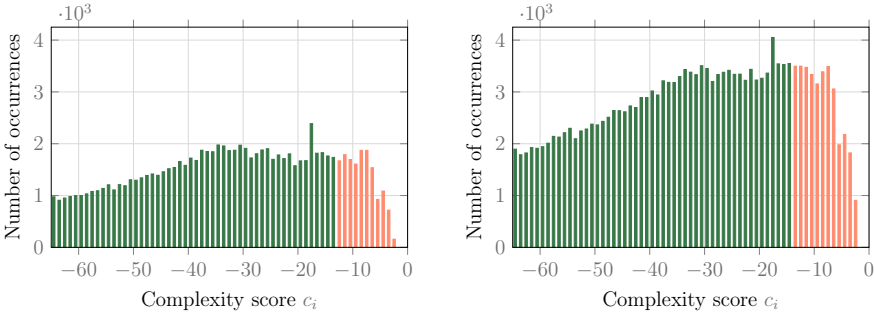


Figure 3: Distribution of c_i for the syllable «ЛОК», calculated over the Russian Wikipedia collection for models with and without sorting. The orange part of the distribution corresponds to $c_i > C_\gamma(x)$, $\gamma=0.95$

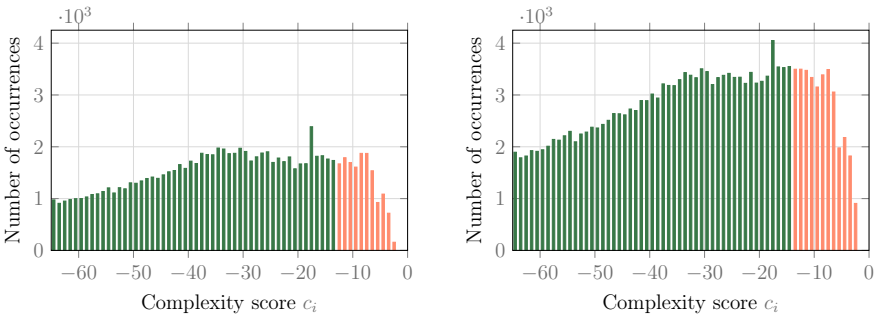


Figure 4: Distribution of c_i for the word «МАТЕМАТИКА», calculated over the Russian Wikipedia collection. The orange part of the distribution corresponds to $c_i > C_\gamma(x)$, $\gamma=0.95$

4.2.2. Counter-based models

We explore two functions here. Firstly, [2] defines the complexity of the word as its length (*lexical length model*). Therefore, the model builds empirical distribution over all words' lengths and counts the word as complex if it is long enough.

Advancing this approach, we consider not the word length, but its counter value $count(x_i)$, which is the number of times word x_i appears in reference collection (*lexical counter model*). The complexity function should be a decreasing function of $count(x_i)$. For example:

$$c_i = -count(x_i) \tag{9}$$

4.3. Syntactic complexity models

To estimate syntactic complexity, we use UDPipe [15] to extract syntactic dependencies, part of speeches (noun, verb, adjective, etc.) and sentence parts (subject, object, attribute, etc.). Using derived information, we propose two models.

4.3.1. Distance-based model

Let A_h be a product of PoS —set of all parts of speech may occur, and SP —set of all sentence parts. Therefore each $a \in A_h$ is a pair (p, s) , where $p \in PoS$ and $s \in SP$ are part of speech and sentence part respectively. We call such pairs *syntgams*.

We apply the distance complexity function (7) to such tokens to receive a distance-based syntactic model (*syntactic syntgam model*).

4.3.2. Counter-based model

Using the syntactic dependencies returned by the parser, we define the complexity function as a length of the dependency (alike using word length [2]) and acquire the counter-based syntactic model (*syntactic length model*). The examples of distributions are shown in figure 5.

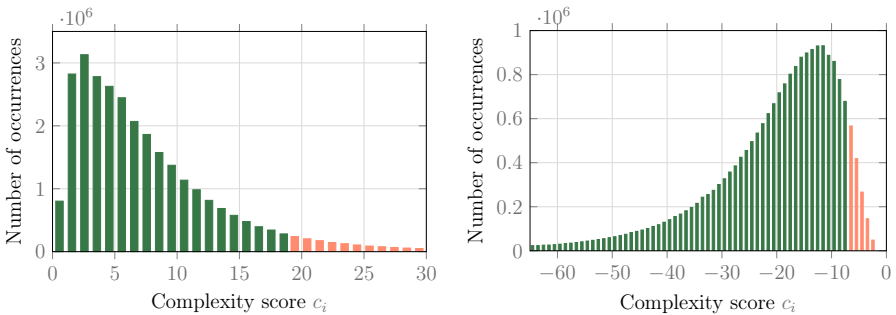


Figure 5: Distribution of syntactic dependencies' lengths and distance-based c_i for syntgam (verb, root), calculated over the Russian Wikipedia dataset. The orange part of the distribution corresponds to $c_i > C_\gamma(x)$, $\gamma=0.95$

4.4. Discursive complexity models

The last but not least language level we consider is the discursive level, initially proposed in [19]. On this level, model evaluates the meaningfulness of text, its coherence, and consistency.

To evaluate the complexity the vocabulary of common connector-words for the Russian language (i.e., «который», «из-за того что», «с тех пор как», etc.) is used. Thus, the more such connectors appear in the document, the more complex it is.

Therefore, we define a counter-based model with sentences as tokens, and complexity function equal to the number of connectors in the sentence (*discursive connectors model*).

5. Dataset

We used a crowdfunding platform Yandex.Toloka to gather a labeled dataset of pairs of Russian Wikipedia pages.

Assessors were asked to label 10K pairs of Russian Wikipedia articles. We suggested them to read both pages carefully and choose which is more challenging to comprehend. The interface consisted of two links to evaluated articles and four options to choose from: “LEFT” or “RIGHT” when an assessor assumes the left or the right document is more complex, “EQUAL” in case the assessor cannot determine which document is more challenging to comprehend and “INVALID” option if the documents in given pair lie in different domains. The interface is shown in **figure 6**.

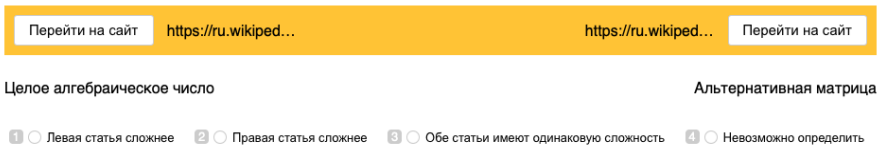


Figure 6: Interface for assessors at Yandex.Toloka

We chose documents from math, physics, medicine, and programming areas. The topic modeling approach [9], namely the Additive Regularization of Topic Models (ARTM) theory [16], was used to cluster the documents by fields. ARTM features an effective way to build structured multimodal topic models [17], [3]. We included the modalities of words and word collocations, obtained with TopMine algorithm [5]. Then, documents from a single domain and with almost identical lengths formed the pairs. Examples of document pairs to assess are introduced in **table 3**.

Each pair was labeled by two assessors to avoid human factor mistakes. We assume that the pair was correctly labeled if labels were not controversial, i.e., one assessor labeled the first document as more complex while others chose the second document. If the pair was labeled as ‘INVALID’ at least by a single person, we also eliminated it from the final dataset.

Table 3: Examples of labeled document pairs

Left Document	Right Document	Which document is more complex
Matrix	Tensor	RIGHT
Rational number	Fraction (mathematics)	LEFT
Proton	Neutron	EQUAL
Mac OS X	Convex Hull	INVALID

So, 8K pairs out of 10K were correctly labeled and formed the dataset

$$D = \{(d, d') \mid d' \text{ is more complex than } d\}.$$

To shorten the calculations and formulas, let’s denote $(d, d') \in D$ as $d < d'$.

6. Ensembling models

Having the dataset, we can train a supervised model to piece together all the proposed models. Such an ensemble combines estimations from all language levels.

We chose a linear combination to be the resulted model:

$$W(d, \alpha) = \sum_{k=1}^K \alpha_k W_k(d), \quad \alpha_k \geq 0, \quad (10)$$

where vector α is the solution to the optimization problem:

$$\sum_{d \prec d'} \underbrace{\mathcal{L}(W(d', \alpha) - W(d, \alpha))}_{\text{pair-wise margin}} \rightarrow \min_{\alpha}, \quad (11)$$

where $\mathcal{L}(M)$ is a smooth, non-increasing function of margin M .

To avoid overfitting, we use ElasticNet [18] method of combining L1 and L2 regularizes:

$$\frac{1}{2|D|} \sum_{d \prec d'} \mathcal{L}(W(d', \alpha) - W(d, \alpha)) + \lambda \left((1 - \beta) \sum_{k=1}^K \alpha_k^2 + \beta \sum_{k=1}^K |\alpha_k| \right) \rightarrow \min_{\alpha}, \quad (12)$$

where β is a mixing parameter between ridge ($\beta = 0$) and lasso ($\beta = 1$), λ controls the regularization impact.

For \mathcal{L} function we consider three options:

- **Negative SE:** $\mathcal{L}(M) = -M^2$
- **Negative sigmoid:** $\mathcal{L}(M) = -\sigma(M)$, where $\sigma(x) = 1/(1 + \exp x)$ —sigmoid function
- **Negative AE:** $\mathcal{L}(M) = -|M|$

The results of testing all models above and the ensemble are described in the **Experiments** section.

7. Experiments

We tested every model and the ensemble trained on the dataset mentioned above. For all experiments, we used Wikipedia as a reference collection. The accuracy score was selected as a quality metric.

$$\text{accuracy}(c) = \frac{\sum_{d \prec d'} [c(d) < c(d')]}{|D|} \quad (13)$$

To validate the ensembles, we preliminarily split the dataset into train D_{train} and test D_{test} parts, so having 6K training objects and 2K testing.

7.1. Single models

We compare all aforementioned quantile-based models to various readability indexes and baselines proposed in [1], [2] and [19]. As for hyperparameters, we used $w_i = c_i/n$ (for text length not to affect the scores), $p = 1$, and $\gamma = 0.95$ for all models proposed. The results are exposed in **table 4**.

Table 4: Comparison of readability indexes performance to proposed models

Model Class	Model	Accuracy
Readability Indexes	Automated Readability Index	50.5%
	Flesch-Kincaid Grade	44.7%
	Gunning FOG	44.4%
	Flesch Reading Ease	50.7%
	Dale-Chall	37.0%
	Linsear Write	45.2%
	Coleman-Liau	52.1%
Morphological	Letter [1]	63.7%
	Syllables	70.9%
	Sorted Syllables	73.1%
Lexical	Length [2]	42.4%
	Distance	75.0%
	Counter	71.2%
Syntactic	Length	62.0%
	Syntagm	64.2%
Discursive	Connectors [19]	62.5%

The lexical distance model demonstrates the best performance in terms of accuracy among all the described models. Moreover, all quantile-based models, except for lexical distance one, outperform readability indexes. The sorted-syllables model performs better than unsorted, which proves the assumption about the sustainability of distributions in the sorted-syllables model.

7.2. Ensembles

Table 5: Comparison of ensembles with different margin functions to the best models on different language levels

Model	Margin Function	Accuracy
Coleman-Liau	—	52.1%
Morphological Sorted Syllables	—	73.1%
Lexical Distance	—	75.0%
Syntactic Syntagms	—	64.2%
Connectors	—	62.5%
Ensemble	Negative SE	88.1%
Ensemble	Negative sigmoid	84.6%
Ensemble	Negative AE	85.1%

To validate ensembles trained on D_{train} , we first evaluate all models on D_{rest} part of the dataset to get comparable results. In **table 5**, we compare the best models from all language levels with ensembles with various margin functions. We set the hyperparameters equal $\beta = 0.5$ and $\lambda = 10$ for all models.

As can be seen, Negative SE works best for fitting an ensemble, while all ensembles demonstrate quality growth compared to other models.

7.3. Noosphere Reference Collection

Here we explore the impact of the reference collection on the models' performance. We fitted the models with Noosphere corpora as a reference collection. This collection is less scientific and formal, featuring diverse literary works. We still evaluate the models on the labeled dataset, introduced in [Section 5](#). The results are exposed in [table 6](#).

Table 6: Comparison of models fitted on Noosphere reference collection

Model Class	Model	Accuracy
Morphological	Letter [1]	60.3%
	Syllables	69.2%
	Sorted Syllables	70.5%
Lexical	Length [2]	39.8%
	Distance	72.1%
	Counter	66.9%
Syntactic	Length	63.1%
	Syntgam	66.4%
Discursive	Connectors [19]	60.2%
Ensembles	Negative MSE	83.1%

All scores are lower, except for the syntactic models. There are understandable reasons for that. Firstly, the Noosphere collection is smaller than Wikipedia, resulting in less accurate empirical distribution estimations. Secondly, the collection consists of the non-scientific documents and does not contain specialized terms. Nevertheless, syntactic models improve their performance, mainly because of the absence of formulas in the reference collection.

Overall, the ensemble's accuracy is still higher than 80%, which outperforms both the readability indices and cognitive model baselines.

8. Conclusion

In conclusion, we presented new quantile-based models to measure cognitive text complexity. All models are based on psychophysiological assumptions. We explored models dealing with tokens from morphological, lexical, syntactic, and discursive language levels. All complexity scores are calculated with respect to the reference collection—a set of adequately simple documents used to obtain the empirical distributions of the token complexities. The reference collection should be chosen carefully and be large enough, but it gives high flexibility to the discussed approach. By varying the reference collection, we can obtain complexity scores concerning a particular domain. We introduced the way to measure the quality of the cognitive complexity

models, based on crowdsourcing. By ensembling models from various language levels, we attain an accuracy score of more than 88% and 83% using Russian Wikipedia and Noosphere reference collections, respectively. Suggested models outperform the readability indices and previously proposed cognitive complexity models.

9. Acknowledgements

This work is supported by the Russian Foundation for Basic Research, grant 20-07-00936.

References

1. *A. A. Birkin*: Speech Codes. Hippocrat, Saint-Peterburg, 2007.
2. *A. A. Birkin*: Nature of Speech. Likbez, Moscow, 2009.
3. *N. A. Chirkova.*: Additive regularization for hierarchical multimodal topic modeling. *Machine Learning and Data Analysis*, 2:187–200, 01 2016.
4. *Aryna Dzmitryieva*: The art of legal writing: A quantitative analysis of russian constitutional court rulings. *Sravnitel'noe konstitucionnoe obozrenie*, 3:125–133, 01 2017.
5. *Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare Voss, and Jiawei Han*: Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8, 06, 2014.
6. *M. A. Eremeev*: and Konstantin Vorontsov. Lexical quantile-based text complexity measure. In *RANLP*, 2019.
7. *R. Flesh*: How to test readability. New York, Harper and Brothers, 1951.
8. *Robert Gunning*: The technique of clear writing. McGraw-Hill, New York, 1952.
9. *Thomas Hofmann*: Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99*, pages 50–57, New York, NY, USA, 1999. ACM.
10. *J. Peter Kincaid, Robert P. Fishburne, Richard Lawrence Rogers, and Brad S. Chissom.*: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.
11. *William Lidwell, Kritina Holden, and Jill Butler*: Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design. 2010.
12. *Irina Osborneva*: Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov [semiautomatic evaluation of the complexity of academic texts on the base of statistic parameters]. 2006.
13. *R. J. Senter and E. A. Smith*: Automated readability index. *AMRL-TR*, 66(22), 1967.
14. *Valery Solovyev, Vladimir Ivanov, and Marina Solnyshkina*: Assessment of reading difficulty levels in russian academic texts: Approaches and metrics. *Journal of Intelligent and Fuzzy Systems*, 34:1–10, 04 2018.

15. *Milan Straka and Jana Strakova*: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. pages 88–99, 01 2017.
16. *K. V. Vorontsov and A. A. Potapenko*: Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*, 101(1):303–323, 2015.
17. *Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina*: Non-bayesian additive regularization for multimodal topic modeling of large collections. 10 2015.
18. *Hui Zou and Trevor Hastie*: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.
19. *В. М. Тютюнник, А. А. Биркин и Ю. Г. Гуцин*: Основы лингвистической психофизиологии. изд-во МИНЦ «Нобелистика» Тамбов; М.; СПб.; Баку; Вена; Гамбург, 2016.

SPARQL QUERY GENERATION FOR COMPLEX QUESTION ANSWERING WITH BERT AND BILSTM-BASED MODEL

Evseev D. A. (dmitrij.euseew@yandex.ru),
Arkhipov M. Yu. (arkhipov@yahoo.com)

Neural Networks and Deep Learning Lab, Moscow Institute of Physics and Technology, Moscow, Russia

In this paper we describe question answering system for answering of complex questions over Wikidata knowledge base. Unlike simple questions, which require extraction of single fact from the knowledge base, complex questions are based on more than one triplet and need logical or comparative reasoning. The proposed question answering system translates a natural language question into a query in SPARQL language, execution of which gives an answer. The system includes the models which define the SPARQL query template corresponding to the question and then fill the slots in the template with entities, relations and numerical values. For entity detection we use BERT-based sequence labelling model. Ranking of candidate relations is performed in two steps with BiLSTM and BERT-based models. The proposed models are the first solution for LC-QUAD2.0 dataset. The system is capable of answering complex questions which involve comparative or boolean reasoning.

Key words: knowledge base, complex question answering, query generation, entity detection, relation prediction

DOI: 10.28995/2075-7182-2020-19-270-282

ГЕНЕРАЦИЯ SPARQL-ЗАПРОСОВ ДЛЯ ОТВЕТА НА СЛОЖНЫЕ ВОПРОСЫ С ПОМОЩЬЮ BERT И BILSTM

Евсеев Д. А. (dmitrij.euseew@yandex.ru),
Архипов М. Ю. (arkhipov@yahoo.com)

Лаборатория нейронных систем и глубокого обучения,
Московский физико-технический институт (национальный
исследовательский университет), Москва, Россия

В данной работе описывается вопросно-ответная система для ответа на сложные вопросы по базе знаний Wikidata. В отличие от простых вопросов, для ответа на которые требуется найти один факт в базе

знаний, сложные вопросы требуют извлечения более 1 триплета, а также логические или сравнительные рассуждения. Предложенная система переводит вопрос на естественном языке в запрос на языке SPARQL, выполнение которого дает ответ. В состав системы входят модели, которые определяют шаблон SPARQL-запроса, соответствующего вопросу, и затем заполняют пустые места в шаблоне сущностями, отношениями и численными значениями. Для извлечения сущностей мы использовали модель маркировки последовательностей на основе BERT. Ранжирование возможных отношений для вопроса происходит в два этапа с помощью моделей на основе BiLSTM и BERT. Предложенные модели — первое решение для датасета LC-QUAD2.0. Система способна отвечать на вопросы, требующие сравнительное или логическое рассуждение.

Ключевые слова: база знаний, ответ на сложные вопросы, генерация запросов, извлечение сущностей, извлечение отношений

1. Introduction

Question answering has been an active area of research over past decades. Question answering systems can use two kinds of sources to find an answer: unstructured text corpora [11], [10], and knowledge bases (KB). KBs are an important source of information which integrates information from different sources [15]. Question answering models using KBs are compact and interpretable [16].

Knowledge base question answering (KBQA) requires matching of a subgraph with a question. If the question corresponds to a single triplet in a KB, the task is called simple question answering [1]. Complex question answering requires matching several triplets and logical, quantitative and comparative reasoning over knowledge graphs [13], [4].

One of the key approaches to complex question answering is SPARQL query generation. LC-QUAD [13] is a dataset with 5,000 questions and corresponding SPARQL queries over DBpedia, which involve logical and quantitative reasoning. LC-QUAD2.0¹ [4] is a dataset of 30,000 questions compatible with both DBpedia and Wikidata², which contains more types of SPARQL queries compared with previous version. The queries involve ranking of graph edges, boolean reasoning over more than one triplet and comparative constraints.

In this paper we describe models for SPARQL query generation trained on LC-QUAD2.0 dataset. For translation of a question to a SPARQL query, we first define the type of the query template. Then we fill the empty slots in the template with entities, relations from Wikidata and constraints. For entity detection we use BERT sequence labeling model. Relation ranking is performed by BiLSTM, path ranking—by BERT-based ranking model. We use pretrained cased 12-layer BERT-Base³. For extraction of compar-

¹ <https://github.com/AskNowQA/LC-QuAD2.0>

² <https://www.wikidata.org>

³ <https://github.com/google-research/bert>

ative constraints we use regular expressions. Our KBQA system is capable of answering complex questions with logical or comparative reasoning. The proposed KBQA system was released as a component of open-source DeepPavlov library⁴.

2. Related work

The first approaches to KBQA considered single-fact questions. Simple Questions [1] is the most widely used dataset for training models to answer single triplet questions. The model of [1] uses memory networks to store candidate facts and then score them by cosine similarity between question and fact vectors (each vector is a product of trainable embedding matrices and bag-of-ngrams representations of the question and fact). In [2] relations and entities in candidate triplets are separately ranked. Dot product of trainable relation embedding and vector representation of the question (final hidden states of BiGRU + linear layer) is used for scoring. Dot product of TransE entity embedding and vector representation of the question is used for entity scoring. Another approach is generation of the query with character-level encoder-decoder architecture [5]. Encoding of questions, entities and relation labels with BiGRU at word and character level is described in [7].

Decomposition of knowledge-base question answering into entity detection, linking, relation prediction and answer parsing components is a simple approach but it is competitive with more complicated architectures [14]. KBQA system proposed in our work consists of the similar steps and several other steps specific for complex questions. The approach of [14] utilises vanilla RNNs for entity detection and relation prediction. These subtasks of KBQA can be solved with BiLSTM and BiGRU [9] and improve accuracy of [14] on Simple Questions dataset.

Query building for complex question answering includes query generation and ranking steps. Model proposed by [8] generates candidate paths in the knowledge graph starting from extracted entities (entity detection and linking is omitted with the assumption that correct entities are given). The question and candidate paths are encoded with BiLSTM. Dot products of vectors representing the question and candidate paths are used to rank candidate paths. The approach of [18] uses Tree-LSTM which considers tree representations of candidate walks and the question with respect to the syntactical structure. Assuming that the lists of candidate entities and relations are given, Tree-LSTM produces latent representations of the question and candidate queries, which are ranked by the similarity function.

The model of [15] uses message passing for query ranking, which means propagation of confidence scores from candidate entities and relations to the adjacent nodes in the extracted subgraph. The model also includes entity and relation extraction steps. The substrings in the questions in LC-QUAD dataset, corresponding to entities and relations, are tagged “E1”, “E2”, “P1”, “P2”, “C1”, “C2”, which means “first entity” in the question, “second entity” (if exists), “first relation”, “second relation” (if exists), “class of first entity”, etc. BiLSTM + CRF network was trained for labeling of question tokens sequence with the corresponding tags. After entity and relation linking, for all

⁴ <https://github.com/deepmipt/DeepPavlov>

entities in the subgraph the confidence scores are aggregated from adjacent entities and the entity with the highest score is considered as the answer.

The work of [12] presents Complex Imperative Program Induction from Terminal Rewards, a model which can perform set, logical and arithmetic operations on the extracted subgraph assuming that the list of gold entities and relations is given. The query is generated with an imperative sequential program. Each step of the program selects the atomic operator and a set of previously defined variables (for example, entities and relations), and writes the result to memory, which is used in subsequent steps. The model achieves state-of-the-art performance for the Complex Sequential Question Answering dataset.

KBQA system, proposed in this work, can perform all the steps of complex question answering from entity extraction to query generation and is capable of answering to both simple questions and complex questions with boolean, quantitative and comparative reasoning from LC-QUAD2.0 dataset.

3. Overview of LC-QUAD2.0 dataset

Numbers of questions and percentage of the total number of questions for different query template types in train and test sets are shown in **Table 1**.

Table 1: Percentage of different query template types in the dataset

Query template type	Percentage of the total number of questions	Number of questions, train set	Number of questions, test set
statement_property	25.5	5,852	1,484
right-subgraph	15.6	3,574	854
center	14.0	3,220	824
Simple question left	7.0	1,604	438
Simple question right	6.5	1,494	378
string matching simple contains word	6.4	1,466	338
left-subgraph	6.2	1,418	373
boolean with filter	5.8	1,331	341
rank	4.0	921	210
string matching type + relation contains word	2.9	662	148
two intentions right subgraph	2.6	599	141
boolean double one_hop right subgraph	1.8	411	89
boolean one_hop right subgraph	1.7	399	101

“statement_property” are complex questions which deal with a numerical value or date as an answer or one of the entities (1).

(1) When did Jean-Paul Sartre move to Le Havre?

“center” are single-fact questions (one entity and one relation).

“simple question right” and “simple question left” are single-fact questions and the answer entity is connected with one of the entities in the question with the relation “P31” (“instance of”).

“left-subgraph” questions require finding paths in subgraph of the length of 2.

“right-subgraph” are questions with two entities and two relations.

“two intentions right subgraph” questions contain one entity and two relations and these questions have two answers corresponding to two facts about the grounding entity.

“boolean one_hop right subgraph” and “boolean double one_hop right subgraph” require determining whether one or two facts are true or false. To solve these questions we need to look for these facts in Wikidata and if the facts exist in the knowledge base, we consider the statement true, otherwise the statement is false.

“boolean with filter” questions require comparison of the object entity with the numerical value from the question.

“string matching simple contains word” and “string matching type + relation contains word” are questions where the answer entity should contain a particular letter or word.

“rank” questions require ordering of answer entities by ascending or descending.

4. Components of proposed KBQA system

4.1. KBQA pipeline

We decompose the task of KBQA into query template prediction, entity detection, entity linking, relation ranking, path ranking, constraint extraction (if the question has constraints) and generation of query from extracted entities, relations and constraints.

Let us consider as an example the steps of KBQA (Figure 1) for “statement property” question with SPARQL query template (2).

```
(2) SELECT ?obj WHERE { wd:Q1 p:P1 ?s . ?s ps:P1 ?obj .  
      ?s pq:P2 ?x filter(contains(?x, N)) }
```

On entity detection step we extract the entity substring S from the question. After entity linking step we obtain candidate entities E_1, \dots, E_N with corresponding confidences P_{E_1}, \dots, P_{E_N} . Then we extract relations R_1^1, \dots, R_M^1 , connected to entities E_1, \dots, E_N , rank them with BiLSTM ranking model, and leave 15 relations R_1^1, \dots, R_{15}^1 with maximal confidences $P_{R_1^1}, \dots, P_{R_{15}^1}$. Number N for the expression “filter(contains(?x, N))” is extracted from the question with regular expressions. Then we execute SPARQL queries (3)

(3) SELECT ?obj ?p2 WHERE { wd: E_i p: R_j ?s . ?s ps: R_j ?obj .
 ?s ?p2 ?x filter(contains(?x, N)) }

for combinations $\langle E_i, R_j \rangle$ of entities E_1, \dots, E_N and relations R_1^1, \dots, R_{15}^1 and obtain the list of candidate second relations R_1^2, \dots, R_K^2 . Combinations of relations $\langle R_j^1, R_k^2 \rangle$ are ranked with BERT-based ranking model. The model outputs confidences $P_{R_{jk}}$. Entity E_i and relations R_j and R_k with maximal confidences product $P_{E_i} \cdot P_{R_{jk}}$ are filled in the slots of the SPARQL query template (4):

(4) SELECT ?obj WHERE { wd: E_i p: R_j^1 ?s . ?s ps: R_j^1 ?obj . ?s pq: R_k^2
 ?x filter(contains(?x, N)) }

Other types of questions are processed similarly: first we find candidate entities, then extract and rank candidate relations with BiLSTM, if necessary, extract numerical values with regular expressions, find valid combinations of entities and relations according to the query template, rank combinations of relations with BERT and consider combination of entities and relations with maximal product of confidences as the required query.

4.2. Classification of questions by query template type

Query template types “right-subgraph”, “simple question right”, “simple question left”, “left-subgraph”, “center” we united into one class.

“statement_property” questions can be translated into 5 types of SPARQL queries, “rank” questions—into 2 types. Each type is considered as a separate class. All other types of questions are put into a separate class. Total number of classes is 14.

Classification of questions is performed with BERT-based model from DeepPavlov library. Output representation of BERT [CLS] token is fed into a dense layer for classification into 14 classes. For comparison we used tf-idf+SVC model (Table 2).

Table 2: Accuracy of question classification by query template types

BERT	TF-IDF+SVC
90.8	85.5

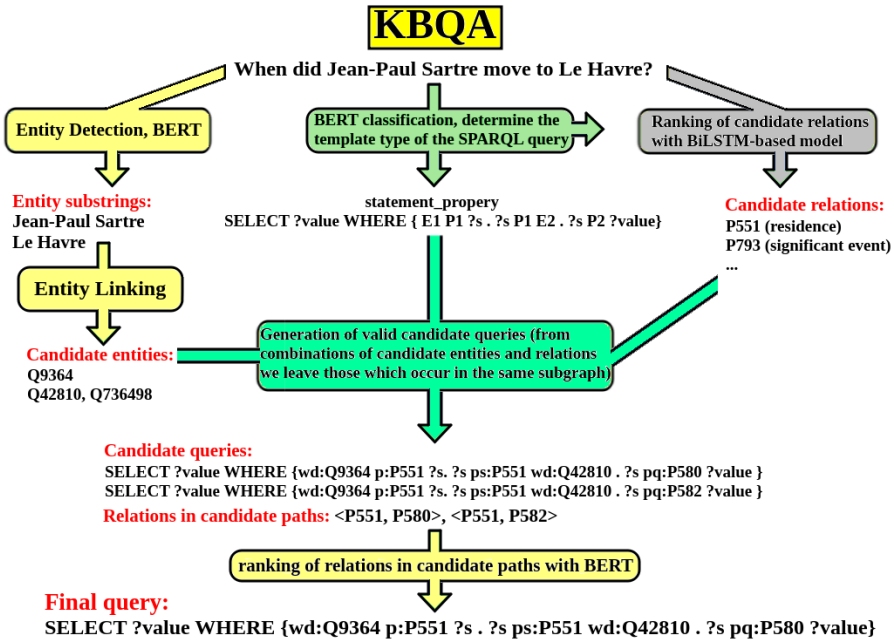


Figure 1: KBQA pipeline

4.3. Entity Detection and Entity Linking

Entity Detection is implemented as labeling of sequence of question tokens

$$q_{seq} = \{w_1, w_2, \dots, w_n\} \tag{1}$$

with one of two labels: “I-TAG” if the token w_i is in the entity substring and “O-TAG” otherwise. For example, in question

(5) When did Jean-Paul Sartre move to Le Havre?

tokens “Jean-Paul”, “Sartre”, “Le”, “Havre” are labelled with “I-TAG”, the other tokens with “O-TAG”. We prepared the dataset from LC-QUAD2.0 for Entity Detection using labels of gold entities to find substrings in questions, corresponding to entities and annotated matched tokens as “I-TAG”. This dataset is used for training of BERT-based sequence labeling model from DeepPavlov library. Output representations of question sub-tokens are fed into a dense layer for classification of sub-tokens into classes, corresponding to two tags (Figure 2). We obtained F1-score of 87 on test-set.

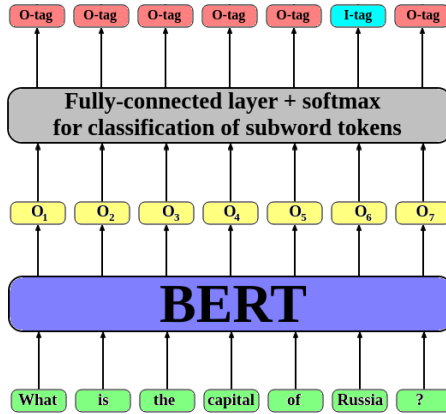


Figure 2: BERT for sequence tagging

For all entities in Wikidata we built an inverted index over unigrams in entity’s label (a dictionary where keys are tokens and values are lists of entities containing these tokens). Entity Linking is implemented using fuzzy matching of the string extracted at Entity Detection step with inverted index. For example, tokens “Jean-Paul” and “Sartre” from the substring “Jean-Paul Sartre” are used as keys to obtain the list of candidate entities, and candidate entity “Q9364” with the label “Jean-Paul Sartre” has the maximum fuzz ratio of 100. Candidate entities, extracted from inverted index dictionary, are ranked by fuzz ratio of their titles with the entity substring and number of relations (the more relations an entity has, the more popular it is).

4.4. Model of relation ranking

The model of relation ranking is inspired by [17] (Figure 3). The sequence of question tokens q_1, \dots, q_n is passed through an embedding lookup layer. The sequence of Word2Vec embeddings e_1, \dots, e_n is the input of 2-layer BiLSTM to encode the token sequence with hidden representations h_1, \dots, h_n . For linked entities we extract all relations from Wikidata which these entities have and consider them as candidate relations. Candidate relations are encoded with PyTorch-BigGraph embeddings [6] $rel_emb_1, \dots, rel_emb_k$. Dot products of each candidate relation embedding and hidden states $rel_emb_i \cdot h_1, \dots, rel_emb_i \cdot h_n$ are passed through softmax layer to obtain coefficients $\alpha_1, \dots, \alpha_n$. Then we sum hidden states weighted with coefficients:

$$q = \sum_{j=1}^n \alpha_j \cdot h_j \quad (2)$$

The model is trained to maximize dot product $q \cdot rel_emb_i$ if rel_emb_i is the embedding of the right relation and minimize if rel_emb_i is the embedding of the wrong relation.

$q \cdot rel_emb_i$ is the confidence that rel_emb_i is the right relation. For example, for the question

(6) What periodical literature does Delta Air Lines use as a mouthpiece?

with the corresponding SPARQL query

```
(7) SELECT DISTINCT ?obj WHERE wd:Q188920 wdt:P2813 ?obj . ?obj wdt:P31 wd:Q1002697
```

the model is trained to output maximal dot product for embeddings of relations P2813 and P31.

For every question Q_k in test set we extracted candidate relations R_1^k, \dots, R_n^k for gold entities E_1^k, \dots, E_m^k . Candidate relations are ranked with relation ranking model and we check if candidate relation with the maximum score is one of the gold relations $R_{1g}^k, \dots, R_{mg}^k$. We measure the percent of questions in test set which have one of the gold relations ranked with the highest score (84% of questions) (Table 3). The model is more accurate if PyTorch-BigGraph embeddings of relations are replaced with average embeddings of relation title tokens (89%).

Table 3: Percent of questions in test set with one of the gold relations ranked with highest score

Relation embeddings used in the model	% of questions
PyTorch-BigGraph	84
Word2Vec	89

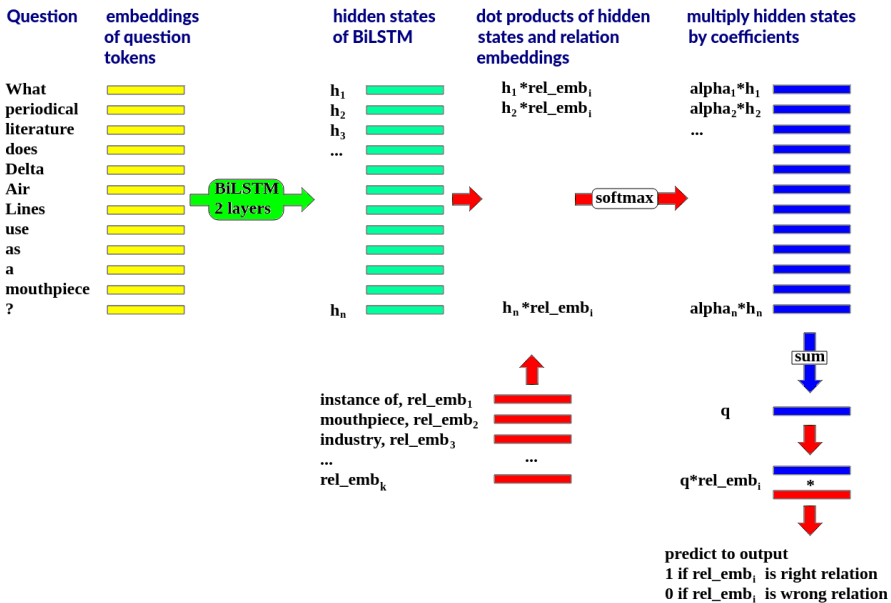


Figure 3: Relation ranking network

4.5. BERT for path ranking

Path ranking model is inspired by [8]. The input to the model is the following: the question q followed by [SEP] token and candidate path

$$C_i = \{R_1, \dots, R_L\}, L \in \{1, 2\} \quad (3)$$

from the set of candidate paths C_1, \dots, C_n . For example, one of the candidate paths for the question

(8) What is stable version of user interface of Amazon Kindle?

is $\{P1414, P348\}$, where the label of relation with identifier in Wikidata “P1414” is “GUI toolkit or framework” and the label of “P348” is “software version”. So, the input to BERT is the question q and relation titles “GUI toolkit or framework” and “software version” (Figure 4).

[CLS] what is stable version of amazon kind ##le ? [SEP] g ##ui tool ##ki ##t or framework [SEP] software version [SEP]

Figure 4: BERT input representation

Output representation of BERT [CLS] token is fed into a dense layer for binary classification into 2 classes: 1 if the candidate path is the gold path for the question (positive sample) and 0 otherwise (negative sample). For training of the model we generated negative samples in the ratio of 20:1 to positive samples. The model achieves F1 of 87.2 on the test set.

4.6. Using regular expressions

Regular expressions are used in “statement_property” questions for extraction of dates and numerical values. “rank” questions require determination of the order of answer ranking (ascending or descending). For example, words “What is the highest”, “the biggest”, “the longest”, etc. point at ascending order (corresponding to “ORDER BY ASC(?obj)” in the SPARQL query) and “the smallest”, “the lowest”, etc. point at descending order (“ORDER BY DESC(?obj)”). Such keywords are extracted with regular expressions. In “boolean with filter” questions regular expressions are used for extraction of numerical values and comparison operators. For example, in the question

(9) Is the maximum wavelength of sensitivity of the human eye equal to 700?

the numerical value is “700” and “equal to” corresponds to “=”. So the SPARQL query for the question is

(10) ASK WHERE wd:Q430024 wdt:P3737 ?obj filter(?obj = 700)

Table 4: Question answering accuracy

Query template type	Answering accuracy
statement_property	51.5
right_subgraph	33.3
center	78.1
Simple question left	67.3
Simple question right	68.7
string matching simple contains word	80.4
left-subgraph	27.9
boolean with filter	75.9
rank	48.5
string matching type + relation contains word	46.9
two intentions right subgraph	43.4
boolean double one_hop right subgraph	63.8
boolean one-hop right subgraph	59.1
Total	56.3

4.7. Results of the KBQA system on LC-QUAD2.0 dataset

Question answering accuracy for different types of questions is shown in **Table 4**. The answer is considered correct if the answer entities and numerical values or dates match with gold answers. The proposed KBQA system gives correct answers to almost one-half of double-fact questions with numerical values or dates (“statement_property”) and questions with ranking of answers (“rank”). The system achieves high scores on single-fact questions (“center”, “boolean with filter”, “string matching simple contains word”). Two-hop questions (“left-subgraph” and “right-subgraph”) present difficulties to the system and are the subject of further research and improvement of the model.

4.8. Results of the KBQA system on LC-QUAD1.0 dataset

We divide questions in LC-QUAD1.0 into the following types: simple (one entity and one relation in the SPARQL query), simple with type (one entity, one relation and entity, which defines the type of answer entities), double (two entities and two relations), 2-hop (one entity and two relations) and boolean (the SPARQL query contains two entities, one relation and “ASK WHERE” keywords). BERT-based model is used for classification of the question by 5 query template types. Extraction of keywords, such as “how many”, “count”, etc. is used to define whether the question requires counting of number of answer entities.

The other details of the solution are the same as in KBQA system for LC-QUAD2.0, excepting additional tag “T-tag” in BERT sequence labeling model for extraction of substrings corresponding to the type of the entity.

Our model outperforms QAmP [15] and WQAqua [3]⁵. We did not compare our model with [8], because their work does not consider entity detection and linking steps.

⁵ <http://lc-quad.sda.tech/lcquad1.0.html>

Table 5: Question answering accuracy

System	Precision	Recall	F1 score
Our model	0.60	0.66	0.63
QAmp	0.25	0.50	0.33
WQAqua	0.22	0.38	0.28

5. Conclusion

In this work, we have described question answering system over Wikidata knowledge base. The system translates a natural language question into a query in SPARQL language, execution of which gives an answer. The proposed KBQA system is capable of answering complex questions which require logical or comparative reasoning. The system is the first solution to LC-QUAD2.0 dataset, and we evaluated the performance of the system for different types of questions in LC-QUAD2.0.

References

1. *Bordes, A. et al.*: Large-scale simple question answering with memory networks. CoRR. abs/1506.02075, (2015).
2. *Dai, Z. et al.*: CFO: Conditional focused neural question answering with large-scale knowledge bases. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 800–810 Association for Computational Linguistics, Berlin, Germany (2016).
3. *Diefenbach, D. et al.*: Towards a question answering system over the semantic web. CoRR. abs/1803.00832, (2018).
4. *Dubey, M. et al.*: Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In: International semantic web conference. pp. 69–78, Springer (2019).
5. *He, X., Golub, D.*: Character-level question answering with attention. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 1598–1607, Association for Computational Linguistics, Austin, Texas (2016).
6. *Lerer, A. et al.*: Pytorch-biggraph: A large-scale graph embedding system. arXiv preprint arXiv:1903.12287. (2019).
7. *Lukovnikov, D. et al.*: Neural network-based question answering over knowledge graphs on word and character level. In: Proceedings of the 26th international conference on world wide web. pp. 1211–1220, International World Wide Web Conferences Steering Committee, Republic; Canton of Geneva, CHE (2017).
8. *Maheshwari, G. et al.*: Learning to rank query graphs for complex question answering over knowledge graphs. In: International semantic web conference. pp. 487–504, Springer (2019).

9. *Mohammed, S. et al.*: Strong baselines for simple question answering over knowledge graphs with and without neural networks. In: Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers). pp. 291–296, Association for Computational Linguistics, New Orleans, Louisiana (2018).
10. *Rajpurkar, P. et al.*: Know what you don't know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822. (2018).
11. *Rajpurkar, P. et al.*: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 2383–2392, Association for Computational Linguistics, Austin, Texas (2016).
12. *Saha, A. et al.*: Complex program induction for querying knowledge bases in the absence of gold programs. Transactions of the Association for Computational Linguistics. 7, 185–200 (2019).
13. *Trivedi, P. et al.*: Lc-quad: A corpus for complex question answering over knowledge graphs. In: International semantic web conference. pp. 210–218, Springer (2017).
14. *Ture, F., Jojic, O.*: No need to pay attention: Simple recurrent neural networks work! In: Proceedings of the 2017 conference on empirical methods in natural language processing. pp. 2866–2872, Association for Computational Linguistics, Copenhagen, Denmark (2017).
15. *Vakulenko, S. et al.*: Message passing for complex question answering over knowledge graphs. In: Proceedings of the 28th acm international conference on information and knowledge management. pp. 1431–1440 (2019).
16. *Wilcke, X. et al.*: The knowledge graph as the default data model for learning on heterogeneous knowledge. Data Science. 1, 1–2, 39–57 (2017).
17. *Xiong, W. et al.*: Improving question answering over incomplete kbs with knowledge-aware reader. arXiv preprint arXiv:1905.07098. (2019).
18. *Zafar, H. et al.*: Formal query generation for question answering over knowledge bases. In: European semantic web conference. pp. 714–728, Springer (2018).

COMBINING FACTS, SEMANTIC ROLES AND SENTIMENT LEXICON IN A GENERATIVE MODEL FOR OPINION MINING

Feldman D. G. (daniil.feldman@phystech.edu)

Moscow Institute of Physics and Technology, Moscow, Russia

Sadekova T. R. (str12.01.94@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

Vorontsov K. V. (vokov@forecsys.ru)

Moscow Institute of Physics and Technology, Moscow, Russia

Opinion mining is a popular task, that is applied, for example, to determine news polarisation and identify product review classes. Our task is unsupervised clusterization of opinionated texts, in particular news on political events. Many papers that tackle this issue use generative models based on lexical features. Our goal is to determine the entities defying an opinion amongst lexical, syntactic and semantic features as well as their compositions. More specifically, we test the hypothesis that an opinion is determined by the composition of the mentioned facts (SPO triples), the semantic roles of the words and the sentiment lexicon used in it. In this paper we formalise this task and prove that using a composition of the above features provides the best quality when clusterising opinionated texts. To test this hypothesis we have gathered and labelled two corpuses of news on political events and proposed a set of unsupervised algorithms for extracting the features.

Keywords: opinion mining, SPO triplets, Fillmore roles, sentiment lexicon

DOI: 10.28995/2075-7182-2020-19-283-298

КОМБИНИРОВАНИЕ ФАКТОВ, СЕМАНТИЧЕСКИХ РОЛЕЙ И ТОНАЛЬНЫХ СЛОВ В ГЕНЕРАТИВНОЙ МОДЕЛИ ДЛЯ ПОИСКА МНЕНИЙ

Фельдман Д. Г. (daniil.feldman@phystech.edu)

Московский физико-технический институт, Москва, Россия

Садекова Т. Р. (str12.01.94@gmail.com)

МГУ имени М. В. Ломоносова, Москва, Россия

Воронцов К. В. (vokov@forecsys.ru)

Московский физико-технический институт, Москва, Россия

1. Introduction

Every important political event is vastly covered in the news. Most sources provide polarised texts expressing the opinion of one of the sides. As a result, readers only get to know one side of the problem. We would like to provide them with a tool to clusterize news feed according to the opinions they express. Often there are more than two opinions on an issue, and a more general approach suggests finding the number of opinions as well. We, however, will be solving a simpler problem where this variable is given.

In this paper we study different approaches to classify opinionated texts on a given news story. We work with datasets of news covering political events. The question we answer is what computationally seizable entities define the opinion of a text. In contrast to the work done traditionally in opinion mining, which focuses on proposing a more complex generative model, in this work, we focus on determining the best features for classification and finding the entities that authors use to express opinions.

Most studies use lexical features (words or sentiment lexicon) as tokens and extract dependencies on a word-level. We assume that syntactic and semantic patterns combined capture the essence of the author's opinion. To be more specific, we propose three entities: subject-predicate-object (SPO) triples, semantic roles and sentiment lexicon. To understand how they capture opinions that are indistinguishable on a word-level, let us consider two extracts from actual news covering the enterprise nationalisation in LPR and DPR (Lugansk and Donetsk People's Republics). These texts have similar word distributions but the mentioned semantic and syntactic features are used differently.

...Президент Петр Порошенко заявил, что Россия де-факто конфисковала украинские предприятия, которые находятся на неподконтрольной Киеву территории. Сегодня ДНР и ЛНР «национализировали» украинские предприятия ... При этом Кремль защитил конфискацию предприятий в ЛДНР ... Украина требует расширить санкции ... За все эти действия обязательно наступит наказание. Украина потребует расширения санкций на тех, кто украл украинские предприятия ...

...По словам Захарченко, Киев встретит свой «ужасный конец»... Киев возьмется за ум, и в целях спасения собственной промышленности снимет блокаду... Обстановка, которую искусственно создала Украина с блокадой Донбасса, вынудила ... кошмарит свой народ ... если в Киеве были приняты какое-либо постановление ... положительные результаты, как для республик, так и для России ... Если им удастся сместить Порошенко и при этом не развалить Украину, то все вернется на свои места...

The words *Poroshenko*, *Russia*, *Ukraine* and others are used in both texts, so we wouldn't be able to tell the opinions apart based on lexical features. One way to solve this is to use subject-predicate-object triples. For example, in the first text *Poroshenko* is used as a subject, but in the second one as an object. The second approach we mentioned is using semantic roles. For example, the word *Russia* is an agent in the first text and a patient in the second one. Lastly, we use sentiment lexicon to tell the opinions apart: the first text uses such negative lexicon as *occupation*, *steal* (aggressor) while the second one has *hostage*, *threat* (victim). To clarify, word distributions do matter but they differ mostly in sentiment lexicon. In this paper we study what feature: SPO triples, semantic roles or sentiment lexicon captures the differences between opinions more accurately. In [section 3](#) we describe these features and the unsupervised algorithms we use to extract them in detail. In [section 4.1](#) we describe the probabilistic models we use for opinion clusterization. In [section 5](#) we present the openly available datasets of news labelled with opinions we have gathered. [Section 5.3](#) contains comparisons of models built on SPO triples, semantic roles, sentiment lexicon and their combinations. Finally, we present our conclusion that an opinion, in fact, is best defined as a combination of all three features.

2. Related work

Opinion mining has been vastly studied in recent years. A general survey of methods is presented in [\[2\]](#) and [\[3\]](#). Earlier works [\[1\]](#), [\[4\]](#) focused on mining opinions in product reviews, but in more recent papers focus has shifted towards texts on political events and the issue of political polarisation [\[5\]](#), [\[6\]](#), [\[12\]](#). Most of the discussed works rely on probabilistic models, which are described in [\[2\]](#). Topic models were used both in supervised [\[5\]](#) and unsupervised approaches [\[6\]](#). When analysing news feed, supervised approaches are hard to apply as big datasets with labelled texts are difficult to come by and cannot be gathered while the news are still relevant, so a supervised approach was not an option for us. All in all, unsupervised generative models are the most preferred and popular in opinion mining [\[5\]](#), [\[4\]](#), [\[14\]](#). Some works tackle a more

general problem of finding topics and opinions simultaneously. The authors of [7] propose a topic-aspect model that mines topics and aspects, where the latter can be interpreted as opinions. We will be focusing on a more localised problem of clustering opinionated texts on a given political event or topic. In other words, we will be building the second layer of a hierarchical topic model of the news feed, which corresponds to opinions. Unsurprisingly, the overwhelming majority of opinion mining studies worked with English texts, while we will be experimenting with Russian news.

SPO triples (or facts) have been used for opinion mining, and they allowed to increase quality in similar problems. For example, the study [8] was aimed at solving a different task: ontology mining. Its authors built a similar generative model (LDA). They proved that using SPO triples to build hierarchical topic models provides a gain in quality. That gave us the idea that a similar heuristic could be useful in opinion mining.

Sentiment analysis is close to opinion mining. Sentiment lexicon is a strong indicator of an opinion and can be accurately extracted without a train dataset, making it a popular tool for opinion mining. Many studies use polarised words to classify product reviews and political texts. The most basic approach is lexicon-based, but its main drawback is that context can massively affect word polarity. A necessary enhancement is rule-based approaches [16] that take into account neighbouring words, their part of speech, sentence role and other features. Works such as [17] propose a system based on a large set of heuristical rules, and our model follows a similar approach.

Semantic roles are powerful in capturing connections in sentences and they have been used for solving many NLP tasks such as question answering, information extraction and information search. Study [15] showed that they can provide a significant quality increase in opinion mining. There are several approaches for extracting semantic roles, most of them use some type of neural network architecture [19], [20], [21] and a hand-accessed database of semantic frames, such as FrameNet or VerbNet [20], [21], for training. A model for extracting semantic roles in Russian was proposed by Shelmanov and Devyatkin in [18]. It used a similar neural networks approach and the Russian database FrameBank, similar to FrameNet.

SPO triples and semantic roles have been used for opinion mining. However, a combination of semantic and syntactic features have not.

To build topic models we will be following an approach called additive topic modelling regularizers (ARTM), a technique that allows to configure different topic models by adding regularizers. It is described in detail and analysed in [9].

3. Features

To build the composite probabilistic model we will be using several syntactic and semantic features. This section will provide their description and describe the unsupervised algorithms we proposed to extract them.

3.1. SPO triples

In subject-predicate-object triples predicates are the words that express situations, they are usually verbs or verbal nouns. Naturally, in order to extract SPO triples,

we would first have to find predicates and then find subjects and objects connected to them. In order to do that we will be building syntactic dependency

trees, an example for the sentence «Радикалы зажгли файеры возле украинской дочки Сбербанка» is given in Figure 1. At its nodes are words connected with edges representing syntactic tags in the Universal dependencies format.

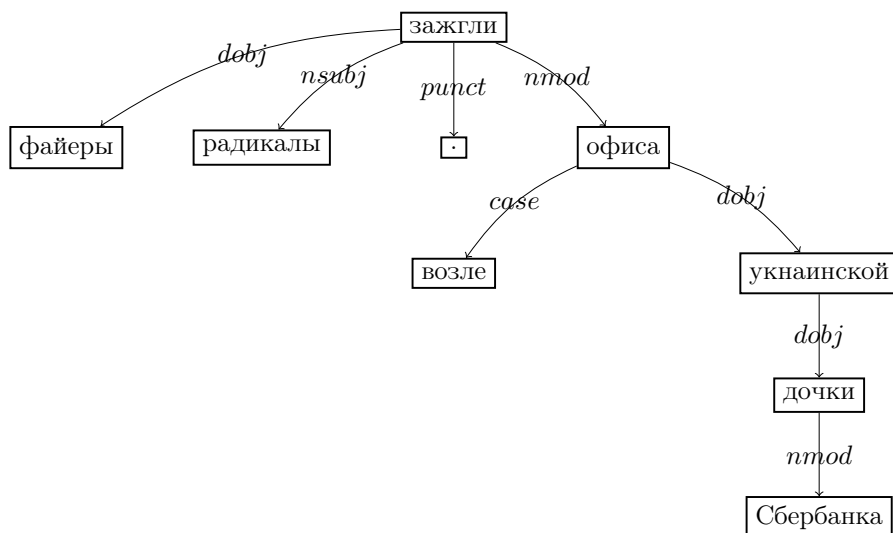


Figure 1: Dependency tree example

To extract the triplets we take into account a number of features such as part of speech, position, syntax tag and others. The most basic SPO triplets are noun-predicate-noun. Besides explicit triplets there are some implicit examples that do not contain a verb or a verb noun, such as noun-noun triplets.

We developed a set of heuristics allowing us to capture the following types of triplets:

- Explicit triplets: noun-verb-noun.
Example: *the congress passed a law* → (congress, pass, law)
- Noun-noun triplets.
Example: *president Putin* → (Putin, is, president)
- Adverb triplets.
Example: *The meeting held by Navalny...* → (Navalny, hold, meeting)
- Adjectives triplets.
Example: *Ukrainian branch* → (branch, is, Ukrainian)

To build the dependency trees we used a dockerized version of Google's SyntaxNet parser¹ pre-trained on a corpus of Russian texts.

¹ <https://github.com/tensorflow/models/tree/master/syntaxnet>

3.2. Fillmore roles

Semantic roles capture the meaning of words and show what the author meant to express. They are a way do describe numerous word relations by a limited set of states. There are many different sets of semantic roles of various sizes. On one end of the spectrum are domain-specific roles such as From_Airport, To_Airport, Depart_Date, they are fixed to a frame (Flight) and often to a small set of words. They are not suitable as features as their distributions in any given text are very sparse. On the other end of the spectrum are macro-roles: Proto_Agent and Proto_Patient, they are too general and usually match subjects and objects. In between lay many sets of semantic roles such as Fillmore’s nine: Agent, Experiencer, Instrument, Goal, Location, Object, Source, Time, and Path. Their description is given in **Table 1**. We decided to choose Fillmore’s [13] set of roles as they are cross-domain and describe complex semantic relations.

To extract Fillmore roles we will be using the model proposed by Shelmanov A. O. and Devyatkin D. A. in 2018 [18]. It uses a neural network using syntactic features gained from a dependency tree (built with SyntaxNet). The model extracts multiple semantic roles which we then clusterise into Fillmore’s 9 roles.

Table 1: Fillmore’s semantic roles

Role	Description
Agent	The instigator of the action denoted by the predicate
Patient	The ‘undergoer’ of the action or event denoted by the predicate
Theme	The entity that is moved by the action or event denoted by the predicate
Experiencer	The living entity that experiences the action or event denoted by the predicate
Goal	The location or entity in the direction of which something moves
Benefactive	The entity that benefits from the action or event denoted by the predicate
Source	The location or entity from which something moves
Instrument	The medium by which the action or event is carried out
Locative	The specification of the place where the action or event is situated

3.3. Sentiment lexicon

Opinions on an object is often supported by sentiment lexicon, and under sentiment we will be understanding two polarities: positive and negative. The problem of extracting sentiment lexicon in news can be broken down into two tasks: gathering a dictionary with sentiment lexicon and tagging words with respect to their context in the texts with +1, -1 or 0.

As a basis dictionary we used the one collectd by the resource Linis Crowd ([10]), which was gained through crowd-accessing of texts on political and social topics. The dictionary contains 2,454 words tagged with sentiment. We then enriched the dictionary using an approach that suggests that synonyms and hyponyms of a sentiment

word have the same polarity while antonyms have the opposite. For that we used Ru-WordNet² [11] and widened the dictionary to 3,419 words.

When tagging words with sentiments we took their context into account. There are several ways to do that but the best one is syntactic dependency trees. We used the following rules to tag words:

- If a tagged word is a noun, an adjective or an adverb, its parents are tagged with the same polarity.
- If a verb is tagged, subjects and objects connected to it are tagged as well.
- If a negative particle is a child of a tagged word, its polarity is changed to the opposite.

4. Multimodal opinion mining model

Our algorithm consists of several steps. Firstly, we build syntactic dependency trees of each sentence in the document corpus using SyntaxNet. The same trees are then used in all three algorithms that extract SPO triples, Fillmore roles and sentiment lexicon. The features are then broken down into 5 modalities, section 4.2 covers the process. Those modalities are used to built a regularised topic model. Finally, we train the topic model on a corpus of documents and split them into clusters corresponding opinions. Figure 2 shows a general plan of the algorithm. Next we will be describing every step in detail.

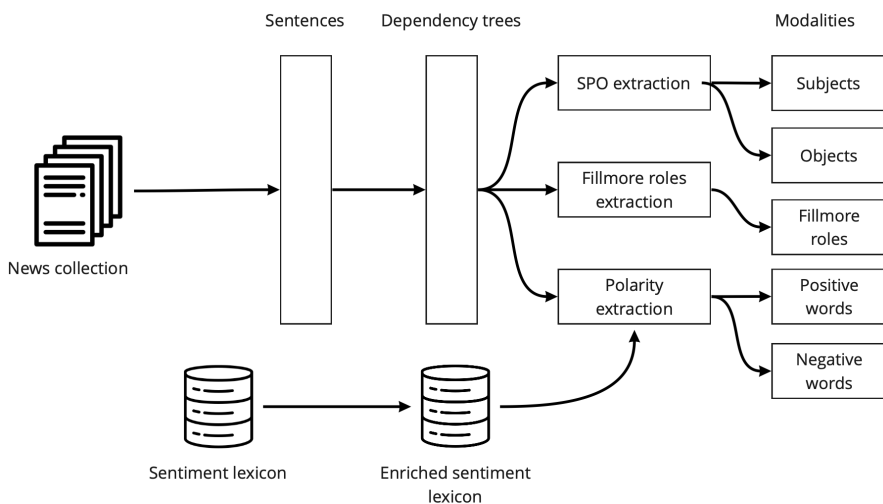


Figure 2: Modalities extraction algorithm

² <http://www.ruwordnet.ru/ru>

4.1. ARTM topic model

In the current section we will describe the basics of additive regularization of topic models. Let D be the corpus of documents and W the set of words (tokens) in them. We will consider every document a bag of words. We assume that each word relates to some topic from T and the corpus is an i.i.d. $(w_i; d_i; t_i)_{i=1}^n$ from a distribution $p(w; d; t) \in W \times D \times T$. Having also assumed that the appearance of a word w in a document depends only on the topic, we can draw:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td} = \Phi \cdot \Theta$$

Φ and Θ are stochastic topic-word and document-topic matrices, our goal is to find them. To do that we will be maximizing the likelihood logarithm with probability distribution constraints:

$$L(\Phi, \Theta) + R(\Phi, \Theta) = \ln \prod_{d \in D} \prod_{w \in d} p(w|d)^{n_{dw}} + R(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\Phi, \Theta)$$

Instead of words we can break the documents down into of several types of tokens called modalities (i.e. subjects, objects, polarised words). In this case the equation takes the form:

$$L(\Phi, \Theta) + R(\Phi, \Theta) = \sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\Phi, \Theta)$$

Here M is the list of all modalities. W^m is the dictionary of m-th modality. By adjusting weights τ_m we control the influence of each modality. The resulting optimisation task is:

$$\begin{aligned} \min_{\Phi, \Theta} & \sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\Phi, \Theta) \\ \text{s.t.} & \sum_{w \in W^m} \varphi_{wt} = 1, m \in M; \varphi_{wt} \geq 0 \\ & \sum_{t \in T} \theta_{td} = 1, m \in M; \theta_{td} \geq 0 \end{aligned}$$

Here $R(\Phi, \Theta)$ is the regulariser we add to the likelihood. If $R(\Phi, \Theta) = 0$ it turns into a PLSA model. In this case the problem has an infinite number of solutions, so regularisers provide additional constraints. $R(\Phi, \Theta)$ is usually a sum of regularisers with coefficients: $\sum \tau_i R_i(\Phi, \Theta)$. We will be using the following regularizers in our work:

- Smooth/sparse regularizer:

We presume that every document and every word relates to a small number of topics, so the distributions $\varphi_t = (\varphi_{wt})_{w \in W}$; $\theta_d = (\theta_{td})_{t \in T}$ should be sparse. At the same time there are some common vocabulary topics that are present in every document, their distributions are smooth. Common vocabulary does not carry much information for our experiments, so we add a smooth regulariser for

those words to be gathered in common vocabulary topics. Other topics will then have sparse distributions where only a few words are assigned high probabilities. Those words can be called kernel words. The model we build has one extra topic that holds common vocabulary terms.

- Decorrelation regulariser:

In opinion mining it is necessary that the opinions found in the text corpus differ significantly, otherwise we cannot classify the texts. We add the decorrelation regulariser so that is sparse.

4.2. Modalities

We will be building probabilistic models based on five modalities: $M = \{M_s, M_o, M_r, M_p, M_n\}$. The first two modalities are subjects and objects, derived from SPO triples. At first we used the whole triples as a modality, but the distribution of terms over texts in any given corpus was too sparse: most triples were unique to the collection. Usage of SPO triples for opinion mining was based on the heuristic that news with similar lexicon had differing distributions of subjects and objects, that is why we decided to use them as the two modalities. Let $T_d = \{(s_1, p_1, o_1), \dots, (s_n, p_n, o_n)\}$ be the list of all triplets in document d , then we create two lists $\{s_1, \dots, s_{n_s}\} \in W^s$ and $\{o_1, \dots, o_{n_o}\} \in W^o$ and define n_{dw} as the number of triples containing w :

$$n_{dw_s} = \sum_{(s,p,o) \in T_d} [s = w], w \in W^s$$

$$n_{dw_o} = \sum_{(s,p,o) \in T_d} [o = w], w \in W^o$$

The next modality M_r corresponds to roles. Its dictionary W^r contains all pairs (w, r) , $w \in W$, $r \in R$, where W is the corpus vocabulary and R is the set of nine Fillmore roles. Here n_{dw} is calculated as the number of times the pair (w, r) occurs in document d . The last two modalities M_p and M_n correspond to sentiment lexicon. W^p is a dictionary that contains all positively polarised words while W^n contains all negatively polarised words.

4.3. Assigning opinions

Having obtained the matrix θ with topic distributions for every document we have to assign some opinion to every text, that is to clusterise them. In order to do that we will consider θ_d as a vector of features for d . We can say that feature x_i shows “how much” of topic t is contained in d . If we assume that every text expresses a single opinion, it can be defined as

$$o_d = \arg \max_t \theta_d$$

It must be noted that an opinion does not necessarily express a positive or negative attitude to an entity, as there can be any number of opinions in a text corpora. It is rather a combination attitudes on different entities.

5. Experiments

To evaluate our algorithm we have collected and labelled two corpuses of news:

1. 82 news considering enterprise nationalisation in LPR and DPR. The texts average at 200 words. They were extracted from multiple news sources: Russian as well as Ukrainian. We selected texts expressing two opinions: Moscow's opinion and Kiev's opinion, most texts on the topic belong to one of them.
2. 220 news considering Donald Trump's decision of quitting the Paris Climate Agreement. The text's sizes once again averaged at around 200 words. The news were equally distributed between two opinions: one of Trump's supporters, those who oppose him (such as Elon Musk).

The corpuses were labelled by two independent accessors who were given selected news on a topic. They first read the whole collection and determined the number of opinions present. Then every text was marked with an opinion. The two accessors reached an agreement with an average rate of 91% when assigning opinions. An additional accessor was brought to evaluate uncertain texts. When collected, the corpuses were being extended until news began to be derivative from those already added. The datasets are available for public use at [our repository](#)³. From here on we will be referring to the datasets as Corpus 1 and Corpus 2.

To evaluate clusterisation quality over two classes we will be measuring precision, recall and F1-score.

5.1. Adjusting hyperparameters

Our model has several hyperparameters:

- Weights of modalities $\{\tau_m\}$, $m \in \{s, o, r, p, n\}$, $\sum_m \tau_m = 1$
This is the most important parameter—the weight distribution of modalities that defines an opinion. One of the goals of our research is to get the optimal distribution of features: SPO triples, semantic roles and sentiment lexicon that provides the best clusterisation quality and thus defines an opinion.
- Regularization coefficients τ
These coefficients determine how sparse or smooth the resulting distributions in Φ and θ are.
- Minimal TF for modalities
It is possible to take into account words with term frequency above a threshold.

To find the optimal hyperparameters we follow the steps:

1. Fix the regularization parameter at $\tau = 1.0$ and adjust the minimal TF for all modalities. We start with this hyperparameter as it takes only a small discrete set of values.

³ https://github.com/newfteddy/opinion_mining_features/tree/master/data

2. With fixed minimal TF optimise the regularization coefficients for each modality. τ has a noticeable effect on overall performance.
3. Find the optimal modalities weights $\tau_m, m \in \{s, o, r, p, n\}$ with all other parameters fixed.

The first step yielded the optimal minimal TF for filtering the dictionaries M_i : for semantic roles and SPO triplets it is 2, for sentiment lexicon—1. Results of the second step: optimisation with regard to τ are presented in **Figure 3** for the first corpus and in **Figure 4** for the second corpus.

Each plot shows the value of F1-score of a model trained on a single feature with regard to the parameter τ . The results may seem different at first, but they are in fact quite similar. On both corpuses there is a clear maximum point around 0.8 for all three features. The difference is that on the second corpus there is a distinct quality prevalence of SPO and Fillmore roles features over the sentiment lexicon model. One way to explain this is that the second dataset is significantly larger than the first one so there is better convergence. In any case, the goal of our research is not to find out what feature works best to classify opinionated texts but whether their combination provides a noticeable increase of quality.

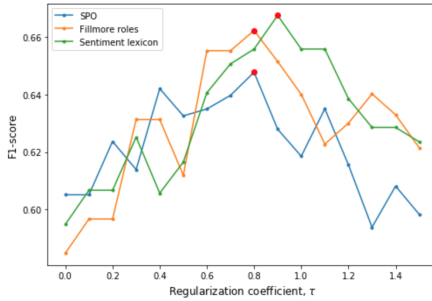


Figure 3: Optimal τ for Corpus 1

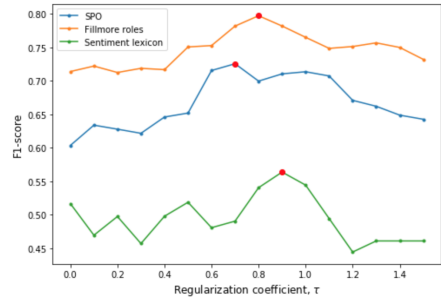


Figure 4: Optimal τ for Corpus 2

5.2. Lexical baseline

To demonstrate the sensibility of using probabilistic models with complex features we will compare our model with a simple bag-of-words baseline. First, we represent each document with a vector of its words.

Thus the corpus can be viewed as a matrix, where each column is a feature vector $\{tf\text{-idf}(w, d)\}_{W \times D}$. Next, we clusterize the documents using the k-means algorithm.

The result of the k-means depends on the initial point, so to obtain a more representative result we conducted 100 experiments with the baseline and averaged the results for each corpora. **Figure 5** shows the average f1 measure for the baseline algorithm. The number of iterations show how many experiments were conducted. The average quality is 0.67 on Corpus 1 and 0.72 on Corpus 2. The result variance is around 0.5, it is higher on Corpus 1 due to its smaller size.

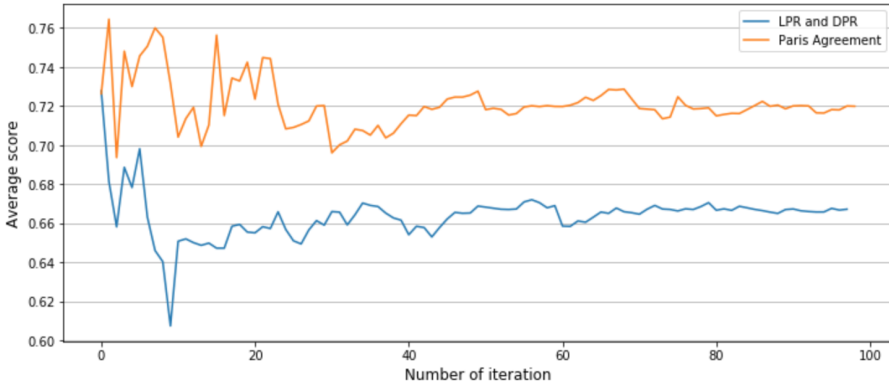


Figure 5: Average F1-measure for lexical baseline

5.3. Evaluation results

To evaluate clustering we will be measuring precision, recall and F1-score. Both datasets have two classes, so the quality for a constant model would be 0.5. The goal is to determine what defines an opinion. In order to do that we will be training our probabilistic model with different sets of modalities. They will be comprised of:

- Subject-predicte-object triplets (SPO)
- Fillmore roles (FR)
- Sentiment lexicon (Sent)

In the beginning we train probabilistic models with a single feature to see how well they can clusterise opinionated news on their own. Then we try combining features in pairs: SPO triples together with Fillmore roles (SPO+FR), SPO triples with sentiment lexicon (SPO+Sent) and Fillmore roles with sentiment lexicon (FR+Sent). Finally, we evaluate the model using a combination of all three features. The results are presented in **Tables 2** and **3**.

Table 2: LPR and DPR enterprises

Modalities	Pr	Rec	F1
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
SPO+FR+Sent	0.77	0.97	0.86

Table 3: Trump leaving the Paris Agreement

Modalities	Pr	Rec	F1
TF-IDF	0.57	0.97	0.72
SPO	0.56	0.99	0.72
FR	0.67	0.97	0.79
Sent	0.56	0.55	0.55
SPO+FR	0.72	0.99	0.83
SPO+Sent	0.57	0.99	0.72
FR+Sent	0.73	0.97	0.83
SPO+FR+Sent	0.77	0.94	0.85

The resulting models that used a combination of all three features significantly outperformed the lexical baseline. However, single-feature models showed similarly low quality.

On the first corpus all three models trained with a single feature showed quality around 0.65. This shows that in the first dataset a single syntactic or semantic feature is unable to separate news relating to different opinions, which is not surprising. However, the features gave false classifications on separate groups of news. This fact allowed models trained on a combination of two features to show a significant growth of quality with a combination of SPO triples and sentiment lexicon showing the F1-score of 0.81. Combining all three features yields the best result extending the quality growth and bringing the F1-score up to 0.86 mainly on account of notably increasing recall.

Analysing performance on the second corpus, we can see a similar pattern. Of the three single feature models only the one using Fillmore roles showed decent quality. The SPO model had a high recall score but very low precision, while sentiment lexicon had poor performance overall. Once again, combining the features increased the F1-score in all three cases. We note very interpretable results in the sense that combining the two best features yielded the best results of all pairs raising the quality to 0.83. Similar to Corpus 1 the best score of 0.85 was reached by combining all three features bringing precision up significantly from single-feature models.

The resulting scores were reached with one point of the modalities weights τ_m distribution. A logical question arises: how stable are the optimal points of this hyperparameter? To answer it we paired the features and conducted experiments with varying weights distribution. For example, for the feature pair SPO triples and Fillmore roles we set the weight of the latter with values $\tau_r \in \{0, 0.05, 0.1, \dots, 1\}$ and the weight of the triples as $1 - \tau_r$. A similar procedure was done with other pairs of features. Results are illustrated in **Figure 6** for Corpus 1 and **Figure 7** for Corpus 2. Firstly, let us examine **Figure 7**. For the pair roles-SPO (third line) the left end shows a model trained solely with semantic roles, which has the F1-score of 0.79, and the right end reflects a model using only SPO triples with the F1-score of 0.72. As we move right, adding more weight to the SPO modalities, quality starts increasing to the point of weight distribution of $\{\tau_r = 0.7, \tau_s + \tau_o = 0.3\}$ and gradually decreases from there. A similar result is observed with other feature pairs, allowing us to conclude that the optimum is stable with regard to feature weights. Moving on to **Figure 6** we see, that on the first corpus the optimal solution is not as stable. All in all, a similar pattern appears: moving from right to left increases quality to a point, just not as gradually. Tweaking the weights by 0.1 can change the score notable. There are several ways to explain this behaviour. Firstly, the dataset itself is quite small so results are less stable. Secondly, all the single-feature models have relatively low performance on their own. Further investigation of this issue requires widening Corpus 1 and gathering additional datasets which we leave for future work.

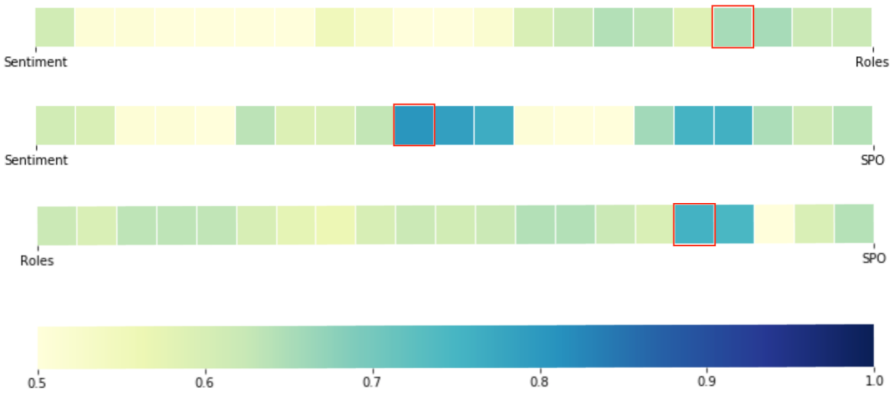


Figure 6: F1-score distribution over modalities weights for Corpus 1

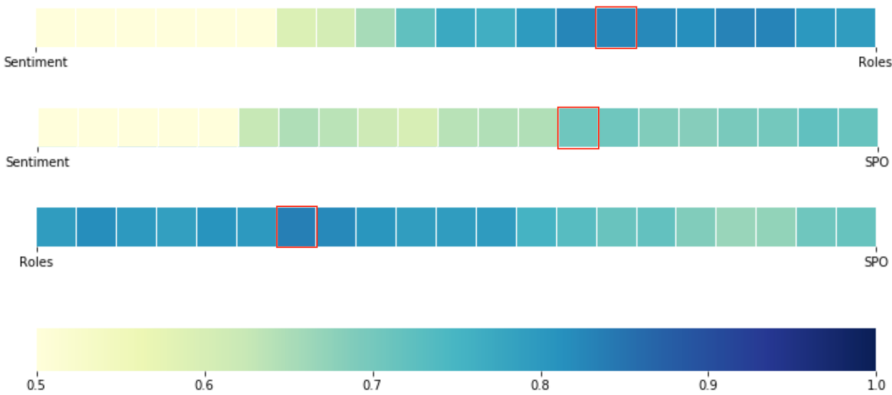


Figure 7: F1-score distribution over modalities weights for Corpus 2

6. Conclusion and Future work

We presented a model for opinion mining in Russian texts that was able to perform two-class clusterization with the F1-score above 0.85 on two datasets. We presented algorithms for mining three syntactic and semantic features: SPO triples, Fillmore roles and sentiment lexicon and compared them with a baseline lexical model. The proposed probabilistic model was trained on all three features as well as their combinations. We demonstrated that combining syntactic and semantic features allows to classify opinionated texts accurately even when single-feature models show low quality and provides a significant advantage over the baseline. We collected and labelled two collections of news on political events to conduct the experiments and made them available for public use.

In this work, we did not determine what weights the features should be combined with to provide optimal clustering quality. To put it in simple terms, we proved that an opinion is defined by mentioned facts, sentiment lexicon and semantic roles of the words used but did not determine “how much” each feature contributed to expressing an opinion. Thus, applying the algorithm at its current state on a raw corpus requires some labelled data to optimise the weights. Adapting self-learning techniques requires widening existing datasets and collecting new ones which we plan to do in the future.

7. Acknowledgements

This work is supported by Russian Foundation for Basic Research, grant 20-07-00936.

References

1. *Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai*. “Topic sentiment mixture: Modeling facets and opinions in weblogs”. In Proceedings of the World Wide Conference (2007), pp. 171–180.
2. *B. Pang, L. Lee*. “Opinion Mining and Sentiment Analysis. Foundations and Trends”. In Information Retrieval (2008), pp. 1–135.
3. *M. S. Hajmohammadi, R. Ibrahim, Z. A. Othman* “Opinion Mining and Sentiment Analysis: A Survey”. In International Journal of Computers & Technology Vol. 2, No. 3 (2012).
4. *M. J. Paul, R. Girju* “Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models”. In Proc. of EMNLP ’09 (2009), pp. 1408–1417.
5. *Y. Fang, L. Si, N. Somasundaram, Z. Yu* “Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model”. In: Proc. of WSDM ’12 (2012), pp. 63–72.
6. *R. Balasubramanyan, W. W. Cohen, D. Pierce, D. P. Redlawsk* “Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News?”. In: Proc. of the Sixth International AAAI Conference on Weblogs and Social Media (2012), pp. 18–25.
7. *M. J. Paul, R. Girju* “A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics”. In: Proc. of AAAI ’10 (2010), pp. 545–550.
8. *X. Zhu, D. Klabjan, P. N. Bless* “Unsupervised Terminological Ontology Learning based on Hierarchical Topic Modeling”. In: Proc. of ACL 17 (2017).
9. *E. I. Bolshakova, K. V. Vorontsov and others* “Automatic word processing in natural language and data analysis”. pp. 195–228.
10. *Koltsova, O. Yu, S. Alexeeva and S. Kolcov* “An opinion word lexicon and a training dataset for russian sentiment analysis of social media.”. In: Proc. of the International Conference “Dialogue 2016” (2016).
11. *Lashevich G. et al.* “Creating Russian WordNet by Conversion.”. In: Proc. of the International Conference “Dialogue 2016” (2016).

12. *P. Sobkowicz, M. Kaschesky, G. Bouchard* “Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web”. In: *Government Information Quarterly* vol. 29 (2012) pp. 470–479.
13. *Charles J. Fillmore* “Some problems for case grammar”. In: *22nd Annual Round Table. Linguistics: Developments of the Sixties Viewpoints of the Seventies. Volume 24 of Monograph Series on Language and Linguistics.* (1971) pp. 35–56.
14. *H. Wang and C. Zhai* “Generative Models for Sentiment Analysis and Opinion Mining”. In: *Springer International Publishing AG* (2017).
15. *S. M. Kim, E. Hovy* “Extracting opinions, opinion holders, and topics expressed in online news media text”. In: *Proc. of the Workshop on Sentiment and Subjectivity in Text* (2006) pp. 1–8.
16. *K. Moilanen, S. Pulman* “Sentiment Composition”. In: *Proc. of RANLP-2007* (2007) pp. 378–382.
17. *I. A. Karpov, M. V. Kozhevnikov, V. I. Kazorin, N. R. Nemov* “Entity Based sentiment analysis using syntax patterns and convolutional neural network”. In: *Proc. of the International Conference “Dialogue 2016”* (2016) pp. 378–382.
18. *A. O. Shelmanov, D. A. Devyatkin* “Semantic role labeling with neural networks for texts in Russian”. In: *Proceedings of the International Conference “Dialogue 2017”* (2017).
19. *D. Marcheggiani, A. Frolov, and I. Titov* “A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling”. In: *arXiv preprint arXiv:1701.0259* (2017).
20. *C. A. Thompson, R. Levy, and C. D. Manning* “A Generative Model for Semantic Role Labeling”. In: *Springer-Verlag Berlin Heidelberg* 2003 (2003), pp. 397–408.
21. *A. Giuglea and A. Moschitti* “Semantic Role Labeling via FrameNet, VerbNet and PropBank”. In: *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* (2006), pp. 929–936.

EVENT2MIND FOR RUSSIAN: UNDERSTANDING EMOTIONS AND INTENTS IN TEXTS. CORPUS AND MODEL FOR EVALUATION

Fenogenova A. S. (alenuh@gmail.com)

Sberbank, Moscow, Russia

Tikhonova M. I. (m_tikhonova94@mail.ru)

Sberbank; National Research University Higher School
of Economics, Moscow, Russia

Filipetskaya D. V. (dafi913@yandex.ru)

Moscow Institute of Physics and Technology, Moscow, Russia

Mironenko F. D. (fomius2000@yandex.ru)

Saint Petersburg State University, Saint-Petersburg, Russia

Tabisheva A. O. (anastasiatabisheva@yandex.ru)

National Research University Higher School of Economics,
Moscow, Russia

The paper provides a comprehensive overview of the corpus for the Russian language for the commonsense inference task. Namely, we construct event phrases, which cover a wide range of everyday situations with labelled intents and reactions of the event main participant and emotions of other people involved. The dataset consists of two parts: a crowdsourced corpus of 6,756 examples from Russian sources and a translated into Russian part of the original corpus of 23,409 examples. Apart from this, we use the collected data in order to train the event2mind model for the Russian language. The paper presents careful description of the best Russian model and the results of the conducted experiments.

Key words: event2mind, chatbots, emotion detection, intents generation, Natural Language Processing, dialog systems

DOI: 10.28995/2075-7182-2020-19-299-309

EVENT2MIND ДЛЯ РУССКОГО ЯЗЫКА: КОРПУС И МОДЕЛЬ — ПОНИМАНИЕ ЭМОЦИЙ И ИНТЕНДОВ В КОРОТКИХ ТЕКСТАХ

Феногенова А. С. (alenush@gmail.com)

Сбербанк, Москва, Россия

Тихонова М. И. (m_tikhonova94@mail.ru)

Сбербанк; Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Филипецкая Д. В. (dafi913@yandex.ru)

Московский физико-технический институт, Москва, Россия

Мироненко Ф. Д. (fomius2000@yandex.ru)

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

Табишева А. О. (anastasiatabisheva@yandex.ru)

Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Распознавание интента (намерения) субъекта является естественным для живого человека и весьма сложной задачей для компьютера. В данной работе представлен корпус для русского языка для задачи распознавания намерения субъекта: по полученному на вход короткому тексту-событию определяются причины, по которым субъект совершил действие, а также эмоции субъекта и других участников события. Формат корпуса соответствует формату оригинального корпуса на английском языке. Собранный корпус состоит из двух частей: размеченный русский корпус из 6756 примеров и переведенная автоматическим переводчиком с английского отфильтрованная часть английского корпуса из 23409 примеров. Помимо этого был проведен ряд экспериментов по обучению модели для русского языка и получена модель, сравнимая по качеству с английской. Это доказывает воспроизводимость алгоритма для языков с более сложным по сравнению с английским морфологическим составом.

Ключевые слова: event2mind, чатботы, диалоговые системы, детектирование эмоций, Natural Language Processing

1. Introduction

Common sense reasoning tasks have received significant attention in Natural Language Processing and attempts [2], [8], [12], [14], [17], [18] to solve them have been made in recent years. Such type of automatic pragmatic reasoning can be useful for a wide range of NLP applications that require anticipation of people's reasoning and emotions. For instance, incorporating such model into a dialog system could make its reactions more emphatic and appealing for users. Moreover, commonsense reasoning is often regarded as a necessary step towards human understanding.

Most of the proposed models [2], [12], [17], [18] are supervised and, therefore, require large datasets, often with quite nontrivial markup, for training. Constructing such datasets is difficult and expensive. For English there exist several corpora¹ [12], [15], [16] for commonsense reasoning tasks. For Russian, however, the situation is not so good. The absence of data is one of the main obstacles to adaptation most of the models for the Russian language.

Common sense reasoning problem could be stated in different ways. The one considered in the paper, originally proposed in [12], is formulated as follows: given a short free-form text describing an event ("*PersonX eats breakfast in the morning*") a model makes reasoning about the agent's intents ("*X wants to satisfy hunger*"), reactions ("*X feels satiated, full*") and possible reactions of the other event's participants.

In [12] the authors presented their model, *event2mind*, which supports commonsense inference on events phrases. In their work a specific focus was made on modeling stereotypical intents and reactions of people. Another contribution of their work was a crowdsourced corpus that supports commonsense inference in English language, which is publicly available².

The main difficulty in adapting *event2mind* model for the Russian and other languages is that it requires a large corpus of the event phrases with labelled intents and reactions for training. Thus, in order to train a model for any other language besides English it is necessary to construct a dataset.

The main contribution of the paper is a text corpus suitable for *event2mind* training in Russian which consists of two parts:

1. 6,756 event phrases covering a diverse range of everyday events and situations in Russian,
2. a subset of 23,409 event phrases from English corpus translated via Google translator³.

In addition to that, we used the obtained corpus in order to train Russian model of *event2mind*. The article provides a careful description of the conducted experiments with different model's versions and configs and presents the best one.

¹ http://nlpprogress.com/english/common_sense.html

² <https://uwnlp.github.io/event2mind/data/>

³ <https://translate.google.ru>

2. Dataset

One of the goals of the work was to collect a labelled corpus of short free-form texts, which are further referred to as *event phrases*, in Russian in the format suitable for event2mind training. We define events like authors in the original work and try to collect phrases that contain a diverse range of common everyday events and situations.

2.1. Crowdsourced corpus

First step was to collect a sufficient amount of events. For this purpose we gathered texts from several sources:

1. short episode descriptions of TV series and soap operas. We manually selected 50 TV series from the Internet portal KisTV⁴ making the focus on series about present everyday life situations. In total descriptions of 50 TV series were used among which are Friends, Sex and the City, Santa Barbary, Univer (Универ), Kitchen (Кухня) and others. We did not use fantasy, science fiction or series about medicine as long as they contain a lot of specific words which are not commonly used.
2. Book summaries from Briefly.⁵ Total number of downloaded summaries equals 1,512.
3. Texts from SynTagRus corpus⁶ [19], [3]—a subcorpus of Russian National Corpus, with fiction and news, with manual syntactic annotation.

Маша пьет кофе в модном кафе

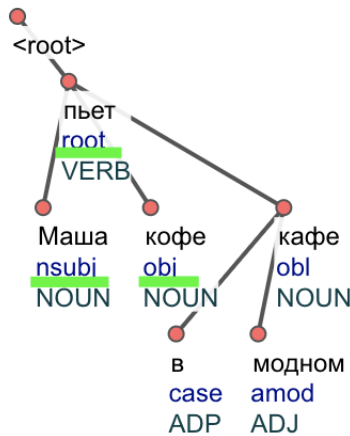


Figure 1: Events in the tree are green-labelled

⁴ <http://kistv.ru/>

⁵ <http://briefly.ru>

⁶ https://github.com/UniversalDependencies/UD_Russian-SynTagRus

From these three sources event phrases are extracted. We select event phrases as a combination of a verb predicate with partially instantiated arguments, like in the **Figure 1**. Events from row texts are derived using UDPipe [13] syntactic parser, model trained on Russian Syntagrus, version UD2.4. Namely, we automatically search for syntactic patterns of verb with its arguments in a syntactic tree that satisfy one of the following rules:

- *nsubj + root + obj*
- *nsubj + root + iobj*
- *nsubj + advmod + root*
- *nsubj + root + case + obl*
- etc.

Similar to the original paper we introduce type variables for generalisation of event phrases. In particular, predicate subjects corresponding to pronouns and name entity mentions are replaced with typed variables such as PersonX, PersonY or PersonZ (if there is more than one person in the event). Named Entity Recognition is conducted to replace named entities with the aliases according to the grammatical information from UDPipe. Namely, we have considered all the tokens that match the syntactic pattern above marked with tags PROPEN and PRONOUN. Then we work only with text fragments that matched the rule patterns. In addition, phrases which do not contain inanimate subjects are also filtered using grammar information from UDPipe. Following the original work our corpus contains only events with person named entities.

After the depersonalization frequency analysis and Levenshtein distance are used in order to select most common event phrases and to filter nonstandard examples which occur only once. First, we take all the event phrases which occur more than once. For the rest of the data for every pair of event phrases pairwise Levenshtein distance $L(\text{phrase}_1, \text{phrase}_2)$ is computed and for every pair with $L(\text{phrase}_1, \text{phrase}_2) \leq 5$ the shorter phrase is chosen for the final dataset. Thus, we obtain 4523 unique event phrases.

2.2. Crowdsourcing

In order to annotate raw events with intents and reactions we design a Yandex.Toloka⁷ task supplemented with annotation instruction. A snippet of the task is shown in **Figure 2**. For each event phrase we ask:

- *weather it contains a meaningful event,*
- *is it possible to find some reasons that cause such event,*
- *can subjects of the event have some emotions and reactions to it.*

In addition to that we provide possible answers generated by the trial event-2mind model trained only on automatically translated part of the dataset. This model performs poorly, nevertheless in some cases it generates reasonable answers.

In case of positive answer for the first question in the toloka's task we ask annotators to verify trial model's answers and then to give their own variants for possible intents and reactions of the agent (PersonX) and reactions of other event participants if any.

⁷ <https://toloka.yandex.ru/>

We encourage annotators to write more than one variant for every event phrase and, in addition, we get verified or corrected variant from the trial model. Thus, we get at least two valid answers for an example. We set the toloka’s task with annotator overlap equal 3. Then we drop the events that were considered to be wrong by at least 2 annotators and set label “none” if in the example there were only intents and no reactions or vice versa.

It should be mentioned that for label standardization we recommend annotators to use infinitive word forms such as *хотеть есть* (be hungry), *быть счастливым* (be happy), *грустный* (sad).

ЧеловекX выражает свою признательность ЧеловекуY

Присутствует ли в этом тексте осмысленное событие, у которого можно выделить предпосылки или реакции действующих лиц?

Да Нет

Можно ли по этому тексту понять, какие намерения были у действующего лица (если оно есть)?

Да Нет

Может ли **благодарен** быть причиной этого события?

Да Нет

Напишите наиболее вероятную причину этого события

ЧеловекY что-то сделал хорошее

Можно ли по этой ситуации понять реакцию главного действующего лица (ЧеловекX)?

Да Нет

Может ли **мощный** быть реакцией ЧеловекаX на это событие?

Да Нет

Напишите наиболее вероятную реакцию ЧеловекаX на это событие

благодарный

Может ли это событие вызвать реакцию окружающих?

Да Нет

Может ли **лучше** быть реакцией окружающих на это событие?

Да Нет

Напишите наиболее вероятную реакцию окружающих на это событие

счастливый

Figure 2: Toloka’s markup for event2mind task

As a result we collect almost 7k examples of good Russian events and corresponding intents and reactions for them.

2.3. Translated English corpus

Besides the annotated Russian dataset we prepared an automatically translated into Russian and then cleaned English corpus. The motivation for this was that we did not possess enough resources to annotate as many events in Russian as the authors did. In order to find the dependence between the model’s performance and the corpus size

and to estimate the minimum size of Russian dataset sufficient for the model's training we performed a number of experiments. The original English model was trained on different subsets of the original corpus, which were of different size. The results of the experiments are presented in **Table 1**. It should be mentioned that we were unable to obtain exactly the same scores as in paper though we used the original Allennlp⁸ event2mind config of the best model loaded in official framework's repository.

Table 1: English model performance in dependence on the dataset size. Evaluation of intents/reactions is measured similarly to the original paper by *recall@10* (percentage of times the gold falls within the top 10 decoded; higher is better) on development set

data size	val loss	intent	xreact	oreact
46k (full)	2.52	0.38	0.41	0.65
30k	2.60	0.36	0.39	0.65
20k	2.74	0.32	0.37	0.63
5k	3.22	0.31	0.35	0.55

It could be observed that small dataset of 5,000 examples is obviously not enough for the model to obtain reasonable quality. However, with the set of 30 thousand events the validation loss and intent recall metrics do not critically decrease compared to the full English corpus. Thus, it could be concluded that the annotated Russian corpus of about 7,000 events is not enough for training a model and the data should be to augmented.

For this purpose we create a supplementary dataset from English corpus automatically translated into Russian. The original dataset is translated via Google translator. The English corpus contains data from several sources: ROC Story training set [11], the GoogleSyntactic N-grams [4], the Spinn3r corpus [5], and idioms. However, we take only examples from the source ROC Story as it has the highest annotation agreement statistics according to the original research [12] and correspondingly it contains more clear events which are less exposed to the automatic translation errors.

The translated data is further filtered and phrases containing English or transcribed phrases are removed. Finally, the data is checked by reviewers and sentences which were translated wrongly are deleted.

It should be noted that the translated corpus is of poorer quality compared to the crowdsourced one due to the imperfections of the automatic translation and the lack of coherence between the phrases and the labels.

2.4. Final dataset

The final Russian corpus⁹ is a union of the Russian crowdsourced dataset and the automatically translated part of English corpus which contains 30,165 events in the event2mind format.

⁸ <https://allennlp.org/>

⁹ https://github.com/Alenush/russian_event2mind/tree/master/dataset

3. Model and experiments

The event2mind model aims to generate three entity-specific pragmatic inferences (subject’s intent, subjects’s reaction, and others events participants’ reactions) given an event phrase in free-form text. First, the input is encoded as a vector $h^E \in \mathbb{R}^H$. This vector is further used to predict the output which consists of three sequences of words. Event2mind is a neural encoder-decoder model. The system is multitask learning, simultaneously minimizing the loss for all three decoders at each iteration.

RNN decoder generates the textual description. The event phrase embedding h^E is set as the initial state h_{dec} of three decoder RNNs, which then output the intent/reactions one word at a time (using beam-search at test time). An event’s intent sequence ($v_i = v_i^{(0)}, v_i^{(1)}, \dots$) is computed by the following formula:

$$v_{(t+1)i} = \text{softmax}(W_i \text{RNN}(v_{(t)i}, h_{(t)i,dec}) + bi).$$

Thus, a model can successfully compose embedding representations of previously unseen events and reactions. Though the event sequences are typically rather short (3.6 tokens on average), event2mind model still benefits from the BiRNN’s ability to compose words.

In the conducted experiments we used the full final Russian dataset (both translated and crowdsourced parts). In order to make translated and annotated example distribution more uniform the dataset was shuffled before training the model. It should be mentioned that we also experimented with training a model only on the translated part of corpus, however, it did not perform well. In the original article the authors performed experiments with different encoders and decoders but the most promising one were found to be BiRNN with GRU and vector of size 100 in encoder and sequence decoder. We tried LSTM and GRU in encoder and different embeddings models: such as fasttext [1], [7] and word2vec [9], [10] from RusVectores source.¹⁰ The results are presented in the **Table 2**.

Table 2: Scores on Russian dataset with different model configurations. Evaluation of intents/reactions is measured similarly to the original paper by *recall@10* (percentage of times the gold falls within the top 10 decoded; higher is better) on development set.

Best models by the average *recall@10* are highlighted in bold.

vectors	architecture	val loss	intent	xreact	oreact
araneum fasttext	LSTM	0.9704	0.818	0.725	0.92
araneum fasttext	GRU	0.9695	0.819	0.725	0.92
ruscorpora fasttext	LSTM	0.9508	0.821	0.725	0.9195
ruscorpora fasttext	GRU	0.9517	0.822	0.7255	0.9195
araneum word2vec (skipgram)	LSTM	0.916	0.816	0.707	0.924
araneum word2vec (skipgram)	GRU	0.919	0.827	0.727	0.92
ruscorpora word2vec (skipgram)	LSTM	0.9205	0.816	0.70	0.92
ruscorpora word2vec (skipgram)	GRU	0.917	0.825	0.725	0.923

¹⁰ <https://rusvectors.org/ru/models/>

From the obtained results the following conclusions could be made:

- word2vec embeddings perform a little better than fasttext ones,
- ruscorpora fasttext embeddings perform better compared with the araneum ones,
- with word2vec embeddings GRU shows better results than bidirectional LSTM.

The best model¹¹(areneum word2vec + GRU) repeats the English results on the Russian data. **Table 3** shows several examples of the model's work on real event phrases. It could be seen that the model's predictions for intents are sensible and reflect common knowledge and reasoning. Still person's reactions are of poorer quality. It may be explained by the fact that the set of emotions depends on a larger piece of text than one short phrase. Therefore, event phrase without the context is not enough for making precise decisions about peoples' feelings or emotions.

Table 3: Examples of model's commonsense inference

Event	Intent	PersonX's reaction
PersonX выпил кофе	пить, разбудить, проснуться	счастливый, гордый, довольный
(PersonX drank some coffee)	(to drink, to awaken, to wake up)	(happy, proud, satisfied)
PersonX позвал на свидание PersonY	любовь, привлечь, благодарен	довольный, счастливый, облегчение
(PersonX invited PersonY for a date)	(love, attract, grateful)	(satisfied, happy, relief)
PersonX идет в школу	учить, веселиться, чтобы получить образование	гордый, грустный, довольный
(PersonX goes to school)	(to teach, to have fun, to get education)	(proud, sad, satisfied)

4. Conclusion

In the paper a new Russian corpus for event2mind task was introduced. Our corpus supports learning representations over a diverse range of events and reasoning about the likely intents and reactions of previously unseen events. In addition to that the results of the experiments with the Russian model are described and the code for the best model is provided. It was demonstrated that architecture works for more grammatically complicated language than English and it is still performing commonsense inference on textually described everyday events. The dataset and Russian model are provided in the repository¹².

¹¹ https://github.com/Alenush/russian_event2mind/tree/master/model

¹² https://github.com/Alenush/russian_event2mind

In the future we plan to increase the model's performance by experimenting with different language models such as BERT and ELMo, for example. In addition to that we have noticed that the number of subject's reactions is quite limited and, therefore, it could be regarded as a classification problem. Thus, we plan to train a separate classification model on top of BERT or ELMo embeddings.

References

1. *Bojanowski, Piotr, et al.* "Enriching word vectors with subword information". *Transactions of the Association for Computational Linguistics* 5 (2017): 135–146.
2. *Devlin, Jacob, et al.* "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805 (2018).
3. *Droganova, Kira, Olga Lyashevskaya, and Daniel Zeman.* "Data conversion and consistency of monolingual corpora: Russian UD treebanks". *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, December 13–14, 2018, Oslo University, Norway. No. 155. Linköping University Electronic Press, 2018.
4. *Goldberg, Yoav, and Jon Orwant.* "A dataset of syntactic-ngrams over time from a very large corpus of english books". (2013).
5. *Gordon, Andrew S., and Reid Swanson.* "StoryUpgrade: Finding Stories in Internet Weblogs". *ICWSM*. 2008.
6. *JBenko, V., & Zakharov, V. P.* "Very large Russian corpora: new opportunities and new challenges. In *Computational linguistics and intellectual technologies* (pp. 79–93) (2016).
7. *Joulin, Armand, et al.* "Bag of tricks for efficient text classification". arXiv preprint arXiv:1607.01759 (2016).
8. *Liu, Quan, et al.* "Combing context and commonsense knowledge through neural networks for solving winograd schema problems". *2017 AAAI Spring Symposium Series*. 2017.
9. *Mikolov, Tomas, et al.* "Efficient estimation of word representations in vector space". arXiv preprint arXiv:1301.3781 (2013).
10. *Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig.* "Linguistic regularities in continuous space word representations". *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2013.
11. *Mostafazadeh, Nasrin, et al.* "A corpus and evaluation framework for deeper understanding of commonsense stories". arXiv preprint arXiv:1604.01696 (2016).
12. *Rashkin, Hannah, et al.* "Event2mind: Commonsense inference on events, intents, and reactions". arXiv preprint arXiv:1805.06939 (2018).
13. *Straka, Milan, Jan Hajic, and Jana Straková.* "UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.
14. *Trinh, Trieu H., and Quoc V. Le.* "A simple method for commonsense reasoning". arXiv preprint arXiv:1806.02847 (2018).

15. Wang, Alex, et al. “Glue: A multi-task benchmark and analysis platform for natural language understanding”. arXiv preprint arXiv:1804.07461 (2018).
16. Wang, Alex, et al. “Superglue: A stickier benchmark for general-purpose language understanding systems”. Advances in Neural Information Processing Systems. 2019.
17. Yang, Zhilin, et al. “Xlnet: Generalized autoregressive pretraining for language understanding”. Advances in neural information processing systems. 2019.
18. Zellers, Rowan, et al. “Swag: A large-scale adversarial dataset for grounded commonsense inference”. arXiv preprint arXiv:1808.05326 (2018).
19. Дяченко, П. В., et al. “Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус)”. Труды Института русского языка им. ВВ Виноградова 6 (2015): 272–300.

ИМПЛИЦИТНЫЕ ЛОГИКО- СЕМАНТИЧЕСКИЕ ОТНОШЕНИЯ И МЕТОД ИХ ПОИСКА В ПАРАЛЛЕЛЬНЫХ ТЕКСТАХ

Гончаров А. А. (a.gonch48@gmail.com)

ИПИ ФИЦ ИУ РАН, Москва, Россия

Инькова О. Ю. (Olga.Inkova@unige.ch)

ИПИ ФИЦ ИУ РАН, Москва, Россия;

Женевский университет, Женева, Швейцария

DOI: 10.28995/2075-7182-2020-19-310-320

IMPLICIT LOGICAL-SEMANTIC RELATIONS AND A METHOD OF THEIR IDENTIFICATION IN PARALLEL TEXTS

Goncharov A. A. (a.gonch48@gmail.com)

Institute of Informatics Problems, FRC CSC RAS,
Moscow, Russia

Inkova O. Yu. (Olga.Inkova@unige.ch)

Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia;
University of Geneva, Geneva, Switzerland

One of the main characteristics of logical-semantic relations (LSRs) between two fragments of a text is that these relations can be either explicit (expressed by some marker, e. g. a connective) or implicit (derived from the interrelation of these fragments' semantics). Since implicit LSRs do not have any marker, it is difficult to find them in a text (whether automatically or not). In this paper, approaches to analysing implicit LSRs are compared, an original definition for them is offered and differences between implicit LSRs and LSRs expressed by non-prototypical means are described. A method is proposed to identify implicit LSRs using a parallel corpus and a supracorpora database of connectives. Based on the well-known statement that LSRs can be explicitated by adding connectives in the translation, it is argued here that through selecting pairs in which fragments where a connective is used to express an LSR in the translation correspond to those containing any of the translation stimuli standard for this connective in the source language, it is possible to get an array of contexts in which this LSR is implicit in the

source text (or expressed by means other than connectives). This method is then applied to study the French causal connectives *car*, *parce que* and *puisque* using a Russian-French parallel corpus. The corpus data are analysed to obtain information about LSRs particularly about cases where the causal LSR in Russian is implicit, as well as about the use of causal connectives in French. These results are used to show that the method proposed allows to quickly create a representative array of contexts with implicit LSRs, which can be useful in both text analysis and in machine learning.

Key words: corpus linguistics, parallel corpora, logical-semantic relations, implicitness, discourse connectives

1. Введение

Под логико-семантическими отношениями (ЛСО) мы понимаем такие отношения между двумя фрагментами текста, которые потенциально могут быть выражены специализированным показателем, коннектором (ср. отношения причины, следствия, сопоставления и др., выражаемые союзами). Уже тот факт, что в этом определении фигурирует слово «потенциально», свидетельствует о том, что ЛСО могут быть как эксплицитными, т. е. выраженными соответствующим показателем, так и имплицитными, т. е. выводимыми на основе смыслового соотношения следующих друг за другом фрагментов текста, коммуникативной ситуации или знаний об устройстве мира. Если коннекторы и, соответственно, эксплицитные ЛСО довольно хорошо описаны, то в изучении имплицитных ЛСО пока остаётся много белых пятен. Специального рассмотрения заслуживает вопрос об их идентификации в тексте, что становится особенно актуальным при разработке систем автоматического анализа и генерации текста. Как известно, уровень консенсуса при разметке имплицитных отношений гораздо ниже, чем при разметке эксплицитных, а сами разметчики часто опираются в своих решениях на возможность подстановки того или иного коннектора между двумя высказываниями. Кроме того, данных о частотности случаев, когда то или иное ЛСО является имплицитным, практически не существует, что обусловлено сложностью получения представительного массива аннотаций с имплицитными отношениями.

Мы начнём с обзора существующих подходов к исследованию имплицитных ЛСО и их определения, после чего опишем разработанный нами метод их поиска с использованием параллельных текстов и надкорпусной базы данных (НБД) коннекторов, а также результаты его применения.

2. Имплицитные ЛСО: подходы и определения

Имплицитные отношения между двумя следующими друг за другом фрагментами текста признаются всеми без исключения существующими теориями анализа текста и дискурса, а также грамматиками. Однако содержание понятия имплицитного отношения в значительной степени различается от подхода к подходу. В грамматиках имплицитные ЛСО (т. е. в данном случае те, которые

потенциально могут быть выражены сочинительными и подчинительными союзами или относительными словами) рассматриваются в разделе, посвящённом бессоюзной синтаксической связи (ср. для русского языка [Виноградов, Истрина 1960а]; [Виноградов, Истрина 1960б]; [Шведова 1970]; [Шведова 1980], аналогичная ситуация наблюдается и в грамматиках основных европейских языков). В то же время отмечается, что «круг отношений, выражаемых бессоюзными сложными предложениями, не совпадает с соответствующими функциями сложносочинённых и сложноподчинённых предложений» [Виноградов, Истрина 1960а: 102]. Основным критерием для выделения данного типа сложных предложений служит, тем не менее, отсутствие показателя синтаксической связи между его частями, о чём говорит его название.

В наиболее известных теориях анализа структуры текста — Теории риторической структуры (RST) У. Манна и С. Томпсон, Теории сегментной репрезентации дискурса (SDRT) Н. Ашера и теоретических положениях, лежащих в основе разработки Пенсильванского аннотированного корпуса (PDTB), — признаётся возможность имплицитных отношений между дискурсивными единицами (ДЕ). Но сами термины «отношение» и, следовательно, «имплицитное отношение» понимаются по-разному.

RST и SDRT можно определить как теории, основанные на «глобальном» подходе к связности текста. Это означает, что текст представляет собой единый дискурсивный объект, все элементы которого связаны между собой «риторическими» (в терминах RST) отношениями. Следовательно, понятие «риторическое» отношение является родовым по отношению к понятию ЛСО (подробнее см. [Инькова 2019а]), а коннектор рассматривается лишь как один из видов показателей риторических отношений, причём как показатель малозначимый, поскольку его функция заключается только в том, чтобы сузить спектр возможных интерпретаций отношений между ДЕ [Mann, Thompson 1986: 71]. При этом допускается, что риторическое отношение может не иметь никакого показателя [Там же: 68]. Однако, поскольку в литературе, посвящённой маркированию отношений, коннекторы долгое время считались единственными показателями отношений связности, немаркированными (имплицитными) считались только такие отношения, которые не выражены коннектором; см. [Martin 1992]; [Renkema 2004]; [Taboada 2009]. Тем не менее, как показано в [Taboada 2006], коннекторы маркируют только 30–40% (в зависимости от типа текста, письменного vs. устного) риторических отношений. Поэтому вопрос о средствах маркирования отношений был пересмотрен, в частности в связи с применением RST для создания аннотированных корпусов текстов (см. [Carlson et al. 2003]) и автоматической идентификации риторических отношений. К коннекторам добавляются показатели самой разнообразной природы: лексические, морфологические (временные формы), семантические (синонимия, антонимия и др.), синтаксические (различные виды придаточных и др.), графические (знаки препинания и др.) и т. д.; они могут также комбинироваться между собой [Das, Taboada 2013: 5–6]; [Taboada, Das 2013: 257]; [Das, Taboada 2014: 8–9]. Сам термин «имплицитное» отношение исчезает из пособий по аннотированию; см., например, [Carlson, Marcu 2001]; [Das, Taboada 2014]. В случае, когда никакой

из потенциальных показателей отношения не может быть идентифицирован, проставляется метка *unsure*; ср. (1), приводимый в [Das, Taboada 2014: 33]:

- (1) {*This hasn't been Kellogg Co.'s year.*} {*The oat-bran craze has cost the world's largest cereal maker market share.*} Cause [wsj_610: 1/2]¹

В SDRT, также характеризующейся глобальным подходом к связности текста [Asher 1993], признаётся, что показателями дискурсивных отношений являются коннекторы. Таблица «показателей отношений» в [Muller et al. 2012: 27–28] содержит только единицы данного класса, причём подчёркивается, что коннекторы — многозначны и что указанные для них отношения (одно или более для каждого показателя) являются скорее ориентиром для разметчиков.

PDTB отличает «локальный» подход к связности текста (подробнее о различиях в аннотировании между RST и PDTB см. [Poláková et al. 2017]). Этот корпус задумывался с целью аннотировать коннекторы и выражаемые ими отношения [Prasad et al. 2006: 1]. Коннекторы подразделяются на эксплицитные и имплицитные. В последнем случае разметчик должен попытаться понять, какое отношение может быть установлено между двумя фрагментами текста, а аннотирование состоит в том, чтобы вставить в текст наиболее подходящий по смыслу коннектор. Так, в (2), согласно [Prasad et al. 2007: 17], между фрагментами текста могут быть вставлены два коннектора: *because* и *for example*, потому что второй фрагмент текста вводит одно из объяснений утверждения «The third principal in the S. Gardens adventure did have garden experience»:

- (2) {*The third principal in the S. Gardens adventure did have garden experience.*}
{(Импл. = ***because for example***) *The firm of Bruce Kelly/David Varnell Landscape Architects had created Central Park's Strawberry Fields and Shakespeare Garden.*}

Кроме того, PDTB проводит различие между имплицитным коннектором (*implicit connective*) и имплицитным отношением (*implicit relation*), поскольку случаи, когда между двумя фрагментами текста невозможно вставить коннектор разделяются на 1) альтернативные лексикализации (AltLex), когда отношение выражено другими лексическими средствами, традиционно не относящимися к классу коннекторов, а подстановка коннектора приводит к избыточности высказывания; 2) случаи, когда между фрагментами существует недискурсивное отношение, а именно, когда второй фрагмент текста описывает некоторый аспект ситуации, описанной в первом фрагменте текста (EntRel, *entity-based coherence* [Knott et al. 2001]); 3) случаи, когда между фрагментами текста не может быть установлено никакого отношения (NoRel).

Аннотации имплицитных коннекторов и контекстов, относящихся к одному из трёх перечисленных выше случаев, изначально объединялись в группу «*implicit relation annotations*» [Prasad et al. 2006: 1], однако затем эксплицитным дискурсивным отношениям были поставлены в соответствие 1) эксплицитные коннекторы

¹ Во всех примерах вне зависимости от исходного оформления связываемые фрагменты заключены в фигурные скобки, а языковые единицы, выражающие ЛСО, выделены полужирным шрифтом.

и 2) альтернативные лексикализации, а имплицитным дискурсивным отношениям — только случаи с имплицитным коннектором [Prasad et al. 2017: 1200–1201].

В разрабатываемом нами подходе и в НБД коннекторов² под имплицитными ЛСО понимаются такие отношения между двумя или более фрагментами текста, которые не выражены ни коннектором, ни каким-либо другим языковым средством (лексическим или грамматическим), ни семантически нагруженным знаком препинания (прежде всего, тире и двоеточием; подробнее см. [Инькова 2019б]), но которые могут быть выражены коннектором без необходимости модифицировать контекст. Последнее положение отличает наш подход от подхода PDTB.

Так, для нас в (3) отношение причины во фрагменте на русском языке выражено деепричастным оборотом, для которого традиционно выделяется это значение [Виноградов, Истрина 1960а: 650–656]; [Шведова 1970: 645]; [Шведова 1980: 181–182].

- (3) {*Вдруг Штольц изменился в лице,*} {*поймав истину.*}
[Гончаров И. А. Обломов (1848–1859)]

В то же время в PDTB аналогичный случай (4) [Webber et al. 2019: 34] относится к случаям с «имплицитным коннектором», и при аннотировании вводится коннектор причины *as result of*, что приводит к превращению формы *present participle* в форму *gerund* и к семантической избыточности, поскольку для подобных конструкций в английском языке также выделяется значение причины (см., например, [Stump 1985], где, в частности, приводится таблица, показывающая, что причинная семантика данных единиц выделяется во всех рассматриваемых автором работах по грамматике английского языка [Там же: 19]).

- (4) {*Treasuries opened lower,*} {(Импл. = *as a result of*) *reacting negatively to news that the producer price index — a measure of inflation on the wholesale level — accelerated in September.*} [wsj 2428]

При этом критерием, на основании которого в PDTB выделена группа AltLex, является как раз возникновение избыточности в случае добавления коннектора ([Prasad et al. 2007: 27]; [Webber et al. 2019: 12])³.

Наше понимание имплицитного ЛСО является, таким образом, более узким, чем принятое в PDTB, наиболее близком нам подходе, соответствуя только части случаев с «имплицитным коннектором».

² Представительный фрагмент НБД коннекторов доступен по адресу: <http://a179.frccsc.ru/RSCF10004/main.aspx>.

³ Заметим, что данный подход ориентирован на английский язык; ср. невозможность такой трансформации в русском и многих других языках: (4) {*Рынок казначейских бумаг открылся снижением,*} {(Импл. = **как результат* / **в результате* / **вследствие*) *негативно отреагировав на новости о том, что индекс цен производителей — показатель инфляции на оптовом уровне — в сентябре вырос.*} При возможном: *вследствие* негативной реакции...

3. Метод поиска имплицитных ЛСО в параллельных текстах

Плодотворность использования параллельных текстов для аннотирования ЛСО, как эксплицитных, так и имплицитных, уже не раз отмечалась исследователями. Так, в [Cartoni et al. 2013] показано, что переводной эквивалент коннектора может дать дополнительную информацию о его семантике и снять неоднозначность интерпретации в случае полисемичных коннекторов. В [Shi et al. 2017: 484] подчёркивается, насколько трудным является распознавание имплицитных ЛСО и что нейронные сети плохо справляются с этой задачей из-за нехватки аннотированных данных. Для решения этой практической проблемы в работе предлагается оригинальный метод использования параллельных корпусов (в данном случае англо-французского корпуса). Авторы исходят из положений о том, что профессиональные переводчики иногда добавляют коннекторы, отсутствующие в оригинале (т.н. эксплицитирование ЛСО), и что при машинном переводе уже имеющиеся в тексте коннекторы, как правило, не опускаются. Поэтому в рамках исследования выполняется обратный перевод текстов с французского на английский язык при помощи системы машинного перевода. После этого отбираются случаи, когда ЛСО в оригинальном фрагменте на английском языке является имплицитным, а во фрагменте, полученном при помощи обратного машинного перевода на английский язык, оно выражено коннектором — из-за того, что соответствующий коннектор был добавлен профессиональными переводчиками при переводе оригинального фрагмента на французский язык. Это позволяет, по мнению авторов, получить информацию высокой степени достоверности о ЛСО, содержащемся в исходном английском тексте, и создать представительный массив аннотаций для машинного обучения.

В работе [Гончаров, Инькова 2019а] был предложен иной метод поиска имплицитных ЛСО в параллельных текстах. В общем случае, если корпус таких текстов позволяет осуществлять поиск языковых единиц (ЯЕ) как в оригинале, так и в переводе, то можно находить 1) пары фрагментов, где в одном из языков присутствует некоторая ЯЕ; 2) пары фрагментов, где в одном языке употреблена ЯЕ x , а в другом языке — ЯЕ y . В первом случае нередко приходится работать с большим массивом случаев, во втором имеется возможность задавать лишь известные (большой частью стандартные) комбинации x и y .

Исходя из этого, представляется полезным сократить массив таким образом, чтобы в нём остались лишь неизвестные или нестандартные соответствия. Для этого в НБД была предусмотрена возможность поиска, позволяющая отбирать пары фрагментов, где во фрагменте на одном языке обязательно должна присутствовать заданная ЯЕ, а во фрагменте на другом языке должна/ы отсутствовать заданная/ые ЯЕ. На первом этапе работы была реализована лишь функция исключения одной или нескольких однословных ЯЕ [Гончаров, Инькова 2019а]. Эта функция показала эффективность такого вида поиска, поэтому ввиду формальных особенностей коннекторов, которые часто являются неоднословными (51,25% аннотаций в НБД на 15.02.2020), она была усовершенствована, что дало возможность исключать из выборки одну или несколько неоднословных единиц в фиксированной форме.

4. Анализ полученных результатов

Для настоящего исследования были использованы тексты параллельного французского подкорпуса НКРЯ (направление перевода «русский — французский») общим объёмом более 3,9 млн словоупотреблений (данные на 15.02.2020). Поиск был применён для анализа русских стимулов перевода трёх французских коннекторов — *car, parce que, puisque*, — выражающих ЛСО причины. Из выборки исключались пары фрагментов, которые содержат коннекторы русского языка *потому (что), оттого (что), поскольку, так как, ибо, за то что, ведь*. Безусловно, не во всех случаях стимул перевода был нулевым (ср. (3) выше), однако в данной работе мы остановимся лишь на таких примерах, которые мы объединяем в кластер Zero. Этот кластер в ходе анализа выборки был разделён на три подкластера:

1. Коннектор отсутствует, но может быть восстановлен (Absent CNT).

- (5) {*Совещанию придавалось особое значение,*} {(Импл. = **поскольку**)
оно связывалось с поездкой Гимmlера в полевую Ставку вождя.}
[Гроссман В. С. Жизнь и судьба (1960)]
{*Cette réunion était particulièrement importante,*} {**puisque** Himmler devait
se rendre ensuite au Quartier général du Führer.} [Пер. Berelowitch A. (1980)]

2. Фрагмент, соответствующий фрагменту с коннектором, имеет иную структуру (DifferStr FragmCNT); так, в (6) при переводе фрагмент текста, содержащий коннектор, меняет свою структуру: ‘никто не хочет брать меня на работу, потому что я из Чечни’.

- (6) {*На работу не хотят брать,*} {“А-а-а... из Чечни?”}
[Алексиевич С. А. Время секунд хэнд (ч. 2) (2013)]
{*Personne ne veut m'employer*} {**parce que** je reviens de Tchétchénie...}
[Пер. Benech S. (2013)]

3. Соответствующий фрагмент с коннектором отсутствует (Absent FragmCNT); ср. (7), где переводчик вводит дополнительный фрагмент ‘вам же хуже’ (*tant pis pour vous*), с которым он связывает отношением причины фрагмент ‘вам же нечего будет в пирожное положить’. В оригинале фрагменты текста *просыплю корицу и вам же нечего будет в пирожное положить* связаны ЛСО условия.

- (7) — *Смотрите, {просыплю корицу;}* {*вам же нечего будет в пирожное положить,*} — *заметила она.* [Гончаров И. А. Обломов (1848–1859)]
— *Attention, je vais renverser la cannelle; {tant pis pour vous} {car je n'aurai rien à mettre dans votre gâteau,}* *fit-elle remarquer.* [Пер. Jurgenson L. (1988)]

Несмотря на то, что изменения контекста при переводе, приводящие к соответствиям типа 2. и 3., могут быть оправданными, мы не рассматриваем их как случаи имплицитных ЛСО. Тем не менее, в [Shi et al. 2017: 491] особо отмечается важность исследования контекстов, сходных с теми, которые мы относим к типу 2. Хотя такие контексты 1) не включают никакого коннектора, 2) имеют структуру,

не позволяющую его добавить, и 3) могут не рассматриваться как содержащие какое-либо ЛСО, они вызывают появление коннектора в переводе, а потому могут дать новую информацию о механизмах функционирования ЛСО в целом и расширить имеющиеся знания о спектре контекстов, включающих ЛСО.

Таблица ниже содержит данные о количестве переводных соответствий, относящихся к кластеру Zero, и их распределение по трём подкластерам (использование различных речевых реализаций коннекторов в таблице не отражается, т.е. случаи, например, с *car*, *car enfin*, *car... et* объединены в кластер *car*). Все процентные показатели приведены по отношению к общему числу пар с этим коннектором в НБД.

Коннектор франц. яз.	Стимул перевода в русском языке								Всего пар в НБД
	Zero		Absent CNT		DifferStr FragmCNT		Absent FragmCNT		
	Абс.	%	Абс.	%	Абс.	%	Абс.	%	
<i>car</i>	178	18,46	148	15,35	21	2,18	9	0,93	964
<i>parce que</i>	44	7,84	34	6,06	10	1,78	0	0	561
<i>puisque</i>	18	9,78	12	6,52	5	2,72	1	0,54	184
	240		194		36		10		1709

Таблица показывает, прежде всего, эффективность предлагаемого метода поиска: подавляющее большинство полученных аннотаций содержат именно контексты с имплицитным ЛСО причины в русском оригинале, т.е. когда коннектор отсутствует, но может быть восстановлен. Во-вторых, можно заметить, что изменение структуры фрагмента текста с коннектором или добавление фрагмента текста в перевод не являются частотными. Наконец, из таблицы видно, что причинные коннекторы имеют различную дистрибуцию: для коннектора *car* нулевой стимул перевода встречается гораздо чаще, чем для других коннекторов причины. На наш взгляд, это вызвано тем, что данный коннектор выражает более «слабое» причинное отношение, которое в оригинале установлено, как правило, между самостоятельными предложениями вне зависимости от их объёма; клаузами в составе сложного предложения, разделёнными запятой или точкой с запятой; частицами «да» или «нет», выражающими согласие или несогласие, и причинами этих ответов; частями одной и той же реплики, разделёнными авторской речью.

5. Выводы

Описанный вид поиска позволяет оперативно получить представительные массивы случаев, когда ЛСО устанавливается без использования прототипических средств. Аннотирование этих случаев даёт возможность получить количественные показатели средствами НБД, а дальнейший анализ позволяет 1) классифицировать случаи неспециализированного выражения ЛСО, играющих, как известно, ведущую роль в обеспечении связности текста (такая классификация является одной из задач теоретической и когнитивной семантики); 2) классифицировать случаи, когда ЛСО является имплицитным, что может быть использовано

при разработке систем автоматической обработки текста, в том числе его семантического анализа; 3) сравнить эффективность существующих методов поиска имплицитных ЛСО. В данной работе было предложено решение только для второй задачи. О методах решения первой задачи см., например, [Инькова 20196]; [Гончаров, Инькова 20196]), где предложена в общем виде и проиллюстрирована типология средств выражения ЛСО, альтернативных коннекторам.

Достоинством предлагаемого подхода является как возможность рассмотрения конкретных языковых единиц (насколько часто стимул перевода конкретного коннектора является нулевым), так и — впоследствии — ЛСО в целом (насколько часто случаи имплицитного ЛСО эксплицируются в переводе; какие ЛСО чаще других являются имплицитными). Привлечение данных русского и французского языков позволит верифицировать гипотезы, касающиеся общих принципов функционирования языка, сформулированные на основе данных английского языка, в частности «causality-by-default hypothesis» [Sanders 2005]. Отметим также, что в работах, содержащих описание методов автоматизации поиска имплицитных ЛСО, степень детализации результатов гораздо ниже, поскольку случаи имплицитных ЛСО распределяются по группам отношений, а их классификации, как известно, в значительной степени не совпадают.

Перспективами исследования является получение количественных данных по коннекторам, выражающим другие ЛСО, что позволит в дальнейшем перейти от исследования случаев, когда ЛСО выражено коннектором в оригинале или в переводе, к целенаправленному поиску случаев имплицитных ЛСО, не ограниченному использованием параллельных корпусов.

Литература

1. Asher N. (1993) Reference to abstract objects in discourse. Kluwer Academic Publishers, Dordrecht.
2. Carlson L., Marcu D. (2001) Discourse Tagging Reference Manual. ISI Technical Report ISI-TR-545, available at: <https://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf>.
3. Carlson L., Marcu D., Okurowski M. E. (2003) Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory, Current and New Directions in Discourse and Dialogue, Kluwer Academic Publishers, Dordrecht, pp. 85–112.
4. Cartoni B., Zufferey S., Meyer T. (2013) Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique, Dialogue and Discourse, Vol. 4, No. 2, pp. 65–86.
5. Das D., Taboada M. (2013). Explicit and Implicit Coherence Relations: A Corpus Study, Proceedings of the 2013 annual conference of the Canadian Linguistic Association, available at: http://homes.chass.utoronto.ca/~cla-acl/actes2013/Das_and_Taboada-2013.pdf.
6. Das D., Taboada M. (2014) RST Signalling Corpus Annotation Manual, available at: https://www.sfu.ca/~mtaboada/docs/publications/RST_Signalling_Corpus_Annotation_Manual.pdf.

7. *Goncharov A. A., Inkova O. Yu.* (2019a) Methods for identification of implicit logical-semantic relations in texts [Metodika poiska implitsitnykh logiko-semanticheskikh otnosheniy v tekste], Informatics and Applications [Informatika i ee Primeneniya], Vol. 13, No. 3, pp. 97–104.
8. *Goncharov A. A., Inkova O. Yu.* (2019b) Means of Expressing Causal Relations in Russian: Analysis Using a Cross-Linguistic Supracorpora Database [Sposoby vyrazheniya prichinnykh otnosheniy v russkom yazyke: opyt analiza s ispol'zovaniem krosslingvisticheskoy nadkorporusnoy bazy dannykh], Russian Grammar: Active Processes in Language and Discourse. Proceedings of International Scientific Symposium [Russkaya grammatika: aktivnye protsessy v yazyke i rechi. Sbornik nauchnykh trudov Mezhdunarodnogo nauchnogo simpoziuma], Yaroslavl, pp. 385–396.
9. *Inkova O. Yu.* (2019a) Logical-semantic relations: classification problems [Logiko-semanticheskie otnosheniya: problemy klassifikatsii], Text coherence: mereological logical-semantic relations [Svyaznost' teksta: mereologicheskie logiko-semanticheskie otnosheniya], Izdatel'skii Dom YaSK, Moscow, pp. 11–98.
10. *Inkova O. Yu.* (2019b) Annotation of Parallel Texts: The Concept of Divergent Translation [Annotirovanie paralel'nykh tekstov: ponyatie «divergentnyy perevod»], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii «Dialog»], Moscow, pp. 227–238.
11. *Knott A., Oberlander J., O'Donnell M., Mellish C.* (2001) Beyond elaboration: the interaction of relations and focus in coherent text, Text Representation: Linguistic and Psycholinguistic Aspects, John Benjamins, Amsterdam, pp. 181–196.
12. *Mann W. C., Thompson S. A.* (1986) Relational Propositions in Discourse, Discourse Processes, Vol. 9, pp. 57–90.
13. *Martin J. R.* (1992) English Text: System and Structure, John Benjamins, Amsterdam and Philadelphia.
14. *Muller Ph., Vergez-Couret M., Prévot L., Asher N., Benamara F., Bras M., Le Draoulec A., Vieu L.* (2012) Manuel d'annotation en relations de discours du projet AN-NODIS, Carnets de Grammaire 21. Rapports internes de CLLE-ERSS, available at: <http://w3.erss.univ-tlse2.fr/textes/publications/CarnetsGrammaire/carnGram21.pdf>.
15. *Poláková L., Mírovský J., Synková P.* (2017) Signalling Implicit Relations: A PDTB—RST Comparison, Dialogue & Discourse, Vol. 8, No. 2, pp. 225–248.
16. *Prasad R., Miltsakaki E., Dinesh N., Lee A., Joshi A., Robaldo L., Webber B. L.* (2007) The Penn Discourse Treebank 2.0 Annotation Manual. IRCS Technical Reports Series, available at: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1203&context=ircs_reports.
17. *Prasad R., Miltsakaki E., Dinesh N., Lee A., Joshi A., Webber B. L.* (2006) The Penn Discourse TreeBank 1.0 Annotation Manual. IRCS Technical Reports Series, available at: https://repository.upenn.edu/cgi/viewcontent.cgi?referer=https://www.google.fr/&httpsredir=1&article=1003&context=ircs_reports.

18. *Prasad R., Webber B., Joshi A.* (2017) The Penn Discourse Treebank: An Annotated Corpus of Discourse Relations, *Handbook of Linguistic Annotation*, Springer Science+Business Media, Dordrecht, pp. 1197–1217.
19. *Renkema J.* (2004) *Introduction to Discourse Studies*, John Benjamins Publishing Company, Amsterdam / Philadelphia.
20. *Sanders T. J. M.* (2005) Coherence, causality and cognitive complexity in discourse, *Proceeding SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, pp. 105–114.
21. *Shi W., Yung F., Rubino R., Demberg V.* (2017) Using Explicit Discourse Connectives in Translation for Implicit Discourse Relation Classification, *Proceedings of The 8th International Joint Conference on Natural Language Processing*, Taipei, pp. 484–495.
22. *Stump G. T.* (1985) *The Semantic Variability of Absolute Constructions*, D. Reidel Publishing Company, Dordrecht.
23. *Taboada M.* (2006) Discourse markers as signals (or not) of rhetorical relations, *Journal of Pragmatics*, Vol. 38, No. 4, pp. 567–592.
24. *Taboada M.* (2009) Implicit and explicit coherence relations, *Discourse, of Course*, John Benjamins Publishing Company, Amsterdam, pp. 127–140.
25. *Taboada M., Das, D.* (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations, *Dialogue and Discourse*, Vol. 4, No. 2, pp. 249–281.
26. *Webber B., Prasad R., Lee A., Joshi A.* (2019) The Penn Discourse Treebank 3.0 Annotation Manual, available at: <https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf>.
27. *Vinogradov V. V., Istrina E. S.* (eds.) (1960a) *Grammar of Russian Language. Vol. II. Syntax. Part 1* [Grammatika russkogo yazyka. T. II. Sintaksis. Ch. 1.], Izdatel'stvo AN SSSR, Moscow.
28. *Vinogradov V. V., Istrina E. S.* (eds.) (1960b) *Grammar of Russian Language. Vol. II. Syntax. Part 2* [Grammatika russkogo yazyka. T. II. Sintaksis. Ch. 2.], Izdatel'stvo AN SSSR, Moscow.
29. *Shvedova N. Yu.* (ed.) (1970) *Grammar of Modern Standard Russian Language* [Grammatika sovremenного russkogo literaturnogo yazyka], Nauka, Moscow.
30. *Shvedova, N. Yu.* (ed.) (1980) *Russian Grammar. Vol. II. Syntax* [Russkaya grammatika. Tom II. Sintaksis.], Nauka, Moscow.

ВИДОВЫЕ ТРОЙКИ РУССКОГО ГЛАГОЛА В ДИАХРОНИИ (НА МАТЕРИАЛЕ НКРЯ)¹

Горбова Е. В. (elena-gorbova@yandex.ru)

РГПУ им. А. И. Герцена, Санкт-Петербург, Россия

DOI: 10.28995/2075-7182-2020-19-321-347

ASPECTUAL TRIPLETS OF THE RUSSIAN VERB IN DIACHRONY: EVIDENCE FROM THE RUSSIAN NATIONAL CORPUS

Gorbova E. V. (elena-gorbova@yandex.ru)

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

The paper deals with the so-called aspectual triplets of the Russian verb. Based on the data from the Russian National Corpus, it proposes a diachronic method to study triplets as well as a two-component model of the Russian aspect as an alternative to the traditional word-based classification model. The first component of the model is a morphological mechanism of the imperfectivizing suffixation of prefixed verbs that is inflectional (*ras-kry-t^{IPFV}* — *ras-kry-va-t^{IPFV2}* ‘disclose, reveal’), but has a limited scope of action (prefixed verbs only). The second component of the model is the actionality (lexical aspect) with a maximal scope. Related to the verb class as a whole, it is especially crucial for non-prefixed simplexes. Actionality enables the functioning and perfective / imperfective characterization of simplexes which do not fall under the inflectional grammatical aspect. The analysis of ten biimperfective triplets resulted in several observations and conclusions. One of them concerns the role of a ‘joker’, which all imperfective simplexes (IPFV₁) have in the aspectual triplets as (quasi)synonyms for corresponding secondary imperfectives (IPFV₂). A working hypothesis on the predominance of IPFV₁ over PFV in every triplet, based on the broader polysemy of the former, has not been confirmed. However, the two-component model has explanatory power for the cases of reverse frequency (PFV over IPFV₁) through its lexical aspect component. Another working hypothesis on a possible increase or a decrease in the number of secondary imperfectives in diachrony was partially confirmed — an increase was noted for the 20/21st century.

Key words: Russian verb aspect, biimperfective aspectual triplets, actionality (lexical aspect), prefixed verbs, primary imperfective, secondary imperfective

¹ Исследование поддержано грантом РФФИ 19-012-00143 А «Суффиксальная имперфективация приставочных перфективов в современном русском языке». Автор благодарит Е. Еникееву за помощь в статистической обработке полученных результатов.

1. Введение и постановка задачи

Предметом анализа являются биимперфективные видовые тройки типа *гореть*^{НСВ₁} — *сгореть*^{СВ} — *сгорать*^{НСВ₂}, в последние четверть века активно обсуждаемые в литературе [Апресян 1995], [Апресян 1997]; [Петрухина 2000], [Петрухина 2009]; [Зализняк, Шмелев 2000]; [Ясаи 2001]; [Храковский 2005], [Храковский 2006]; [Томмола 2008]; [Зализняк, Микаэлян 2010]; [Татевосов 2010]; [Зализняк, Микаэлян, Шмелев 2015]; [Janda, Lyashevskaya 2011]; [Янда 2012]; [Janda et al. 2013]; [Горбова 2014], [Горбова 2015]; [Wiemer 2019]. Однако мы попробуем взглянуть на них под новым углом зрения.

Согласно двухкомпонентной теории русского вида, представленной в [Горбова 2019a], имперфектив-симплекс (НСВ₁) видовой тройки (*гореть*) является моноаспективом (*imperfectivum tantum*) и принадлежит к акциональному классу состояний либо неопредельных процессов. Префигированный перфектив (СВ, *сгореть*) и суффиксально образованный от него префигированный же имперфектив (*сгорать*), т. наз. вторичный имперфектив (НСВ₂), являются формами одной лексемы, образуя видовую пару в рамках бинарной словоизменительной категории вида. По своей акциональной характеристике такая лексема чаще всего является предельным процессом (*сгорать* — *сгореть*). В рамках предложенной теоретической модели «избыточным» в биимперфективной тройке неизменно является первичный имперфектив-симплекс (НСВ₁), а его семантическая близость к префигированному вторичному имперфективу (НСВ₂) объясняется (квази)синонимией.

При этом признается функциональное сходство НСВ₁ и НСВ₂: одинаковый набор временных форм (аналитическое будущее *буду* + инфинитив, настоящее, прошедшее) и сочетаемость с матричными фазовыми глаголами (^{ок}*начать, продолжить, кончить гореть и сгорать* vs. ^{*}*начать, продолжить, кончить сгорать*), что подтверждено исследованием грамматических профилей префиксальных и суффиксальных видовых пар в [Janda, Lyashevskaya 2011]. Однако источники аналогичного функционирования двух имперфективов принципиально различны. Это акциональность (лексический вид) в случае имперфективов-симплексов и морфологический показатель имперфективности (суффикс *-(ы/и)ва-* или *-а-*) в случае префигированных НСВ₂.

НСВ₁ и НСВ₂ различаются не только своими морфологическими, но также синтаксическими и семантическими свойствами. Синтаксическое различие таково: НСВ₂ наследует переходность (при ее наличии) от префигированного СВ, в то время как НСВ₁ может быть непереходным [Карцевский 1962], [Храковский 2005], [Татевосов 2010]. Ср. лабильный по переходности *бить* <камнем по дереву> и *бить* <чашку>, и исключительно переходные *разбить* / *разбивать* <чашку>. Со стороны семантики можно отметить «неравнообъемность семантического потенциала» первичного и вторичного имперфектива [Апресян 1995]: возможные затруднения НСВ₂ в реализации процессного

² Используемые здесь и далее обозначения НСВ₂ для вторичного имперфектива и НСВ₁ — для первичного, а также понятие биимперфективной тройки, были введены в работе [Анна Зализняк, Микаэлян 2010].

значения, а также особую роль НСВ₁ — роль «джокера», т.е. (контекстуального) заместителя НСВ₂, ср.: *бить (врага) — побить — побивать; бить (Часы бьют) — пробить — пробивать; бить (посуду) — разбить — разбивать; бить (масло) — сбить — сбивать* [Апресян 1995: 106].

В [Горбова 2015: 28] были сформулированы общие принципы образования биимперфективных троек с квазисинонимией двух имперфективов. Они сводятся к следующему. Тройка естественна и ожидаема в тех случаях, когда НСВ₁ является либо лексикографически предельным глаголом, либо обозначает неопредельную ситуацию, которая в контексте легко преобразуется в предельную, а НСВ₂ — это соотносенный с пределом процесс: *читать — прочитывать <статью>, грузить — на(по, за)грузить <дрова, вещи, кузов>*. И не ожидаема в том случае, когда НСВ₁ является состоянием (неопредельным процессом), а НСВ₂ образован от СВ-моментатива; в таком случае вторичная имперфективация осуществляет итеративизацию единичной ситуации (этап СВ), причем эта ситуация лексикографически может не соотноситься с НСВ₁: *знать^{НСВ1} → узнать^{СВ} → узнавать^{НСВ2}; идти^{НСВ1} → прийти^{СВ} → приходиться^{НСВ2}: знать <историю> ≠ узнавать <историю>, идти <домой, вдоль улицы> ≠ приходиться <домой>*.

Двухкомпонентная модель русского вида [Горбова 2019а] противопоставляет суффиксальную имперфективацию как словоизменительный механизм видообразования и префиксальную перфективацию как словообразовательный механизм, в норме сопровождаемый телисизацией и перфективацией. Имперфективное поведение большинства симплексов объясняется акциональностью (лексическим видом) глагольной лексемы. Тем самым предлагается новое объяснение сходств и различий между (квази)синонимичными НСВ₁ и НСВ₂ в тройке. Кроме того, модель предсказывает увеличение частотности НСВ₂ как регулярной видовой формы для префигированного СВ в общем случае. Верификация этих положений осуществляется с помощью синхронного (в рамках каждого временного среза) и диахронического анализа материала, полученного путем обращения к НКРЯ, см. [разделы 2–4](#).

2. Методика исследования

Для анализа биимперфективных троек основной корпус НКРЯ был разбит на четыре пользовательских подкорпуса по времени создания текста: 18 в. (1700–1799), 19 в. (1800–1899), 20 в. (1900–1999) и 21 в. (2000–2019). Даты обращения — 09.01.2020–27.01.2020 (использована старая версия НКРЯ).

Объемы созданных подкорпусов см. в [Таблице 1](#).

Табл. 1. Объемы созданных на базе основного корпуса НКРЯ подкорпусов (в документах / словах)

18 в.	19 в.	20 в.	21 в.
1 838 / 5 202 929	5 424 / 54 274 217	68 844 / 152 456 773	39 928 / 76 144 419

Примем объем подкорпуса 19 в. в словах за основание для сравнения (единицу). Тогда объемы остальных подкорпусов в относительных цифрах таковы, см. **Таблицу 2**.

Табл. 2. Относительные объемы созданных на базе основного корпуса НКРЯ подкорпусов (в словах)

18 в.	19 в.	20 в.	21 в.
0,1	1	2,8	1,4

Отбор биимперфективных троек для анализа производился с опорой на «Фрагмент словаря видовых троек» [Зализняк и др. 2015: 242–247], содержащий 34 тройки, из которых были отобраны пять³: *отправиться* — *травиться* / *отравляться*⁴, *переночевать* — *ночевать* / *переночевывать*, *свариться* — *вариться* / *свариваться*, *съесть* — *есть* / *съесть*, *ужалить* — *жалить* / *ужаливать*. Еще пять троек (после соответствующего анализа на (квази)синонимичность НСВ₁ и НСВ₂⁵) были добавлены нами: *оторвать* — *рвать* / *отрывать*, *пробить* — *бить* / *пробивать*, *пробормотать* — *бормотать* / *пробормотывать*, *сгореть* — *гореть* / *сгорать*, *сорвать* — *рвать* / *срывать*. Итак, анализируются 10 видовых троек, в которых СВ и НСВ₂ имеют префиксы: *о-* (1 случай), *от(о)-* (1), *пере-* (1), *про-* (2), *с(о)-* (4), *у-* (1).

Работа с каждой тройкой включала в себя следующие этапы.

1. В каждом из созданных подкорпусов посредством лексико-грамматического поиска⁶ формируется запрос на НСВ₁, СВ и НСВ₂; полученные данные (абсолютные количества вхождений после ручной очистки от «шума») представлены в виде графиков (см. **Рис. 1–10** ниже).

³ С исходной, но непоследовательной выдержанной, ориентацией на префиксы *про-*, *с(о)-*, *у-*, с которыми автор работал в [Горбова 2019b]; тройка с префиксом *про-* [Зализняк и др. 2015] одна: *проворонить* — *воронить* / *проворонивать*, она была заменена на тройку с более частотным симплексом *бормотать* (43 вхождения *воронить* и 8129 — *бормотать* по основному корпусу НКРЯ).

⁴ Графическая подача и порядок следования членов тройки здесь полностью дублирует источник, (Зализняк и др., 2015: 242–247). Ниже порядок изменяется в соответствии с принятой концепцией.

⁵ Анализ предполагал обнаружение в НКРЯ фрагментов текстов, в которых симплекс реализовал бы значение соответствующего НСВ₂, например, для *отрывать* <пуговиц(у/ы)> — *рвать* <пуговиц(у/ы)>: *Она сидела под бруствером, обеими руками рвала, расстегивала пуговицы на груди, видимо, оглушенная, с закрытыми глазами.* [Юрий Бондарев. *Горячий снег*, 1969], см. также (1a) ниже.

⁶ Семантические признаки в запрос не вводились, поскольку последовательное и результативное применение этого инструмента оказалось невозможным. Напр., при выборе для *бить* семантического признака *t:impact:destr* (уничтожение) поиск по подкорпусу 20 в. объемом более чем в 152 млн слов не дает результата, однако двухвалентный *бить* в значении 'убивать' или 'разбивать' при ручном анализе выдачи для *бить* в том же подкорпусе (5 050 документов, 25 347 вхождений) обнаруживается. С другой стороны, полная ручная обработка объемных выдач была сочтена неосуществимой, вследствие чего принято решение в случае объемных выдач ограничиться 330 документами (по 10–15 из каждой сотни документов).

2. Высчитывается процент увеличения / уменьшения вхождений НСВ₁, СВ и НСВ₂ в каждой тройке по созданным подкорпусам (Percentage increase/decrease, см. Brezina 2018: 230⁷), полученные данные сводятся в таблицу (Табл. 3–12). Такое сравнение позволяет оценить изменение частотности НСВ₁, СВ и НСВ₂ на разных временных срезах в каждой тройке и троек между собой.
3. Путем анализа выданных НКРЯ (не менее 330 документов) устанавливаются типичные актанты НСВ₁, СВ и НСВ₂ в каждой тройке. Далее проводится попарное (НСВ₁ и СВ; СВ и НСВ₂; НСВ₁ и НСВ₂) сравнение выявленной сочетаемости как индикатора лексической семантики однокоренных глаголов в каждой тройке.

Далее в **разделе 3** сформулированы рабочие гипотезы. В **разделе 4** приводятся результаты исследования видовых троек по обозначенной выше схеме. В **разделе 5** подведены итоги исследования.

3. Рабочие гипотезы

Кроме ожиданий, естественным образом вытекающих из изложенной в разделе 1 двухкомпонентной теории русского вида, а именно: в биимперфективной тройке именно НСВ₂ (но не НСВ₁) не отличается от СВ в плане лексической семантики, в силу чего неизменно избыточен — как потенциальный «джеркер» — НСВ₁, отметим следующее.

- а. В тройке частотность НСВ₁ превышает частотность не только НСВ₂, но и СВ, поскольку лексическая семантика симплекса (как вокабулы) отличается широкой полисемией, а семантика префиксальных СВ и НСВ₂ сужена за счет ее конкретизации префиксом (равно в синхронии и диахронии).
- б. По отношению к НСВ₂ существует тренд — либо увеличение его частотности в диахронии, либо, напротив, уменьшение. В первом случае можно говорить о последовательном развитии словоизменительной модели видообразования (суффиксальной имперфективизации), во втором — о последовательном же угасании этого механизма и предпочтении образования видовой пары типа НСВ₁ / СВ путем использования нерегулярной префиксации. Последнее — аргумент в пользу словоклассифицирующей (деривационной) модели русского вида на современном этапе развития этой категории.

⁷ Для оценки изменения частотности относительно выбранного основания для сравнения — подкорпуса 19 в. — использовалась следующая формула (модификация формулы из [Brezina 2018: 230]):

$$\frac{\text{кол-во вхождений в подкорпусе}}{\text{относительный объем}} - \text{количество вхождений в подкорпусе 19 в.} \\ \frac{\text{количество вхождений в подкорпусе 19 в.}}{\text{количество вхождений в подкорпусе 19 в.}} \times 100$$

где относительный объем каждого подкорпуса — величина из **Табл. 2**.

4. Полученные в результате корпусного исследования данные

4.1. рвать — оторвать / отрывать

Запрос: *рвать* (граммат. признаки — V, tran).

Сочетаемость⁸: 18 в. — *рвать* что-л. <растения и их части>, <зубы>; чье-л. <тело>, безл. (кого-л. *рвет*); *рвать на себе волосы*;

19 в. — *рвать* что-л. с чего-л. <цепи, одежду, петли, пуговицы>; что-л. <бумагу, текст>; кого / что-л. <на части>; (о ветре) <шляпу, ставни, облака>; что-л. кому-л. чем-л. <уши, ноздри, бороды, рожи; душу, сердце>; что-л. откуда <пистолет, известия, деньги>; (о шипучем напитке) <пробку>; что-л. чем-л. <динамитом, бомбами>; что-л. <из рук>; что-л. <отношения, связи> с кем/чем-л., <злобу> на ком-л.; *рвать и метать*;

20 в. — *рвать* что-л. <тельняшку, свитер; заевшую молнию, круг>, что-л. с кого-л. <белье, одежду>; (себе) что-л. <жилы, глотку>; <мехи (гармони)>, *рвать* (со старта); <ленточку грудью (на финише)>; *рвать подметки на ходу*; 'нарушать' что-л. <тишину, здоровье, оборону>; *поводок* (о животном); *рвать когти*;

21 в. — *рвать* кого-л., что-л. <в клочья>; <изнутри>; <нервы> кому-л.; <кусок> от кого-л.

Запрос: *оторвать* (V, tran).

Сочетаемость: 18 в. — *оторвать* кого-л. от чего / кого-л.; кому-л. чем-л. <часть тела>;

19 в. — *оторвать* что-л. от чего-л.; что-л. <с руками>; (не) *оторвать* <глаз, взора, взгляда> от кого / чего-л.; 'резко ответить, оборвать';

20 в. — *оторвать* <когти>; (от себя) что-л., <песню>; что-л., где-л. ('раздобыть'), *оторви да брось*;

21 в. — (не) *оторвать* кого-л. от чего-л.

Запрос: *отрывать* (V, tran).

18 в. — *отрывать* <куски> от чего / кого-л., кому-л. / от кого-л. <часть тела>;

19 в. — *отрывать* кого-л. от чего-л. <дела, занятия>; (не) *отрывать* <глаз, взора, взгляда> от чего / кого-л., что-л. от /у чего-л. (<вуаль от шляпы, ворот, пуговицу у сюртука>); что-л. по чему-л. <по буковке, по кусочку>, кого / что-л. <от сердца>;

20 в. — *отрывать* (от себя) что-л.; (не) *отрывать* что-л. от чего-л. <перо, карандаш от бумаги, кисть от холста, ноги от земли>; кому-л. <руки> (о тяжести); что-л. <с руками>;

21 в. — *отрывать* <на ходу подметки>; <время> от чего-л., кого-л., что-л. (от себя).

⁸ Для первых двух троек: *рвать* — *оторвать* / *отрывать* и *рвать* — *сорвать* / *срывать* сочетаемость представлена с ограничениями — при повторении в последующем веке, тип сочетаемости не приводится, новые фиксируются.

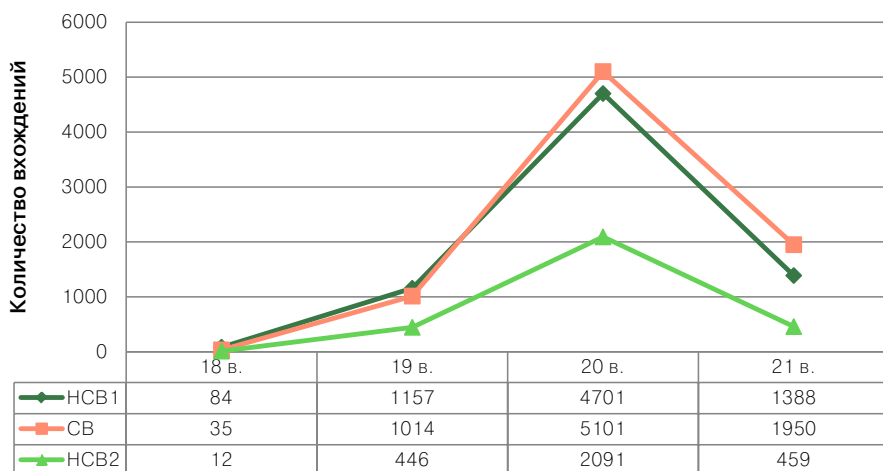
рвать — оторвать/отрывать

Рис. 1. Абсолютные количества HCB₁, CB и HCB₂ тройки *рвать — оторвать / отрывать* в выдачах НКРЯ (в словах)

Табл. 3. Увеличение / уменьшение частот HCB₁, CB и HCB₂ тройки *рвать — оторвать / отрывать* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
HCB ₁	–27%	—	+45%	–14%
CB	–65%	—	+80%	+37%
HCB ₂	–73%	—	+67%	+6%

4.2. рвать — сорвать / срывать

Запрос для HCB₁ *рвать* и данные о его сочетаемости см. в [разделе 4.1](#).

Запрос: *сорвать* (V, tran).

Сочетаемость: 18 в. — *сорвать* что-л. <цветы, плоды>; <вывеску, печать>; что-л. с кого-л. <мундир, маску>; (о ветре, буре) что-л. <мачту>; кому-л. что-л. чем-л. <ядром голову>; безл. (кого-л. *сорвет*);

19 в. — *сорвать* что-л. с чего-л. <колпак с головы; флаг>; с кого-л. / где-л. <(денег) на водку>; <куш, банк>; на ком-л. что-л. <гнев, злость>; <поцелуй с губ> кого-л.; что-л. <мероприятие, планы>;

20 в. — *сорвать* <голос>; <аплодисменты>; кого-л. с чего-л. <с урока>; 'привести в негодность' что-л. <кран, руль>;

21 в. — *сорвать* что-л. <чеку, стоп-кран >; у кого-л. *сорвать* крышу.

Запрос: *срывать* (V, tran).

18 в. — *срывать* что-л. с чего-л. <плоды, фрукты, растения>; что-л. с чего-л. <завесу, покров>; что-л. с кого-л. <одежду, погоны>; (о ветре, вихре, буре) что-л. <кровлю>, с чего-л. *барыши*;

19 в. — *срывать* что-л. <мероприятие, планы>; <злость, характер> на ком-л.; <банк>; <аплодисменты>; <цветы удовольствия>; что-л. <сердце>; что-л. у кого-л. <улыбку>; что-л. с кого-л. <личину, маску>;

20 в. — *срывать* <глотку>;

21 в. — *срывать* <крышу, башню> у кого-л.; кого-л. с чего-л. чем-л. <с места течением; людей с родных мест>; <куш>; <машину с места>.

рвать — сорвать/срывать

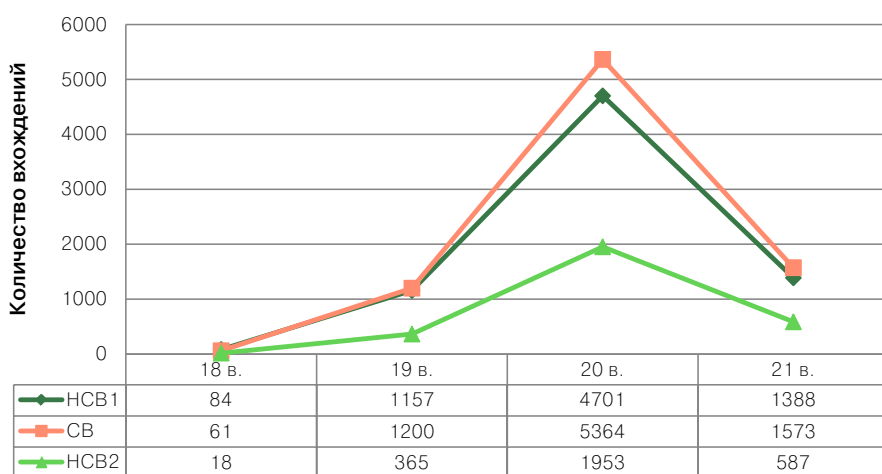


Рис. 2. Абсолютные количества НСВ₁, СВ и НСВ₂ тройки *рвать — сорвать / срывать* в выдачах НКРЯ (в словах)

Табл. 4. Увеличение / уменьшение частот НСВ₁, СВ и НСВ₂ тройки *рвать — сорвать / срывать* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
НСВ ₁	-27%	—	+45%	-14%
СВ	-49%	—	+60%	-6%
НСВ ₂	-51%	—	+91%	+15%

4.3. гореть — сгореть / сгорать

Запрос: гореть

Сочетаемость: 18 в. — *гореть* кто-л., что-л. \emptyset , чем-л. <кто-л. нетерпением; хоромы, степь, свечи, окрестности, спирт, щека; все огнем >; как жар *гореть*;

19 в. — *гореть* кто-л., что-л. \emptyset , чем-л., от чего-л. <кто-л. желанием, нетерпением, патриотическим пламенем; свечи, огни, лес, костер, трава, папиросы, любовь, заря, люки, поручни, орудия, компас; глаза огнем; голова от наплыва мыслей>; *X (весь) горит*;

20 в. — *гореть* кто-л., что-л. \emptyset , чем-л., от чего-л. <я нетерпением; лицо, свет, костер, Москва, пожары, буквы, путевки, уши, Южный Крест, звезды, огни, фары, одежда; глаза ненавистью, злобой; шея от пота>; *гори оно всё ясным пламенем; (это) в огне не горит и в воде не тонет; (кто-л.) гореть, как в огне; (у кого-л.) всё в руках горит; \emptyset горит, что ли?*

21 в. — *гореть* кто-л., что-л. \emptyset , чем-л., от чего-л., в чем-л. <кто-л. желанием; солдаты в танке; свет, красный (свет светофора), дом, камин, поля, лук, тесто, дворец, свечи, габариты; лицо; мухоморы, крыша золотом>; *гори все синим огнем.*

Запрос: сгореть

Сочетаемость: 18 в. — *сгореть* что-л. <дом, завод, хлеб, свечи>; (о ком-л.) *со стыда сгореть*;

19 в. — *сгореть* что-л., кто-л. в чем-л., на чем-л., от чего-л./с чего-л. <двор, дом, керосин, книги, день, часть Москвы; кто-л. от водки, с вина; в бессильной жажде; на работе, на открытом воздухе>; (о ком-л.) *со стыда сгореть*;

20 в. — *сгореть* кто-л. в чем-л., от чего-л., в/за X времени <документы, хата, дом, письма; все; день; путевка; кто-л. в неделю; от смущения; в тифу>; (о ком-л.) *со стыда сгореть*;

21 в. — *сгореть* кто-л. в чем-л., на чем-л., от чего-л./с чего-л. <деньги, вклады, топливо, телевизор, книги; дом, все; кто-л. на солнцепеке, от чахотки>.

Запрос: сгорать

Сочетаемость: 18 в. — *сгорать* кто-л. от чего-л., в чем-л. <от любви; во мщени>; что-л. <свечи, вещи, деревья>;

19 в. *сгорать* — кто-л. чем-л./от чего-л. к кому-л.; <грешник, человек, животное; желанием, нетерпением, любовью, страстью к X, от любопытства, от счастья, от нетерпения, от смущения>; что-л. <вещества, дрова, деревня, золото, масло, тела, огонь, хлеба, порох, Москва>; кто-л. *на костре страсти*;

20 в. *сгорать* — кто-л. чем-л./от чего-л. к кому-л., в чем-л. <от стыда и конфуза, от ненависти, от неловкости, от любви, от нетерпения; на работе; в огне>; что-л. чем-л., от чего-л./в чем-л./на чем-л. <письма, солнце, фирмы, стебель, топливо, керосин, масло, контора, здания, капли жира голубоватым пламенем; листья в пламени; деньги от инфляции; кожа на солнце>;

21 в. — *сгорать* кто-л. от чего-л., в чем-л., под чем-л. <от стыда, от нетерпения, от любопытства; в работе; под лучами солнца; на солнце>; что-л. <закаись азота, все, сумма, топливо, адреналин, спичка, мусор, гнездо, церковь, метеорит, газ, жиры, сахара>.

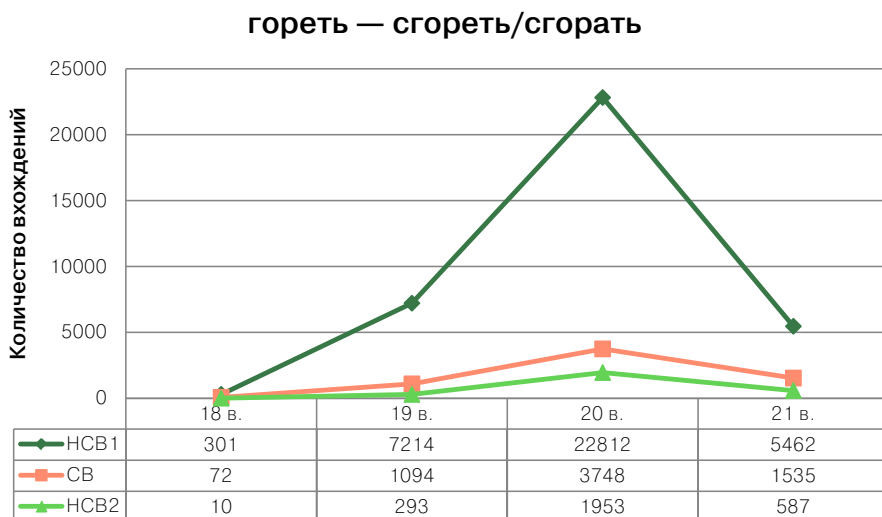


Рис. 3. Абсолютные количества HCB₁, CB и HCB₂ тройки *гореть — сгореть / сгорать* в выдачах НКРЯ (в словах)

Табл. 5. Увеличение / уменьшение частот HCB₁, CB и HCB₂ тройки *гореть — сгореть / сгорать* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
HCB ₁	–58%	—	+13%	–46%
CB	–34%	—	+22%	+71%
HCB ₂	–66%	—	+138%	+43%

4.4. ночевать — переночевать / переночевывать

Запрос: *ночевать*

Сочетаемость: 18 в. — *ночевать* \emptyset , где-л., у кого-л. <в деревне, на станции, тут, здесь, у него>;

19 в. — *ночевать* \emptyset , где-л. у кого-л. <в чулане, на острове, здесь, у Касатского>;

20 в. — *ночевать* \emptyset , где-л., у кого-л. <в келье, у друзей, у сотрудниц>;

21 в. — *ночевать* \emptyset , где-л. у кого-л., у (=возле) чего-л. <дома, там, у телевизора, здесь, в больнице, в избе, на сеновале>.

Запрос: *переночевать*

Сочетаемость: 18 в. — *переночевать* \emptyset , что-л. <ночь>; где-л., у кого-л. <в доме, у меня, у него>;

19 в. — *переночевать* \emptyset ; что-л. <ночь>; где-л., у кого-л. <в участке, у меня>;

20 в. — *переночевать* \emptyset ; что-л. <ночь, несколько ночей>; где-л., у кого-л. <у меня, у соседей, у друзей, в духане, дома, там>;

21 в. — *переночевать* \emptyset ; где-л., у кого-л. <ночь, несколько ночей>, <в интернате, у тебя, там>; как <со всеми удобствами>.

Запрос: *переночевывать*

Сочетаемость: 18 в. — нет;

19 в. *переночевывать* — \emptyset ; где-л. <в комнатах>;

20 в. *переночевывать* — \emptyset ; где-л. <в комнатах>;

21 в. — нет.

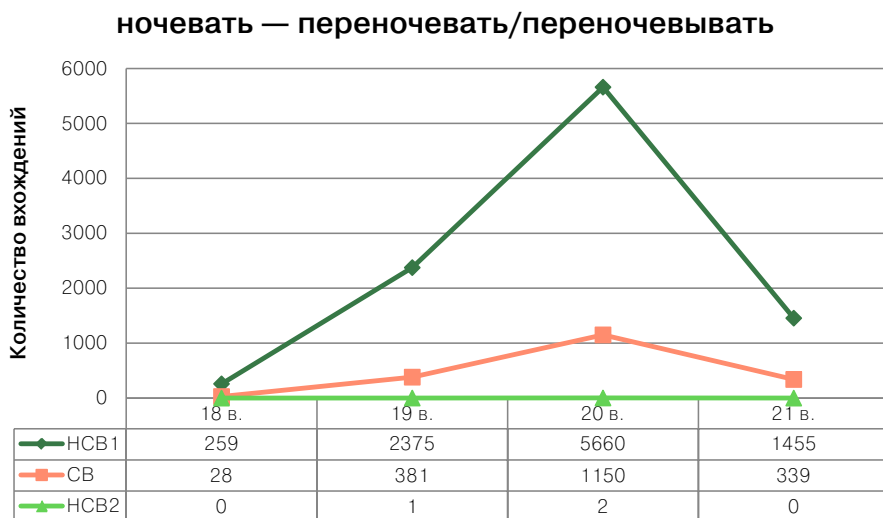


Рис. 4. Абсолютные количества HCB₁, CB и HCB₂ тройки *ночевать* — *переночевать* / *переночевывать* в выдачах НКРЯ (в словах)

Табл. 6. Увеличение/уменьшение частот HCB₁, CB и HCB₂ тройки *ночевать* — *переночевать* / *переночевывать* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
HCB ₁	+9%	—	−15%	−56%
CB	−27%	—	+8%	−36%
HCB ₂	—	—	−28%	—

4.5. *бормотать* — *пробормотать* / *пробормотывать*

Запрос: *бормотать*

Сочетаемость: 18 в. — *бормотать* \emptyset ; с кем-л. <с кем-то>; как <сквозь зубы>;

19 в. — *бормотать* \emptyset ; что-л. <что-то, заговор, слова>; как <невнятно, сквозь зубы, сквозь слезы, про себя, тупо>;

20 в. — *бормотать* \emptyset ; что-л. <что-то>; о чем-л. <о том же>; как <невнятно, себе под нос, дружно>;
 21 в. — *бормотать* \emptyset ; что-л. <что-то>; как <тихо>.

бормотать — пробормотать/пробормотывать

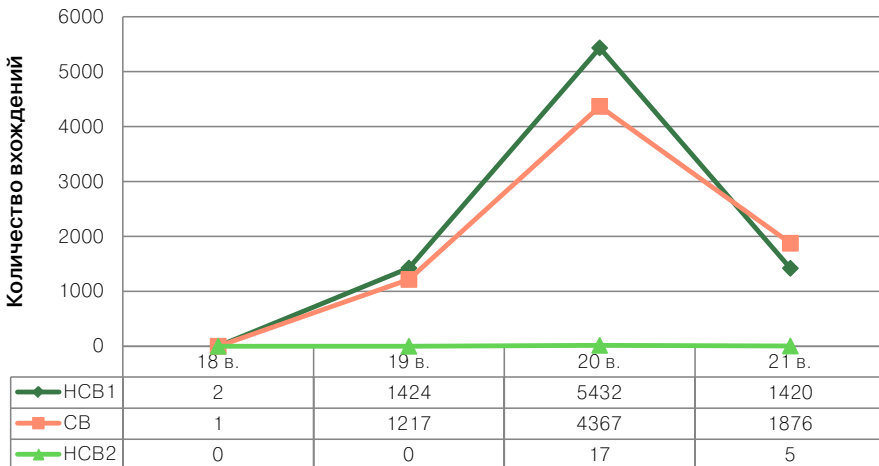


Рис. 5. Абсолютные количества HCB₁, CB и HCB₂ тройки *бормотать* — *пробормотать* / *пробормотывать* в выдачах НКРЯ (в словах)

Запрос: *пробормотать* (1 вхождение)

Сочетаемость: 18 в. — *пробормотать* что-л. <что>;

19 в. — *пробормотать* что-л. <что-то, несколько слов, извинение>; как <сквозь зубы>;

20 в. — *пробормотать* что-л. <что-то>;

21 в. — *пробормотать* что-л. <что-то, молитву, песню, фразу>.

Запрос: *пробормотывать*

Сочетаемость: 18 в. — нет;

19 в. — нет;

20 в. — *пробормотывать* что-л. <что-то, слова, вздохи, куплеты, песенки, текст>;

21 в. — *пробормотывать* что-л. <что-то, слова, песню, слова, суждения>.

Табл. 7. Увеличение/уменьшение частот HCB₁, CB и HCB₂ тройки *бормотать* — *пробормотать* / *пробормотывать* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
HCB ₁	-99%	—	+36%	-29%
CB	-99%	—	+28%	+10%
HCB ₂	—	—	—	—

4.6. вариться — свариться / свариваться

Запрос: *вариться*

Сочетаемость: 18 в. — *варится* что-л. <кушанье, мясо, инбирь> в чем-л. <в желудке, в воде>;

19 в. — *варится* что-л. <груши, рыба, окорок, обед, макароны, сок, мыло, картофель, свекла, мысли, пища, уха> в чем-л. <в воде, в печи, в котле, в чугушке, в желудке, в голове>;

20 в. — *варится* что/кто-л. <еда, уха, глухарь, белье, (политические) блюда, ягоды, гречка, стекло, человек> в чем-л. <в жизни, в баках, в кухне, в Кремле> как <быстро, равномерно, при некоей температуре>; *вариться* в собственном соку;

21 в. — *варится* что/кто-л. <макароны, вермишель, швы, уха, политика, человек> в чем-л. <в кастрюле, в (театральной) жизни, в коммунальной квартире, в семинарах> как <вручную> как долго <много лет>;

Запрос: *свариться*

Сочетаемость: 18 в. — *свариться* что/кто-л. в чем-л. <пища в желудке, человек в котле>;

19 в. — *свариться* что/кто-л. в чем-л. до какого состояния <пища в желудке; рис в воде до мягкости; жженка; каша; человек в котле с кипятком>;

20 в. — *свариться* что/кто-л. в чем-л./у кого-л. до какого состояния <яйца до нужной степени; суп, каша; концентрат в животе; перловая каша в голове; в котле; это у них>; (о ком-л.) *свариться* чем-л. 'обвариться, обжечься' <ребенок кипятком>; *свариться* 'отделиться при варке' что-л. с чего-л. <мясо с костей>;

21 в. — *свариться* что/кто-л. в чем-л. как <картошка в кастрюле; кофе в джезве; человек живьем; пища в желудке; бес в закипевшей воде>.

Запрос: *свариваться*

Сочетаемость: 19 в. — *свариваться* что-л. в чем-л. до какого состояния <рука в кипятке; известка в комья>; что-л. *сваривается* 'слипается' при каких-л. условиях <части при высокой температуре>;

20 в. — *свариваться* что-л. <концы, железо, металлы>; что-л. при каких условиях <пьеса при публике>; кто-л. в чем-л. <хан в котле>; *свариваться* что-л. в чем-л. до какого состояния за какое время <яйца всмятку в восемь минут; гвозди в животе>;

21 в. — *свариваться* что-л. <стальные полосы, каркас, части>; кто-л. как <человек живьем>.

вариться — свариться/свариваться

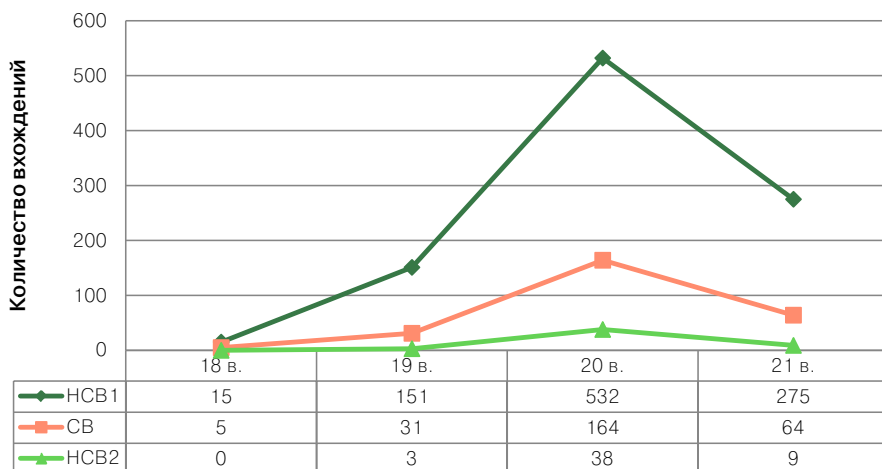


Рис. 6. Абсолютные количества HCB₁, CB и HCB₂ тройки *вариться — свариться / свариваться* в выдачах НКРЯ (в документах / словах)

Табл. 8. Увеличение/уменьшение частот HCB₁, CB и HCB₂ тройки *вариться — свариться / свариваться* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
HCB ₁	+1%	—	+26%	+30%
CB	+61%	—	+89%	+47%
HCB ₂	—	—	+352%	+71%

4.7. бить — пробить / пробивать

Запрос: бить

Сочетаемость: 18 в. — одновалентный: *бить* (о часах, колоколе, курантах, жидкости, сердце); двухвалентный: *бить* кто-л. во что-л. <в барабан, набат, ладоши, ложки>, (о море) <в берега, в землю>, *бить* кто-л. кого-л. ‘убить, забить’ <зверей, животных>; *бить* кто-л. чем-л. <лошадь ногай>; *бить* кто-л. что-л. ‘разбивать’ <стекло, горшки>; *бить* ‘сбивать’ кто-л. что-л. <масло, шерсть>; ‘чеканить’ <монету>; *бить* кто-л. что-л. ‘пробивать’ <шахту>; трехвалентный: *бить* ‘избивать’ кто-л. кого-л. чем-л.; фразеологизмы: *бить* челом, *бить* тревогу, *бить* смертным боем.

19 в. — одновалентный: *бить* (о часах, курантах, колоколе, жидкости); двухвалентный: *бить* ‘пробивать’ кто-л. что-л. <шахту, дорогу>; *бить* что-л. как (об огнестрельном оружии); *бить* кто-л. кому-л. морду; *бить* кто-л. во что-л. <в барабан, набат, калитку>; *бить* ‘чеканить’ <монету>; *бить* ‘забить, убить’ кто-л. кого-л.; *бить* что-л. кого-л. <лихорадка, дрожь, кашель>; трехвалентный:

бить 'избивать' кто-л. кого-л. чем-л.; кто-л. что-л. чем-л. 'сбивать'; кто-л. что-л. чем-л. <конь копытами землю>; *бить* что-л. кого-л. по чему-л./чем-л. <сучья по лицу, вьюга снегом>; *бить* кто-л. кого/что-л. 'убивать' (в том числе — о старшей карте); фразеологизмы: *бить челом*, *бить в глаза*, *бить по нервам* (на нервы); *бить через край*; *бить баклуши*, *бить ключом* (о жизни); *бить на эффект/на жалость*; *бить 'отбивать' такт*; *битый час*;

20 в. — одновалентный: *бить* (о часах, колоколе, курантах, жидкости, сердце); *бить* что-л. (через край); двухвалентный: *бить* кто-л. чем-л. (=хлопать) <крыльями, хвостом>; *бить* кто/что-л. по чему-л. <по мячу, по карману>; *бить* кто-л. что-л. <четку, воблу, рекорды, поклоны>; *бить* что-л. кого-л. <дрожь, озноб, лихорадка>; *бить* кто/что-л. во что-л. <в барабан, в набат, в ладоши>, (о море) <в берега>; *бить* что-л. во что / в кого-л., по кому-л. <ветер, дождь, электричество, гром в землю>, *бить* 'убивать, забивать' кто-л. кого-л. <зверей, животных>; *бить* 'разбивать' кто-л. что-л. <стекло, горшки>; *бить* кто-л. что-л. 'пробивать' <шахту>; трехвалентный: *бить* 'избивать' кто-л. кого-л. чем-л.; фразеологизмы: *бить челом*, *бить тревогу*, *бить смертным боем*;

21 в. — одновалентный: *бить* (о часах, колоколе, курантах, жидкости, сердце); что-л. (через край); двухвалентный: *бить* кто/что-л. по чему-л. <по мячу, по карману, по нервам>; *бить* кто-л. что-л. <четку, воблу, рекорды, поклоны>; *бить* 'хлопать' кто-л. чем-л. <крыльями, хвостом>; *бить* кого-л. чем-л. <током>; *бить* что-л. во что-л./в кого-л., по кому-л. <ветер, дождь, электричество, гром, молния>; *бить* что-л. кого-л. <дрожь, озноб, лихорадка, колотун>; *бить* кто/что-л. во что-л. <в барабан, в набат, в ладоши, в берега (о море), в землю>; *бить* 'убивать, забивать' кто-л. кого-л. <зверей, животных>; *бить* 'разбивать' кто-л. что-л. (стекло, горшки); *бить* кто-л. что-л. 'пробивать' <шахту>; трехвалентный: *бить* 'избивать' кто-л. кого-л. чем-л.; фразеологизмы: *бить челом*, *бить тревогу*, *бить смертным боем*, *бить в голову* (об алкоголе).

Запрос: пробить

Сочетаемость: 18 в. — *пробить* X времени (о часах), *пробить* кто-л. что-л. <сигнал, поход, зорю, дорогу, шахту >, кто-л. что-л. чем-л. <ворота орудиями, шляпу пулей, коня копьем>;

19 в. — *пробить* X времени (о часах, колоколе, склянках), *пробить* кто-л. что-л. <дорогу, тропу>, *пробить* что-л. что-л. <пуля сердце, ядро ворота, трава почве>, *пробить* кто/что-л. что-л. (кому-л.) (пуля Грибоедову ладонь); *пробил* час, звонок;

20 в. — *пробить* X времени (о часах, колоколе, склянках, курантах, барабане), *пробить* кто-л. что-л. <путь, дорогу, тропу, брешь; дробь>, *пробить* 'раздобыть' или 'получить разрешение' кто-л. что-л.; *пробить* кто-л. что-л. в ком/чем-л. (насквозь), *пробить* кому-л. что-л. <бороду (сединай)>; *пробил* час, звонок; *пробить* кого-л. на слезу;

21 в. — *пробить* X времени (о часах, колоколе, склянках, дроби, молнии), *пробить* кто-л. что-л. 'раздобыть' или 'получить' <разрешение, информацию>; кто-л.-л. что в ком/чем-л. (насквозь, через), *пробить* кто-л. что-л. <дорогу, тропу, брешь>, *пробить* кому-л. что-л. <бороду, волосы (сединай)>; *пробить* кого-л. на слезу.

Запрос: пробивать

Сочетаемость: 18 в. — *пробивать* кто-л./что-л. что-л. чем-л. <яйцо носом>;

19 в. — *пробивать* X времени (о часах, колоколе); *пробивать* кто/что-л. что-л. чем-л. <стену, кору, гранит, камень, русскую линию; путь, дорогу, шахту, канавку, брешь, колею>; что-л. пробивает кого-л. <пот>; что-л. пробивает кого/у кого-л. (седина, пушок, слеза); пробивает что-л. что-л. 'прорастает' <трава почву>; что-л. пробивает 'поднимается вверх' (о жидкости), пробивать кого-л. чем-л. 'пронизывать'; пробивать кто/что-л. что/кого-л. (насквозь, через);

20 в. — *пробивать* что-л. <пробил час>; кто/что-л. кому-л. что-л. чем-л. <звание, награду; коньком ногу, себе дорогу>; пробивать что-л. что-л. <свет ладони, луч дым, солнце чащу, пуля кость>; пробивать кто-л. кого/что-л. по чему-л. 'узнавать' <машину по базе>; пробивать 'раздобывать' или 'получать разрешение' кто-л. что-л. кому-л. <ему должность>; пробивать <штрафной (удар)>; пробивать кто-л. что-л. где-л. <чек в кассе, билет >;

21 в. — *пробивать* X времени (о часах, колоколе); пробивать кто-л. что-л. кому-л. чем-л. <стену, кору, гранит, камень, линию; (себе) путь, дорогу, шахту, канавку, брешь, колею>; что-л. пробивает кого-л. <пот, испарина, колотун>; что-л. пробивает кого-л./у кого-л. <седина, пушок, слеза>; что-л. пробивает что-л. 'прорастает'; пробивать 'подниматься вверх' (о жидкости), пробивать 'пронизывать' кого-л. чем-л.; пробивать кого-л. на что-л. <на смех, хохот, ржач, измену, на пожрать>; пробивать кто-л. кого-л. где-л./по чему-л. 'получать информацию'; пробивать кто/что-л. что/кого-л. (насквозь, через) <дорогу, тропу, брешь, заслон, барьер>; пробивать кто-л. что-л. <пенальти, гол, чек>; пробивать кто-л. кому-л. что-л. <звание, награду>.

бить — пробить/пробивать

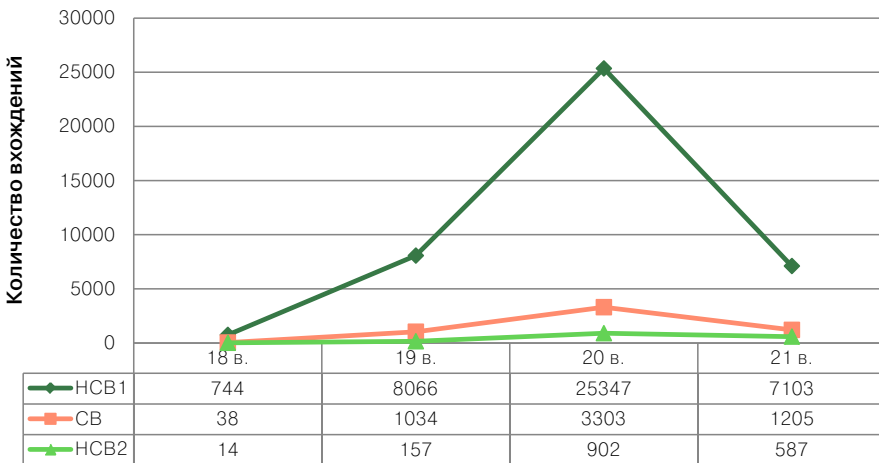


Рис. 7. Абсолютные количества NCB₁, CB и NCB₂ тройки *бить — пробить / пробивать* в выдачах НКРЯ (в словах)

Табл. 9. Увеличение/уменьшение частот НСВ₁, СВ и НСВ₂ тройки *бить* — *пробить* / *пробивать* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
НСВ ₁	–8%	—	+12%	–37%
СВ	–63%	—	+14%	–17%
НСВ ₂	–11%	—	+105%	+167%

4.8. травиться — отравиться / отравляться

Запрос: травиться

Сочетаемость: 19 в. — *травиться* кто-л. чем-л. (о человеке) <мышьяком>; пассив от *травить* кто-л. кем-л. (о животном) ‘затравливать’ <медведя собаками>;

20 в. — *травиться* кто-л. чем-л. (о человеке) <вероналом, спичками, цианистым калием, едой, табаком, газом>; пассив от *травить* кто-л. что-л. <анекдоты>; пассив от *травить* кто-л. кого/что-л. кем/чем-л. <добродетель псами>; пассив от *травить* кто-л. что-л. чем-л./в чем-л. <стекло в кислоте; поля скотом>;

21 в. — *травиться* кто-л. чем-л. (о человеке) <грибами, антибиотиками>.

Запрос: отравиться

Сочетаемость: 18 в. — *отравиться* кто-л. чем-л. (о человеке);

19 в. — *отравиться* кто-л. чем-л. (о человеке) <опиумом, ядом, синильной кислотой, фосфором, политическим принципом>; что-л. чем-л. <удовольствие нехорошим рядом>;

20 в. — *отравиться* кто-л. чем-л. (о человеке) <медным купоросом, грибами, едой, угарным газом, представлениями о счастливой любви>;

21 в. — *отравиться* кто-л. чем-л. (о человеке) <газом, нефтью, грибами, килькой, водкой>.

Запрос: отравляться

Сочетаемость: 19 в. — *отравляться* кто-л. чем-л. (о человеке) <рыбой, морфином, тоскою, литературой, мыслью>; *отравляться* что-л. чем-л. <удовольствие>;

20 в. — *отравляться* кто-л. чем-л. (о человеке) <нитратами, бледной поганкой, водкой, знанием, оперой>; *отравляться* что-л. чем-л. <сознание, воздух, жизнь грязью, радость общения страхом>;

21 в. — *отравляться* кто-л. чем-л. (о человеке) <газом, семенами белены, испарениями, ядами>; *отравляться* что-л. чем-л. <реки, озера, болота; природная среда; русское письмо Прустом; организм ядом, существование вшами>.

травиться — отравиться/отравляться

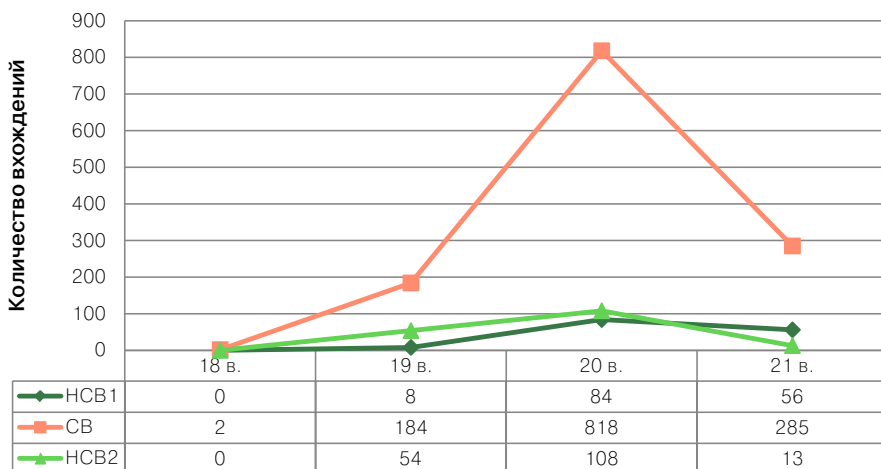


Рис. 8. Абсолютные количества НСВ₁, СВ и НСВ₂ тройки *травиться — отравиться / отравляться* (после очистки от «мусора» выдач НКРЯ; в словах)

Табл. 10. Увеличение/уменьшение частот НСВ₁, СВ и НСВ₂ тройки *травиться — отравиться / отравляться* (после очистки от «мусора» выдач НКРЯ) относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
НСВ ₁	—	—	+275%	400%
СВ	-89%	—	+59%	+11%
НСВ ₂	—	—	-29%	-83%

4.9. жалить — ужалить / ужаливать

Запрос: жалить

Сочетаемость: 18 в. — жалить кто-л. кого-л. <пчелы слона, шмель лису>;

19 в. — жалить кто/что-л. кого/что-л. чем-л. <змея змею; комары Нехлюдова; поступок самолюбие; мысли мозг; пули кого-л.; что-л. укором совести>; жалить до смерти;

20 в. — жалить кто/что-л. кого/что-л. чем-л. <боль; мысли; змеи ребенка; медуза; думы; роса подошвы; пчелы; змея>;

21 в. — жалить кто/что-л. кого/что-л. чем-л. во что-л. <смех героя; копые тело; крапива ноги; клаксон пятерней; жук брамина в щеки; муравьи добычу; солнце голову; в самое уязвимое место>.

Запрос: ужалить

Сочетаемость: 18 в. — *ужалить* кто-л. кого-л. < крыса королевича; змея, пчела человека >;

19 в. — *ужалить* кто/что-л. кого/что-л. чем-л. во что-л. < нечто ее в сердце; змейка руку; змея, тарантул, пчела человека; насекомое человека в лоб >; как *ужаленный*;

20 в. — *ужалить* кто/что-л. кого/что-л. чем-л. < змея человека; страх и сомнение чью-то гордость; оса в горло; пчела, шмель человека >; *будто ужаленный*;

21 в. — *ужалить* кто/что-л. кого/что-л. во что-л. < пуля кого-л. в плечо; солнце; борщевик в лицо; красота богослужения; скорпион ее в родинку; стихи >; как *ужаленный*.

жалить — ужалить/ужаливать

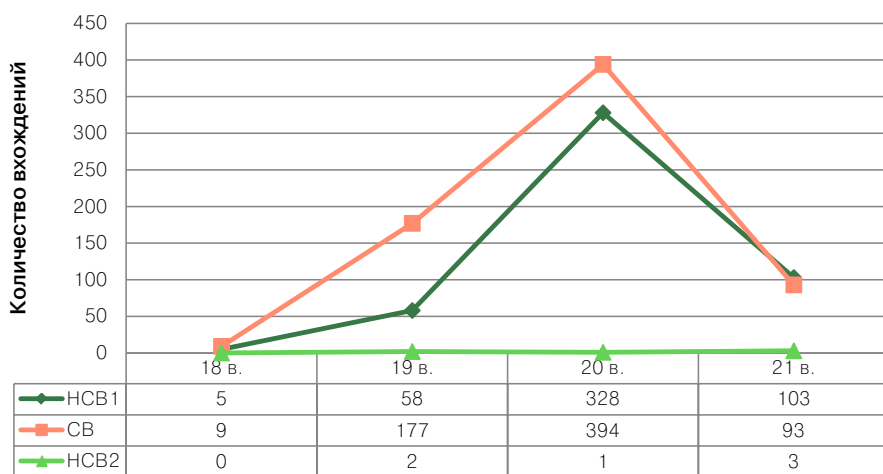


Рис. 9. Абсолютные количества НСВ₁, СВ и НСВ₂ тройки *жалить — ужалить / ужалить* в выдачах НКРЯ (в словах)

Запрос: ужалить

Сочетаемость: 19 в. — *ужаливать* кто-л. кого/что-л. < змеи людей, лошадей >;

20 в. — *ужаливать* кто-л. кого-л. чем-л. < скорпион человека жалом >;

21 в. — *ужаливать* кто-л. кого/что-л. чем-л. < нечто тело горячими заусенцами; пчела человека >.

Табл. 11. Увеличение/уменьшение частот НСВ₁, СВ и НСВ₂ тройки *жалить — ужалить / ужалить* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
НСВ ₁	-14%	—	+102%	+27%
СВ	-49%	—	-21%	-62%
НСВ ₂	—	—	-82%	+7%

4.10. *есть* — *съесть* / *съесть*⁹

Запрос: *есть* (V, tran)

Сочетаемость: 18 в. — *есть* кто-л. что/кого-л. <хлеб, масло, раков>;

19 в. — *есть* кто-л. что/кого-л. <хлеб, хлеб-соль; похлебку, щи, колдуны, баранину, блины, бобы, мясо; волк волка; слизняки малину>; *есть* 'разъесть' что-л. что-л. <запах глаза>; *поедом есть*; *есть за двоих (четверых)*;

20 в. — *есть* кто-л. что/кого-л./ø <хлеб, шоколад, утятину, помидоры, плов, котлеты, персики>; *есть* кого-л. *поедом*;

21 в. — *есть* кто-л. что/кого-л./ø чем-л. <кита, глухаря; фондю вилками, картошку, желуди, яблоки, орехи, ягоды, гранаты; торт, икру>; *есть* 'разъесть' что-л. что-л. <махорка глаза>.

Запрос: *съесть* (V, tran)

Сочетаемость: 18 в. — *съесть* кто-л. кого/что-л. <собаки гуся, осел зерно>;

19 в. — *съесть* кто-л. что/кого-л. <сливы; хлеб; кого-л. живьем; гриб, корку>; кто-л. на чем-л. *собаку съел*; кто-л. кого-л. *поедом съел*;

20 в. — *съесть* кто-л. кого/что-л. <собака курицу; он тебя; кто-л. яблоко, мороженое; паек; медикаменты, сосиски; тарелку борща, супу>; что-л. что-л. <опухоль часть легкого; шахматная фигура шахматную фигуру; огонь бор; гром окончание фразы>; кто-л. на этом деле *зубы съел*; *бог не выдаст, свинья не съест*;

21 в. — *съесть* кто-л. кого/что-л. <медведь птенцов, мышку; слона; кусок селедки; по два куска торта; бутерброд, мюсли, хлеб, банан, йогурт, мороженое; львы своих детенышей>; что-л. что-л. <эти щиты пять миллионов долларов>; *кто смел, тот и съел*; кто-л. на чем-л. *собаку съел*;

⁹ В выдачах всех диахронических подкорпусов с НСВ₁ *есть* 'принимать пищу' в значительном количестве присутствует *есть* бытийный (презентная форма глагола *быть*), в том числе в рамках фрагмента выдач из подкорпуса со снятой омонимией. По причине очень большого объема выдач их ручная очистка от этого шума не производится, в силу чего количественные данные по этому симплексу недостоверны.

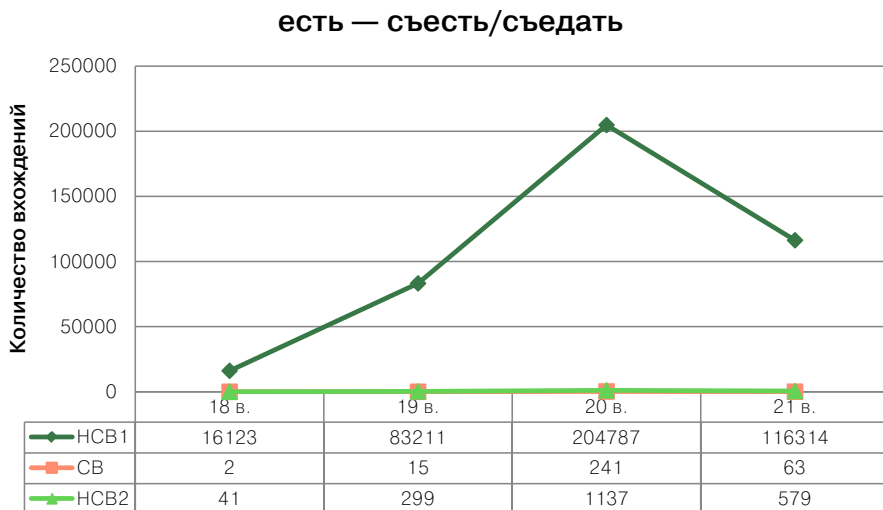


Рис. 10. Абсолютные количества HCB₁, CB и HCB₂ тройки *есть — съестъ / съедать* в выдачах НКРЯ (в словах)

Табл. 12. Увеличение/уменьшение частот HCB₁, CB и HCB₂ тройки *есть — съестъ / съедать* в выдачах НКРЯ относительно 19 в.

	18 в.	19 в.	20 в.	21 в.
HCB1	+93 %	—	−12 %	−0,2 %
CB	+33 %	—	+474 %	+200 %
HCB2	+37 %	—	+37 %	+38 %

Запрос: *съедать* (V, tran)

Сочетаемость: 18 в. — *съедать* кто/что-л. что/кого-л. чем-л. <себя жалостью, заботы батюшку; тоска душу; каши; жена его сердце; время красоту; мот, скупой деньги; заботы, несчастья что-л.; цементная вода железо; черви корни; моль вещи>;

19 в. — *съедать* кто/что-л. что/кого-л. <человек человека; бояре нас; гад другую гадину; ученики кашу; коновал жаркое; дядя по пятисот устриц ежедневно; юноша ватрушку; ситного с луком; барашка; по овце; по шести блюд; 12 котлет; обед, завтрак, пай; два бублика; алчба Вахрушку; тоска, болезнь человека; некто гарантии и субсидии; дворня половину наших доходов; крысы чье-то тело; орлы мелких птиц; в день одну картофелину>; что-л. что-л. <дым сухие иглы; чиновничество бюджет; клетки другие клетки>;

20 в. — *съедать* кто/что-л. что/кого-л. <человек 400 граммов мяса; бутерброд, по яблоку, паек, кашу, довесок, бульон из супа, банку икры за 3 дня; папа съедает сковородку; зависть к другим нас же; каждый человек пуд хлеба; все до крошки; щука уток>; что-л. что-л. <дача много дров и угля; туман последние снега; сумерки краски; нечто душу>; печаль-тоска кого-л. съедает;

21 в. — *съесть* кто/что-л. что/кого-л. чем-л. <завтрак; косу из теста, бутерброд; люди его взглядами; варенье из таза; 400 граммов свежих овощей в день; фрукт на завтрак; немного яда; аппаратчики денежный куш; кашу, банан; по батону каждый день; летучие мыши X насекомых; Минотавр юношей и девушек; невестка печень мужа по кусочку; рыбка червячка; волки больных волков; героиня что-то; рак маму; бизнес, алкоголь, безработица, нервное напряжение мужчин; червь честолюбия, зависть кого-л.>; что-л. что-л. <тропическое дерево загрязнение; проверка тетрадей, служба время; хлопоты день за днем; инфляция деньги; боль желудок; автомобиль топливо; шахматная фигура шахматную фигуру, ладью слон; луна диск солнца>.

4.11. Некоторые наблюдения

4.11.1. Семантика и сочетаемость

Сравнивая сочетаемость СВ и НСВ₂ с одной стороны, и НСВ₁ и СВ, с другой (см. [разделы 4.1–4.10](#)), легко убедиться в том, что синтаксические модели и лексическое наполнение аргументов в первой паре глагольных форм (с одинаковым префиксом) в точности совпадают, а во второй паре (симплекс и префигированный глагол) — существенным образом различаются, причем это верно как для синхронного анализа на каждом временном срезе, так и в диахронии. Показательны две тройки с общим НСВ₁ *рвать*. *Отношения*, например, можно только *рвать*, но не *сорвать* (*срывать*) или *оторвать* (*отрывать*), а уроки — только *сорвать* (*срывать*), но не *рвать* или *оторвать* (*отрывать*). Встречаются, впрочем, и единичные случаи с вариативностью префиксов: *пуговицы* можно *рвать*, *оторвать* (*отрывать*) и — существенно более периферийный выбор — *сорвать* (*срывать*). Эта закономерность легко объясняется наличием (одного и того же) префикса в формах СВ и НСВ₂ и его отсутствием у симплекса НСВ₁.

Отметим, что тройка *ночевать* — *переночевать* / *переночевывать* с единой моделью управления — (*пере*)*ночев(ыв)ать* кто, где / у кого — обладает примечательной синтаксической спецификой. Семантика префикса в этом случае взаимодействует с инкорпорированным аргументом, выраженным корнем¹⁰. По этой причине при глаголе не открывается семантическая валентность для аргумента с синтаксической ролью прямого дополнения, который был бы общим для СВ и НСВ₂, как это имело место в случаях *оторвать* / *отрывать* (что / кого от кого / чего) и *сорвать* / *срывать* (что / кого с кого / чего).

4.11.2. Статистика: общие наблюдения

- i. количество НСВ₁ в абсолютных цифрах во всех без исключения случаях (на десяти тройках) превышает количество НСВ₂;

¹⁰ Изредка встречаются случаи его поверхностной реализации, ср. *Ну, переночевал ночь; наутре чуть зорька занялась, я вскочил...* [В. А. Слепцов. Ночлег (Подгородные сцены) (1863)]; *В 43-м году, после Северо-Западного фронта, после артиллерийского училища, ехал я на Юго-Западный фронт через Москву и одну ночь ночевал у Юлии Федоровны.* [Г. Я. Бакланов. Жизнь, подаренная дважды (1999)].

- ii. превышение количества НСВ₁ над НСВ₂ по результатам оценки увеличения / уменьшения частотности в созданных подкорпусах относительно подкорпуса 19 в. (в процентах) зафиксировано в 4-х тройках: *ночевать* — *переночевывать*, *бормотать* — *проборматывать*, *травиться* — *отравляться* и *жалить* — *ужаливать*¹¹; для остальных шести (*рвать* — *отрывать*, *рвать* — *срывать*, *вариться* — *свариваться*, *гореть* — *сгорать*, *есть* — *съесть*, *бить* — *пробивать*) видим другое: 18 в. — НСВ₁ > НСВ₂ (или НСВ₁ отсутствует, как в *вариться* — *свариваться*), но в 20 и 21 вв. — соотношение обратно: НСВ₂ преобладает над НСВ₁; это может интерпретироваться как рост употребительности НСВ₂ в 20 и 21 вв.;
- iii. количество НСВ₁ чуть более чем в половине случаев (6 на 4) превышает количество префигированного СВ (в абсолютных цифрах); так в случаях *вариться* — *свариться*, *гореть* — *сгореть*, *ночевать* — *переночевывать*, *есть* — *съесть*, *бормотать* — *пробормотать*, *бить* — *пробить*, не так в парах: *рвать* — *отрывать*; *рвать* — *сорвать*; *жалить* — *ужалить*, *травиться* — *отравиться*, где СВ по числу вхождений превышает НСВ₁; есть специфика в случае *бормотать* — *пробормотать*: СВ превосходит по количеству НСВ₁ только в 21 в.

5. Выводы и обобщения

Полученные данные позволяют, как представляется, сделать следующие выводы.

1. Не знающее исключений (на 10 рассмотренных тройках) преобладание НСВ₁ над НСВ₂ (наблюдение i) подтверждает, что имперфектив-симплекс как вокабула, обладая более широкой полисемией, может в качестве той или иной лексемы выполнять роль «джокера» для однокоренного НСВ₂ с конкретизированной префиксом семантикой. Эта роль оказывается доступна при (квази)синонимии беспрефиксного НСВ₁ и префигированного НСВ₂, результата суффиксальной имперфективации, примененной к префигированному же СВ, который в свою очередь оказывается (квази)синонимичным перфективом (обозначением события) для лексемы НСВ₁. Последнее возможно при обозначении лексемой НСВ₁ предельного процесса и совпадении набора ее аргументов с набором аргументов префигированных СВ и НСВ₂, см. (1).

¹¹ Интересно, что все эти тройки «неблагополучны» в отношении НСВ₂: *ужаливать* фиксируется только с 19 в. (нет в 18 в.), вхождения единичны; *проборматывать* — с 20 в., количество вхождений — менее двух десятков; *отравляться* (и *травиться*) не фиксируется в 18 в.; *переночевывать* зафиксировано только в 19 в. (1 вхождение) и в 20 в. (2), но не в 18 и не в 21 вв. (синонимичный *заночевывать* — 12 вхождений по основному корпусу, без разбивки на временные периоды). В остальных тройках картина иная: НСВ₂ фиксируется с 18 в. (кроме *свариваться*) и его количество нарастает, неизменно показывая в каждом последующем временном периоде лучший по сравнению с НСВ₁ результат.

- (1) а. *Музыкант смутился, съезжился и, бросившись одеваться в свой самый лучший кафтан и камзол, торопился, рвал^{НСВ1} пуговицы и парик надел набок.* [Е. А. Салиас. Ширь и мах (1885)]
 б. *Постой, ты мне пуговицу оторвал^{СВ}.* Говори хладнокровно. [А. П. Чехов. Трагик поневоле (1889)]
 с. *<...> говорил он, судорожно отрывая^{НСВ2} одной рукой пуговицу у сюртука, а другой царапая чуть не до крови ухо, — совершенно спокоен!* [И. А. Гончаров. Счастливая ошибка (1839)]

Более общий вывод сводится к тому, что вокабула симплекса НСВ₁ выступает в качестве гиперонима для всех (не только в рамках данной тройки) своих (единожды) префигированных дериватов СВ и регулярно образуемых от последних НСВ₂. Гипероним и гипоним(ы) связаны отношениями деривации и в общем случае (кроме описанной выше (квази) синонимии) различаются лексической семантикой (за счет отсутствия / наличия префикса), при этом степень этого различия может варьировать (ср. *жить* — *прожить* / *проживать* и *знать* — *узнать* / *узнавать*). Словообразовательные отношения между НСВ₁ и СВ ~ НСВ₂ при заведомо большей многозначности (вокабулы-)симплекса, в роли гиперонима и его синтаксической лабильности (на фоне образованных от него префигированных дериватов), предсказывает превышение частотности НСВ₁ (как вокабулы) над частотностью НСВ₂ в любой временной период. И это предсказание подтвердилось. Однако преобладание по частотности НСВ₁ (как лексемы) над (квази)синонимичной ей лексемой НСВ₂ на том или ином временном срезе предложенной теоретической модели не предсказывается. Сравнение частотности двух имперфективов (как в синхронии, так и в диахронии) будет корректным только в рамках доек лексем СВ — НСВ₁ и СВ — НСВ₂, что по техническим причинам (см. сноску 5 выше) в рамках корпусного анализа пока трудно реализуемо.

2. Рабочая гипотеза об абсолютном преобладании по частотности НСВ₁ над СВ не подтвердилась (наблюдение iii). Однако предложенная теоретическая модель русского вида способна предложить объяснение этому факту с помощью одного из своих компонентов — акциональности. Предположительно, в тех случаях (4 из 10), в которых СВ оказался более частотным, чем НСВ₁, референт обозначаемой ситуации чаще реализуется (концептуализируется) как (достигшее кульминации) событие, чем как соотносенный с пределом, но не достигающий кульминации, процесс. То есть говорящий обозначает однократную ситуацию с кульминацией чаще, чем что-либо иное: в нарративе чаще реализуется *ужалить*, чем *жалить*, и *отравиться*, чем *травиться*. При этом симплекс получает явное преимущество перед СВ в контекстах нейтрализации [Маслов 2004: 109], недоступных для перфектива (напр., в презенсе), реализуя в этом случае обобщенно-фактическое значение при единичности ситуации (2a) или неограниченно-кратное значение при итеративности (2b); контексты нейтрализации — это и своеобразный «инкубатор» для НСВ₂, см. (2c).

- (2) а. Змея не **жалит** змея, а сановники только и думают, как бы **ужалить** друг друга. [В. М. Дорошевич. Сказки и легенды (1893–1916)]
 б. Острый же язычок Марии Николаевны проявлялся порою и **жалил** кого полагается и в последующие годы. [В. Д. Алейников. Тадзимас (2002)]
 с. — Он их не боится! Его не **ужаливают!** [А. Эппель. Чреватая идея // «Знамя», 2002]
3. Рабочая гипотеза о наличии в диахронии направленного тренда для НСВ₂ (роста или снижения употребительности, а не движения по синусоиде) подтвердилась частично (см. наблюдение ii): рост употребительности НСВ₂ обнаружен в 6 случаях из 10 (при явной специфичности тех 4-х троек, которые показали другой результат, см. сноску 10). Таким образом, можно уточнить вывод, сделанный ранее в рамках диахронического исследования на базе НКРЯ относительно НСВ₂, но не на биимперфективных тройках: «Явного тренда по увеличению или снижению образования НСВ₂ не прослеживается» (Горбова 2015: 25). Новый материал дает основания с осторожным оптимизмом говорить о тренде увеличения частотности НСВ₂ по сравнению с НСВ₁ в 20 и 21 вв., т. е. об усилении словоизменительного компонента русского вида, который обеспечивается механизмом суффиксальной имперфективации приставочных глаголов.

References

1. *Apresjan Ju. D.* (1995), Interpretation of redundant aspectual paradigms in the defining dictionary [Traktovka izbytochnykh aspektual'nykh paradigm v tolkovom slovare]. Selected papers. Vol. 2: An integral language description and systemic lexicography [Izbrannye trudy. T. II: Integral'noe opisanie jazyka i sistemnaja leksikografija]. Jazyki Russkoi Kul'tury, Moscow, pp. 103–114.
2. *Apresjan Ju. D.* (1997), Lexicographical interpretation of the aspect: non-trivial cases [Leksikograficheskaja traktovka vida: netrivial'nye sluchai]. Proceedings of the aspectological seminar of the Philological faculty of Lomonosov Moscow State University [Trudy aspektologicheskogo seminarana filologicheskogo fakul'teta MGU im. M. V. Lomonosova. T. 2], MAKS Press, Moscow, pp. 7–20.
3. *Brezina V.* (2018), Statistics in corpus linguistics: A practical guide. Cambridge University Press, Cambridge; available at: <https://www.cambridge.org/core/books/statistics-in-corpus-linguistics/4E530F86B328B2287681AD240796D2CF>.
4. *Gorbova E. V.* (2014), Aspect formation of the Russian verb revisited: inflectional interpretation of the aspect [Yeshche raz o vidoobrazovanii russkogo glagola: k slovoizmenitel'noj traktovke vida]. Russian Linguistics, vol. 38 (2), pp. 1–21.
5. *Gorbova E. V.* (2015), Aspectual formation of Russian verb: prefixation and / or suffixation? [Vidoobrazovanije russkogo glagola: prefiksacija ili suffiksacija?], Topics in the study of language [Voprosy jazykoznanija], vol. 1, pp. 7–38.
6. *Gorbova E. V.* (2019a), A two-component model of Russian aspect: statements, predictions, confirmations [Dvukhkomponentnaja model' russkogo vida: polozhenija, predskazanija, podtverzhdenija], Slavistika, vol. 23 (1), pp. 45–61.

7. Gorbova E. V. (2019b), Regularity of the Secondary Imperfectivization of Russian Prefixal Perfectives (the case of *pro-*, *u-* and *s(o)-*verbs) [Reguliarnost' vtorichnoi imperfektivacii russkikh prefigirovannykh perfektivov (na primere glagolov s prefigsami *pro-*, *u-*, *s(o)-*)], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2019" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2019"], Moscow, Issue 18, Supplementary vol., pp. 31–44; available at: <http://www.dialog-21.ru/digest/2019/online/>.
8. Janda L. A. (2012), Russian prefixes as a verb classifier system [Russkie pristavki kak sistema glagol'nykh klassifikatorov], Topics in the study of language [Voprosy jazykoznanija], vol. 6, pp. 3–47.
9. Janda L. A., Endresen A., Kuznetsova Ju., Lyashevskaya O., Makarova A., Nessel T., Sokolova S. (2013), Why Russian aspectual prefixes aren't empty. Prefixes as verb classifiers. Slavica Publishers Indiana University, Bloomington.
10. Janda L. A., Lyashevskaya O. (2011), Aspectual Pairs in the Russian National Corpus. Scando-Slavica, vol. 57 (2), pp. 201–215.
11. Jászay L. (2001), Specifics of secondary imperfectives of aspectual correlations [O specifike vtorichnykh imperfektivov vidovykh korreljatsij], Nedjalkov I. V. (ed.). Linguistics research. Collection of papers on the 70th anniversary of A. V. Bondarko [Issledovanija po jazykoznaniju. Sb. statei k 70-letiju A. V. Bondarko], Saint-Petersburg State Univ., St. Petersburg, pp. 106–118.
12. Karcevski S. (1927/1962), IV. 'Aspect' [IV. 'Vid'], Maslov Ju. (ed). Problems of Verbal Aspect [Voprosy glagol'nogo vida], Foreign Publishing House [Izd-vo Inostrannoj Literatury], Moscow, pp. 218–230. (Karcevski S. (1927), System of the Russian verb. An essay on synchronic Linguistics [Système du verbe russe. Essai de linguistique synchronique], Legiografie, Prague).
13. Maslov Ju. (2004), Selected works: Aspectology. General Linguistics [Izbrannyje Trudy: Aspektologija. Obshchee jazykoznanie], Jazyki Slavianskoi Kul'tury, Moscow.
14. Petrukhina E. V. (2000), Aspectual categories of the verb in Russian versus Czech, Slovak, Polish and Bulgarian [Aspektual'nye kategorii glagola v russkom jazyke v sopostavlenii s cheshskim, slovatskim, pol'skim i bolgarskim], Moscow State Univ., Moscow.
15. Petrukhina E. V. (2009), The Russian verb: categories of aspect and tense (in the context of recent linguistic research) [Russkij glagol: kategorii vida i vremeni (v kontekste sovremennykh lingvisticheskikh issledovanij)], MAKS Press, Moscow.
16. Russian National Corpus [Nacional'nyi korpus russkogo jazyka] (2003–2020), available at: <http://www.ruscorpora.ru>.
17. Tatevosov S. G. (2010), Primary and secondary in the structure of imperfectives [Pervichnoe i vtorichnoe v strukture imperfektivov], Acta Linguistica Petropolitana. Transactions of the Institute for Linguistic Studies [Trudy ILI RAN], Redundancy in the Grammar of Language [Izbytochnost' v grammaticheskom stroe jazyka], vol. VI (2). Kazanskij N. N. (ed). Nauka, St. Petersburg, pp. 299–321.

18. *Tommola H.* (2008), Aspectually significant manners of action. Towards the rehabilitation of the pair purity [Aspektual'no-znachimyje sposoby dejstvija. K rehabilitatsii chistoty pary], *Humaniora: Russian Language [Humaniora: Lingua Russica]*, Works on Russian and Slavic philology [Trudy po russkoj i slavjanskoj filologii. Lingvistika], XI. Language in a functional and pragmatic aspect [Jazyk v funkcional'no-pragmaticheskom aspekte], Tarty Univ., Tartu, pp. 218–232.
19. *Wiemer B.* (2019), About semantically invariant and grammatically trivial in Russian aspect [O semanticheski invariantnom i grammaticheski trivial'nom v russkom vide], Gerasimov D. V., Dmitrenko S. Ju., Zaika N. M. (eds.), Collection of papers on the 85th anniversary of V. S. Xrakovskij [Sbornik statej k 85-letiju V. S. Xrakovskogo], *Jazyki Slavianskoi Kul'tury*, Moscow, pp. 43–66.
20. *Xrakovskij V. S.* (2005), Aspectual triplets and aspectual pairs [Aspektual'nye trojki i vidovje pary], *Russian Language and Linguistic Theory [Russkij jazyk v nauchnom osveshchenii]*, vol. 9, pp. 46–59.
21. *Xrakovskij V. S.* (2006), Aspectological notes [Aspektologičeskije zametki], *Philology, Russian Language. Education [Filologija. Russkii jazyk. Obrazovanie]*, Collection of papers dedicated to the anniversary of professor L. A. Verbitskaja [Sbornik statej, posvyashchennyi jubileju prof. L. A. Verbitskoi], Saint-Petersburg State Univ., St. Petersburg, pp. 202–213.
22. *Zaliznjak Anna A., Mikaeljan I. L.* (2010), About the place of aspectual triplets in the aspectual system of the Russian language [O meste vidovykh troek v aspektual'noj sisteme russkogo jazyka], Proceedings of the International Conference “Dialog 2010” [Trudy Mezhdunarodnoy Konferentsii “Dialog 2010”], available at: <http://www.dialog-21.ru/media/1646/21.pdf>.
23. *Zaliznjak Anna A., Mikaeljan I. L., Shmelev A. D.* (2015), Russian aspectology: in defense of the aspectual pair [Russkaja aspektologija: v zashchitu vidovoj pary], *Jazyki Slavianskoi Kul'tury*, Moscow.
24. *Zaliznjak Anna A., Shmelev A. D.* (2000), An introduction to Russian aspectology [Vvedenije v russkiju aspektologiju], *Jazyki Russkoi Kul'tury*, Moscow.

RELATION EXTRACTION DATASET FOR THE RUSSIAN

Gordeev D. I. (gordeev-di@ranepa.ru),
Davletov A. A. (davletov-aa@ranepa.ru),
Rey A. I. (rey-ai@ranepa.ru),
Akzhigitova G. R. (akzhigitova-gr@ranepa.ru),
Geymbukh G. A. (geymbukh-ga@ranepa.ru)

RANEPA, Moscow, Russia

There are few existing relation extraction datasets for the Russian language and they contain a rather small number of examples. Thus, we decided to create a new Ontonotes-based named entities and relation extraction sentence-level dataset called RURED. The dataset contains more than 500 annotated texts and more than 5,000 labelled relations. We also publish baseline models for relation extraction and named entity recognition trained on the dataset. Our models achieve 0.85 for named entity recognition and 0.78 for relation extraction in F1-score.¹

Key words: relation extraction, named entity recognition, taced, bert

DOI: 10.28995/2075-7182-2020-19-348-360

ДАТАСЕТ ДЛЯ ИЗВЛЕЧЕНИЯ ОТНОШЕНИЙ ДЛЯ РУССКОГО ЯЗЫКА

Гордеев Д. И. (gordeev-di@ranepa.ru),
Давлетов А. А. (davletov-aa@ranepa.ru),
Рей А. И. (rey-ai@ranepa.ru),
Акжигитова Г. Р. (akzhigitova-gr@ranepa.ru),
Геймбукх Г. А. (geymbukh-ga@ranepa.ru)

РАНХиГС, Москва, Россия

На данный момент существует немного размеченных наборов данных для извлечения отношений из текстов. В данной статье мы представляем такой датасет RURED, содержащий разметку именованных сущностей по схеме Ontonotes и отношений между ними на уровне предложений. Датасет содержит более 500 аннотированных текстов и более 5000 размеченных отношений. Также мы публикуем основанные на BERT модели, обученные на этом наборе данных. В задаче автоматического распознавания именованных сущностей модель достигла 0,85 п. п. по метрике F1, для задачи извлечения отношений — 0,78.

Ключевые слова: извлечение отношений, распознавание именованных сущностей, taced, bert

¹ <https://github.com/InstituteForIndustrialEconomics/rured>

1. Introduction

The task of relation extraction is to find entities in a sentence and establish the type of relations between them, i.e. to extract triplets from texts: (entity 1; entity 2; their relationship). For example, in the sentence “Mark Zuckerberg, the founder of Facebook, bought a startup.” there are named entities: “Mark Zuckerberg” and “Facebook”, which are connected by the relation “Founder”. Relation extraction is useful for building taxonomies and extracting facts from texts.

There are several approaches to the problem:

- supervised learning-based methods
- distant supervision-based methods

A popular approach to relation extraction is distant supervision [13]. This method uses an ontology database and a large text corpus to align sentences containing entities. Unfortunately, this method is prone to noisy labels [12]. There have been numerous attempts at fixing problems of distant supervision. However, still, most models ignore categories in the long tail of the distribution [6]. Moreover, classes tend to be of some distinct domains that are typical of the database (e.g. these are locations and nationalities in the case of NYT10 [16] built on top of Freebase [1]). If we are interested in relations that are absent from the database, we have no choice but to resort to supervised or semi-supervised methods.

There is a decent number of named entity datasets for the Russian language. There are traditional Person-Organization-Location (e.g. [5], [7], [14]) datasets as well as more specialized ones that are devoted to a single type of entities (e.g. Persons-1000²). Moreover, it is possible to use transfer learning and zero-shot learning for named entity recognition. It was shown that a multilingual BERT-based model which is fine-tuned on an English NER (named entity recognition) dataset is able to gain reasonable results for Russian³ (unrelated to this work, but we tried a similar transfer learning approach with TACRED and it did not bring us any results. We will explore it in future work). However, for relation extraction the situation is more challenging.

Supervised approaches often treat the problem of relation extraction as a classification task. This approach allows for achieving higher quality predictions. Moreover, the classes may be beyond the scope of knowledge bases. However, the annotation procedure is demanding and tiresome. That is why such datasets are few and exist only for major languages. For the Russian language the only existing annotated dataset was published for the competition FactRuEval 2016 [17] held in conjunction with the conference Dialogue 2016. It contains 1059 facts (which may contain multiple relations). It is often not enough for training a classifier. Even the organizers of FactRuEval 2016 state that “the small size of the demo corpus shut out systems that relied on machine learning and made it difficult to fine-tune rule-based systems”. The best performing system achieved the F1 score of 0.51. At the same time as our work there was created

² <http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000>

³ <http://docs.deeppavlov.ai/en/master/features/models/ner.html#multilingual-bert-zero-shot-transfer>

RuREBus dataset⁴ which was used for RuREBus shared task held together with Dialog 2020 conference. It contains about 300 annotated texts from the domain of Russian municipal legal documents. It can be used as a benchmark for relation extraction and named entity recognition algorithms and it is close in its nature to business domains. However, due to the specificity of the domain it cannot be used for general purposes. Another interesting article focused on extracting relations from Wikipedia pages [10]. It contained both an automatically extracted dataset and manually labelled data. The annotators linked only relations between the ‘main’ company (the topic of the Wikipedia page) and already highlighted mentions of other companies (which have a corresponding Wikipedia page or it is to be created). This procedure works only for Wikipedia and similar domains because we do not usually know the topic of the text and there may be relations between non-topical entities. Thus, we decided to create RURED (RUSSIAN Relation Extraction Dataset). It contains 536 annotated texts and 5,381 relations. The number of labelled named entities is 22,595. Using training data from the dataset we trained several models for named entity recognition and relation extraction.

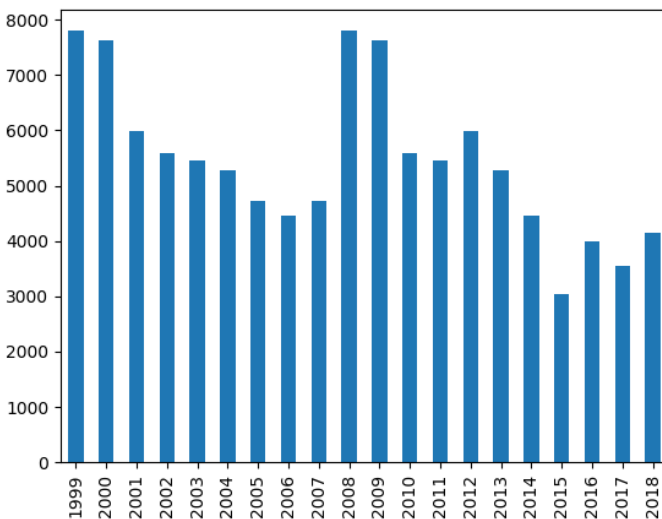


Figure 1: Yearly distribution of economy articles in Lenta.ru

⁴ <https://github.com/dialogue-evaluation/RuREBus/>

2. Dataset

We used Lenta.ru news corpus for annotation.⁵ Only texts with tag “Экономика” (economy) were selected because we were mostly interested in economic events for our future work. Lenta.ru news dataset contains news articles from 1999 till 2019. All texts were selected randomly for annotation.

2.1. Named entities labelling

Named entities were automatically annotated using BERT [3] model trained on English Ontonotes [4] provided by DeepPavlov [2]. The cross-lingual nature of BERT allows us to successfully infer named entities for the Russian language despite the model being fine-tuned on the English dataset. During annotation, named entities were manually corrected if wrong and new entities were added when necessary.

We adhered to Ontonotes 5.0 guidelines and stuck to its annotation procedure.⁶ Several new types and subtypes of named entities were also added (see **Table 1**). For example, GPE was split into several separate types: COUNTRY, CITY, REGION, BOROUGH.

Table 1: New named entity types besides those in Ontonotes

Named entity type	Subtype-of	Description
PROFESSION	—	Professions and people of these professions. Corresponds to ‘title’ in TACRED
COUNTRY	GPE	Names of countries
REGION	GPE	Names of sub-country entities
CITY	GPE	Names of cities, towns and villages
BOROUGH	GPE	Names of sub-city entities
GROUP	—	unnamed groups of people and companies
FAMILY	GROUP	families as a whole
AGE	NUMBER (not used in annotation)	people’s and objects’ ages
NATIONALITY	NORP	names of nationalities
RELIGION	NORP	names of religions
CURRENCY	—	names of currencies

Nested named entities were not labelled—only upper-level entities were annotated. For example, the whole phrase ‘посол РФ в Камеруне’ (ambassador of Russia to Cameroon) was labelled as ‘profession’, child entities ‘РФ’ (Russian Federation) and ‘Камеруне’ (Cameroon) were not labelled. It might be changed in future releases.

Prepositions were included in named entities (usually dates and numbers) if they were vital for the entity (e.g. “since 1992” was labelled as a single entity). However, it resulted in some confusion among annotators due to verb and nouns valency.

⁵ <https://github.com/yutkin/Lenta.Ru-News-Dataset>

⁶ <https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf>

State and organization departments were labelled as “ORGANIZATIONS” and later were connected with “OWNERSHIP” relation (Fig. 4).

Unfortunately, as with many natural datasets we see a heavy-tailed class distribution here (Fig. 2).

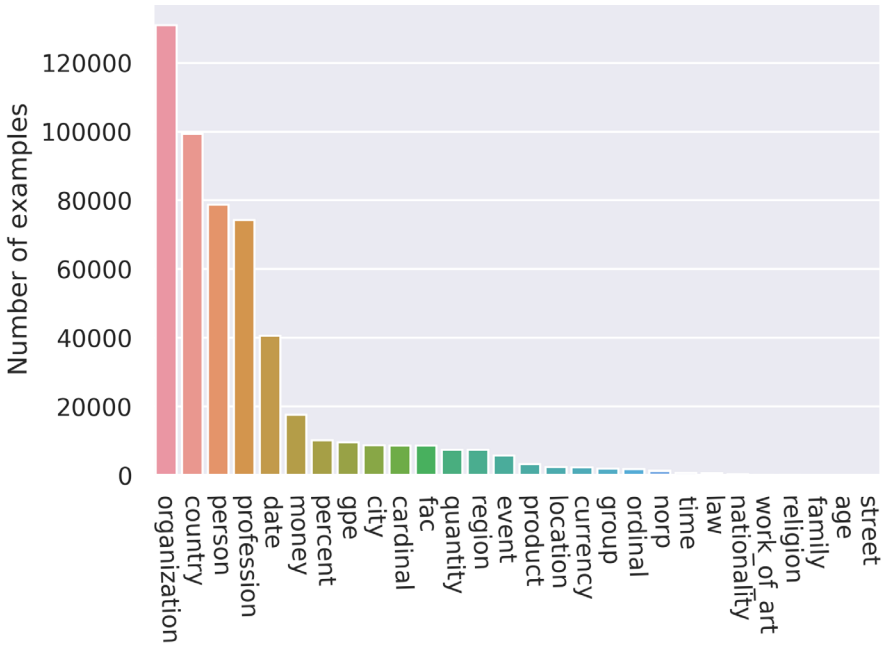


Figure 2: Distribution of named entity classes

2.2. Relation extraction labelling

Our relation extraction dataset is based on TACRED (The TAC Relation Extraction Dataset) [20]. TACRED is an English-language data corpus with labelled entities and relations between them. The markup was done manually using the MTurk HIT distributed annotation system based on data from TAC KBP 2009–2015 competitions. In the competition, each participant is given 100 entities and a large body of texts containing them. Participant systems must extract heterogeneous attributes for each entity and use them to fill a knowledge base. The knowledge base consists of named entities and their attributes. TACRED contains 42 types of relations. 79.9% of examples are of the class ‘no relation’. There are 106,264 examples in the dataset. One of the downsides of TACRED is that approximately 90% of sentences contain only a single annotated relation despite there being multiple possible relations. For unknown texts, this approach also requires to check all named entities against all other named entities ($O(N^2)$ complexity) because relations are intransitive.

To overcome this problem we labelled all relations in a sentence. It allows training models that predict all possible relations for the word in a single run ($O(N)$).

Brat annotation tool [18] has been used for relation and named entities annotation.

All relations were labelled if the fact has been true at some point in time (e.g. if the person is no longer working for the company, but worked for it in the past, the relation is considered valid). It correlates with annotation procedures by other researchers [6].

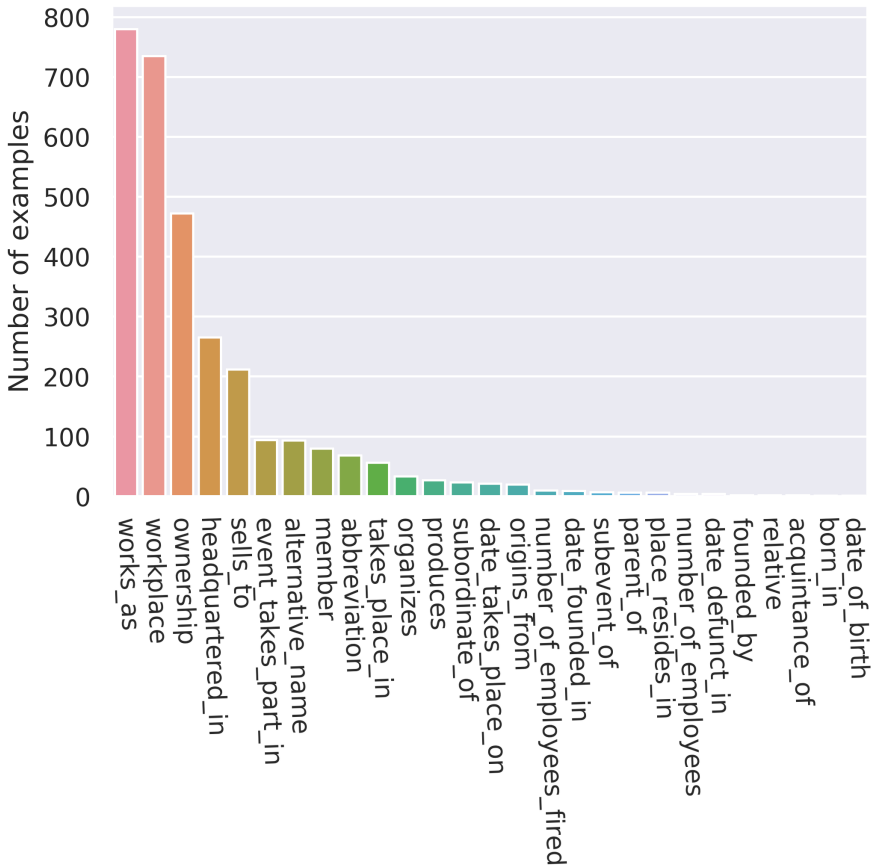


Figure 3: Distribution of relation classes

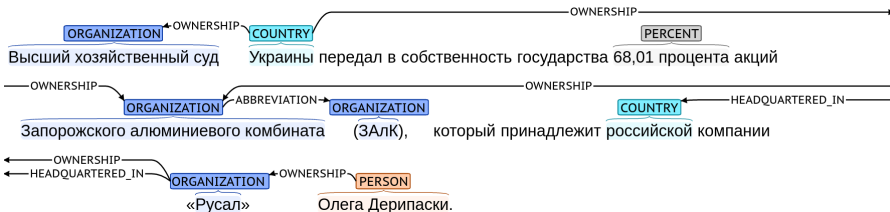


Figure 4: Annotation example. Child organizations and governmental departments are labelled as distinct entities and are connected with relation “OWNERSHIP”. Relations are labelled with disregard to time. Both Ukraine and Rusal own ZAIK

The labelling was performed at the level of sentences. We did not want to complicate the annotation by reference resolution. Moreover, there are existing reference resolution datasets [11] for Russian which can be used together with this corpus. However, in some cases, annotators labelled relations across sentences. They were not removed as they can be easily deleted by post-processing.

Table 2: Relation types

Relation	Parent NERs	Child NERs	Subtype of
PRODUCES	FACILITY, ORGANIZATION	PRODUCT	—
TAKES PLACE IN	EVENT	FAC, ORGANIZATION, GPE	—
DATE TAKES PLACE ON	EVENT	DATE	—
ORGANIZES	GROUP, ORGANIZATION, GPE	EVENT	—
EVENT TAKES PART IN	GROUP, ORGANIZATION, PERSON, GPE	EVENT	—
NUMBER OF EMPLOYEES	FAC, ORGANIZATION	CARDINAL, QUANTITY	—
NUMBER OF EMPLOYEES HIRED	FAC, ORGANIZATION	CARDINAL, QUANTITY	—
NUMBER OF EMPLOYEES FIRED	FAC, ORGANIZATION	CARDINAL, QUANTITY	—
HEADQUARTERED IN	FAC, ORGANIZATION	LOCATION, GPE	—
WORKS AS	PERSON, GROUP	PROFESSION	—
WORKPLACE	PROFESSION, PERSON, GROUP	ORGANIZATION, GPE	—
SUBORDINATE OF	PROFESSION, PERSON, GROUP	PROFESSION, PERSON	—

Relation	Parent NERs	Child NERs	Subtype of
ACQUAINTANCE OF	PERSON, GROUP	PERSON	—
FRIEND OF	PERSON, GROUP	PERSON	ACQUAINTANCE OF
RELATIVE	PERSON, GROUP	PERSON	ACQUAINTANCE OF
PARENT OF	PERSON, GROUP	PERSON	ACQUAINTANCE OF
SIBLING	PERSON	PERSON	ACQUAINTANCE OF
MEMBER	COUNTRY, REGION	GPE	—
OWNERSHIP	LOCATION	LOCATION	—
SELLS TO	GROUP, ORGANIZATION, PERSON, GPE	GROUP, ORGANIZATION, PERSON, GPE	—
ALTERNATIVE NAME	ORGANIZATION	GPE	—
ABBREVIATION	FAC	FAC	ALTERNATIVE NAME
FOUNDED BY	ORGANIZATION	PERSON, ORGANIZATION, FAMILY, GROUP	—
ORIGINS FROM	PERSON, FAMILY, GROUP	NATIONALITY, GPE	—
PLACE RESIDES IN	PERSON	LOCATION, GPE	—
DATE FOUNDED IN	ORGANIZATION, GPE	DATE	—
DATE DEFUNCT IN	ORGANIZATION, GPE	DATE	—
DATE OF DEATH	PERSON	DATE	—
DATE OF BIRTH	PERSON	DATE	—
AGE IS	PERSON	AGE	—
AGE DIED AT	PERSON	AGE	—
BORN IN	PERSON	GPE	—
PLACE OF DEATH	PERSON	GPE, LOCATION	—
SUBEVENT OF	EVENT	EVENT	—

The relation annotation was first individually performed by one of two economists. Afterwards, each annotated text was reviewed and fixed if necessary by one of two linguists to improve annotation quality. Difficult cases were discussed together.

All in all 536 texts were annotated. They contain 6,931 sentences in total, 2,330 of which contain a relation. The average text length is 288 words. In total these texts contain 5,381 relations and 22,595 distinct named entities. The dataset contains 22,846 unique tokens. Class distribution can be seen in the picture (Fig. 3).

We followed the MAMA cycle for dataset labelling [15]. During the annotation procedure, we weekly trained SpanBERT models [8] and inferred predictions for classes with good precision (higher than 0.9) on not-yet annotated data. During inference, only relations with softmax scores higher than 0.9 were selected. This procedure helped us to accelerate annotation, yet it brought some mistakes.

2.3. Inter-annotator agreement

We used Cohen’s Kappa for inter-annotator agreement measurement. For named entities its value is 0.77; for relations it is 0.79. Most disagreements were connected with named entity spans. In many cases, relations labelling disagreement can also be attributed to named entities mismatches (see Fig. 5). It was sometimes the case that the first annotator mistakingly labelled named entity as a whole while our guidelines require to split it into several entities and label a relation between them (if there exists one). Also introducing hierarchy in relations and named entities lead to some disagreement between annotators.

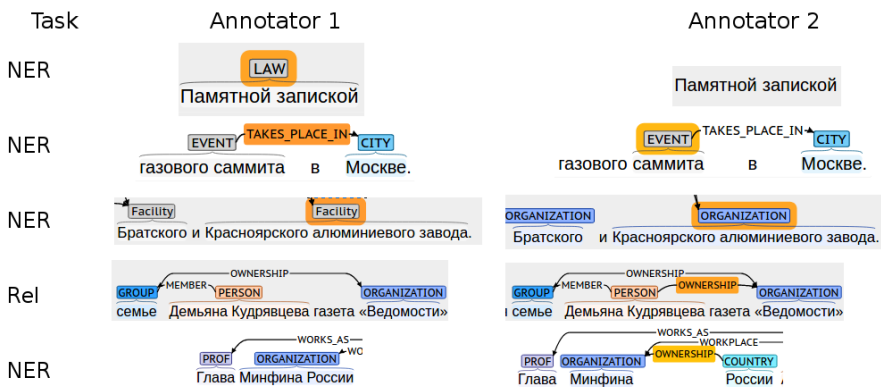


Figure 5: Some disagreement examples

2.4. Annotation Challenges

In many cases, it was difficult to label relations since we did not perform coreference resolution and examined only single sentences. If we had labelled references (e.g. pronouns) as named entities, it would have hindered named entities recognition quality. If we had labelled relations across sentences, it would require to choose among many mentions of the corresponding named entity from the text (sometimes it is not obvious). Moreover, in this case relations tend to span across many sentences which might impose a major difficulty for modern systems [19]. Thus, we decided to skip relations where both entities could not be labelled within a single sentence. In future, we are going to change it.

3. Training models

Using the annotated dataset we trained models for named entity recognition and relation extraction. For all models we randomly separated the sentences from the dataset into train, validation and test datasets using the 0.8: 0.1: 0.1 ratio with disregard to their date and context. We also attempted at fine-tuning them with various learning rates.

3.1. Named entity recognition

For named entity recognition we used a model based on the multilingual BERT model. The BERT [3] system is a transformer-based [19] model that has been pre-trained on a huge text dataset. In the course of pre-training, the tasks of the next sentence prediction and masked language modelling were solved jointly. During the next sentence prediction task, given two input sentences the system determined whether the second sentence is a continuation of the first. In the language modelling problem, the task was to predict masked words using all other words from the sentence. 15% of the words in the original sentence are replaced with a special token [MASK] and, the system predicts words at the positions of the masks. For the next sentence prediction task, a vector representation corresponding to the sentence start token [CLS] was used. For named entity recognition, we use last layer hidden state of the BERT encoder. Afterwards it is passed through a softmax layer. We use cross entropy as our loss function.

To distinguish between the considered entities in the sentence, they are isolated with special tokens representing the beginning and the end of the entity.

3.2. Relation extraction

In this paper, we treat relation extraction as a classification problem. There is a large body of training examples. Each training example is a sentence with a pair of entities and the relationship between them. “No relation” is one of the relations classes.

In this work we use code provided by SpanBERT [7] which is a relation extraction system based on BERT. SpanBERT demonstrated near-SOTA (state-of-the-art) results on the TACRED [21] dataset. Given a sentence, two entities from it and the relation between them (‘no relation’ is a possible outcome) named entities are replaced with their NER-tags. A linear classifier is added on top of [CLS] token to predict the relation type. We also tried replacing BERT-weights with a pretrained RuBERT-model which is finetuned on Russian texts [9]. However, it did not improve our results.

Table 3: Relation extraction and named entity recognition results

task/model	learning rate	dataset	f1	precision	recall
named entities recognition	1e-05	test	0.85	0.83	0.86
relation extraction	1e-05	dev	0.807	0.861	0.760
relation extraction	2e-05	dev	0.805	0.843	0.769
relation extraction	1e-05	test	0.782	0.841	0.731
relation extraction	2e-05	test	0.778	0.814	0.744
relation extraction / rubert	1e-05	dev	0.813	0.861	0.77
relation extraction / rubert	1e-05	test	0.779	0.825	0.739
relation extraction / rubert	1e-05	dev	0.813	0.861	0.77
relation extraction / rubert	1e-05	test	0.779	0.825	0.739

Results for relation extraction and named entity recognition are provided in **Table 3**.

4. Conclusions

In this paper, we have introduced our new dataset containing named entities and relations between them. We also published baseline models that can be used by practitioners and researchers. We hope that this work will enhance NLP-research for the Russian language and will serve as a public baseline for future research on NER and relation extraction. We also hope that it will be of use to the non-scientific NLP community. In future work we would like to broaden dataset domain and to make it even more representative.

5. Licensing

Annotations and trained models are published under MIT license. Lenta.ru news articles are the property of the corresponding copyright holders.

References

1. *Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor.* Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 1247–1249, 2008.
2. *Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lyman, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhрева, and Marat Zaynutdinov.* DeepPavlov: Open-Source library for dialogue systems. In ACL 2018—56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations, pages 122–127, 2018.
3. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. oct 2018.
4. *Ralph Weischedel et Al.* OntoNotes Release 5.0 LDC2013T19. Linguistic Data Consortium, 2013.
5. *Rinat Gareev, Maksim Tkachenko, Valery Solovyev, Andrey Simanovsky, and Vladimir Ivanov.* Introducing baselines for russian named entity recognition. In International Conference on Intelligent Text Processing and Computational Linguistics , pages 329–342. Springer, 2013.
6. *Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy.* SpanBERT: Improving Pre-training by Representing and Predicting Spans. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 13881398, Florence, Italy, 2019. Association for Computational Linguistics.
7. *Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy.* Spanbert: Improving pre-training by representing and predicting spans. arXiv preprint arXiv:1907.10529, 2019.

8. *Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, Omer Levy, and y Allen.* SpanBERT: Improving Pre-training by Representing and Predicting Spans. Technical report.
9. *Yu. Kuratov and M. Arkhipov.* Adaptation of deep bidirectional multilingual transformers for Russian language. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnol.*, volume 2019-May, pages 333–339, 2019.
10. *Artem Kuznetsov, Pavel Braslavski, and Vladimir Ivanov.* Family matters: Company relations extraction from wikipedia. In *International Conference on Knowledge Engineering and the Semantic Web*, pages 81–92. Springer, 2016.
11. *T. A. Le, M. A. Petrov, Y. M. Kurato, and M. S. Burtsev.* Sentence Level Representation and Language Models in The Task of Coreference Resolution for Russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2019)*, pages 341–350, 2019.
12. *Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek.* Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, Atlanta, Georgia, 2013. Association for Computational Linguistics.
13. *Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky.* Distant supervision for relation extraction without labeled data. pages 1003–1011, 2009.
14. *Valerie Mozharova and Natalia Loukachevitch.* Two-stage approach in russian named entity recognition. In *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, pages 1–6. IEEE, 2016.
15. *James Pustejovsky and Amber Stubbs.* Natural language annotation for machine learning. 2013.
16. *Sebastian Riedel, Limin Yao, and Andrew McCallum.* Modeling relations and their mentions without labeled text. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6323 LNAI, pages 148–163, 2010.
17. *A. S. Starostin, V. V. Bocharov, S. V. Alexeeva, A. A. Bodrova, A. S. Chuchunkov, S. S. Dzhumaev, I. V. Emenko, D. V. Granovsky, V. F. Khoroshevsky, I. V. Krylova, M. A. Nikolaeva, I. M. Smurov, and S. Y. Toldova.* FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 702–720, 2016.
18. *Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii.* brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012, Avignon, France, 2012*. Association for Computational Linguistics.
19. *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.* Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 5999–6009, 2017.

20. *Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning.* Position-aware attention and supervised data improve slot filling. In *EMNLP 2017 — Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 35–45, 2017.
21. *Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning.* Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017.

DIALOGUE MANAGEMENT USING EXTENDED DISCOURSE TREES

Ilvovsky D. A. (dilvovsky@hse.ru)

Kazan Federal University, Kazan, Russia; National Research University Higher School of Economics Moscow, Russia

Galitsky B. A. (bgalitsky@hotmail.com)

Oracle Inc, USA

In this paper we learn how to manage a dialogue relying on discourse of its utterances. We consider two complementary approaches of dialogue management based on the discourse text analysis to extend the abilities of the interactive information retrieval-based chat bot.

Keywords: discourse tree, dialogue management, rhetoric structure

DOI: 10.28995/2075-7182-2020-19-361-371

УПРАВЛЕНИЕ ДИАЛОГОМ С ПОМОЩЬЮ РАСШИРЕННЫХ ДИСКУРСИВНЫХ ДЕРЕВЬЕВ

Ильвовский Д. А. (dilvovsky@hse.ru)

Казанский Федеральный Университет, Казань, Россия;
Национальный Исследовательский Университет
Высшая Школа Экономики, Москва, Россия

Галицкий Б. А. (bgalitsky@hotmail.com)

Оракл, США

Ключевые слова: дискурсивное дерево, риторические структуры, управление диалогом

1. Introduction

In this paper we extend the abilities of the interactive chat bot initially developed by [Galitsky and Ilvovsky, 2017] and later improved in [Galitsky, 2019]; [Galitsky and Ilvovsky 2019a]; [Galitsky and Ilvovsky 2019b]. We consider two complementary approaches to the dialogue management both using discourse analysis based on RST [Mann and Thompson, 1988] and. both utilizing Discourse Trees (DT) for the texts.

The first approach is inspired by an idea of a guided search. One source of it is a search methodology designed to show a user an array of different visual possibilities where a searching user may proceed. This is done instead of just navigating to an end point or a terminal answer. We believe that knowledge exploration should be driven by navigating an *extended* discourse tree (EDT) built for the whole corpus of relevant content. It is a combination of discourse trees of individual paragraphs first across paragraphs in a document and then across documents [Galitsky, 2019].

The second approach tries to force the user to request exhaustive information and explanation about the particular topic. We try to achieve this by utilizing discourse tree of the initial piece of a text. At each step of the conversation chat bot analyses remaining topics and try to make user more focused on the initial topic “turning” him back to the undiscussed parts of the text and forcing to request for more details.

As one can see these two approaches are complementary and can be easily combined in one interactive chat bot. In the rest of the paper we discuss these approaches in more details and provide preliminary evaluation for the second approach. More information about general chat bot architecture and evaluation can be found in [Galitsky and Ilvovsky, 2017]; [Galitsky, 2019].

2. Multi Document Navigation Using Extended Discourse Trees

To control the chat bot navigation in a general case, beyond clarification scenarios, one can use the notion of an **extended discourse tree** [Galitsky, 2019]. A conventional discourse tree expresses the author flow of thoughts at the level of paragraph or multiple paragraphs. Conventional discourse tree becomes fairly inaccurate when applied to larger text fragments, or documents. Hence we can extend the notion of a linguistic discourse tree towards an extended discourse tree, a representation for the set of inter-connected documents covering a topic. For a given paragraph, a DT is automatically built by discourse parsers [Joty et.al, 2014]. We then automatically combine DTs for the paragraphs of documents to the EDT, which is a basis of an interactive content exploration facilitated by the chat bot. We apply structured learning of extended DTs to differentiate between good, cognitively plausible scenarios and counter-intuitive, non-cohesive ones. To provide cohesive answers, we use a measure of rhetorical agreement between a question and an answer by tree kernel learning of their discourse trees [Galitsky and Ilvovsky, 2017].

On the web, information is usually represented in web pages and documents, with certain section structure. Answering questions, forming topics of candidate answers and attempting to provide an answer based on user selected topic are the operations which can be represented with the help of a structure that includes the DTs of texts involved. When a certain portion of text is suggested to a user as an answer, this user might want to drill in something more specific, ascend to a more general level of knowledge or make a side move to a topic at the same level. These user intents of navigating from one portion of text to another can be represented as coordinate or subordinate discourse relations between these portions.

We merge the links between logical parts of paragraphs and the links between documents (Fig. 1). If at the current step the user is interested in drilling in, we navigate her through an *Elaboration* relation from nucleus to satellite within a paragraph or *Elaboration* hyperlink to a more specific document. Conversely, if a user decides that the suggested topic is not exactly what he is looking for and wants to return a higher-level view, the system navigates *Elaboration* relation in the inverse order from satellite to nucleus at either paragraph or intra-document level. The other navigation option is relying on *Contrast* or *Condition* relations exploring controversial topics (these rhetorical relations need to be recognized for inter-document case).

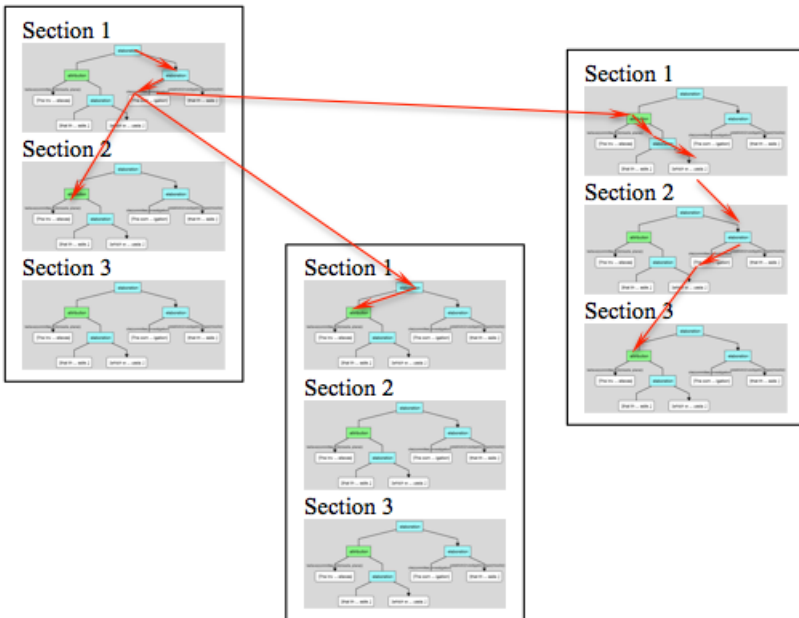


Fig. 1 Illustration for the idea of extended DT: intra-paragraph rhetorical relations are combined with inter-document links also labeled as rhetorical relations

Navigation starts with the route node of a section that matches the user query most closely. Then the chat bot attempts to build a set of possible topics, possible understanding of user intent. To do that, it extracts phrases from elementary discourse units that are satellites of the route node of the DT. If the user accepts a given topic, the navigation continues along the chosen edge; otherwise, when no topic covers the user interest, the chat bot backtracks the discourse tree and proceeds to the other section (possibly of other documents) which matched the original user query second best. Inter-document and inter-section edges for relations such as *Elaboration* play similar role in knowledge exploration navigation to the internal edges of a conventional DT.

3. Using Discourse Tree to Navigate a User through All Aspects of the Topic

3.1. How to Force a User to Request Exhaustive Information and Explanation

In many task-oriented chat bot domains, an objective is to fully inform a user about a particular important piece of information. It is also crucial to make user believe this piece of information, relying on explanation and argumentation in as much degree as possible. In some cases, it is important to make a user believe in a particular short text. This should be done by thoroughly navigating a user through possible disagreements and misunderstanding, to make sure the user is being explained and communicated an issue exhaustively.

- 1) If a text is given, navigating a discourse tree of this text T is one of the most efficient ways to communicate it. The chat bot starts with making an introduction I_T and then making the main statement M_T . Then the user would ask for more details E_T , disagree with the E_T or ask a question on a topic outside of the scope of this text O_T .
- 2) If the user asks for more details I_T , the EDU connected with *Elaboration* with M_T is provided as a reply. We denote this EDU as *Elaboration*(I_T). This is the easiest, most direct situation.
- 3) If the user disagrees, chat bot tries to find an EDU which is connected by *Explanation* or *Cause* with M_T or I_T . This EDU should be returned as a reply.
- 4) If the user asks a different question O_T then it should be answered as a factoid question but nevertheless the chat bot needs to take the user back to T so the reply should end with *Elaboration*(I_T).
- 5) If the user doubts about the validity of a claim in M_T , the chat bot needs to deliver *Attribution*(M_T) as an answer.

The procedure above should iterate until no more EDU in T is left or the user terminates the conversation. If the chat bot persistence is too high in trying to take the user back to T , this user would terminate the conversation too soon. Otherwise, if the chat bot persistence is too low, the user would deviate from T too far so will red less content of T (EDU(T)). We want to optimize the chat bot to maintain the optimal persistence to maximize the number of delivered EDU(T) till the conversation is abandoned by the user.

The chat bot can use different modes interacting with the user:

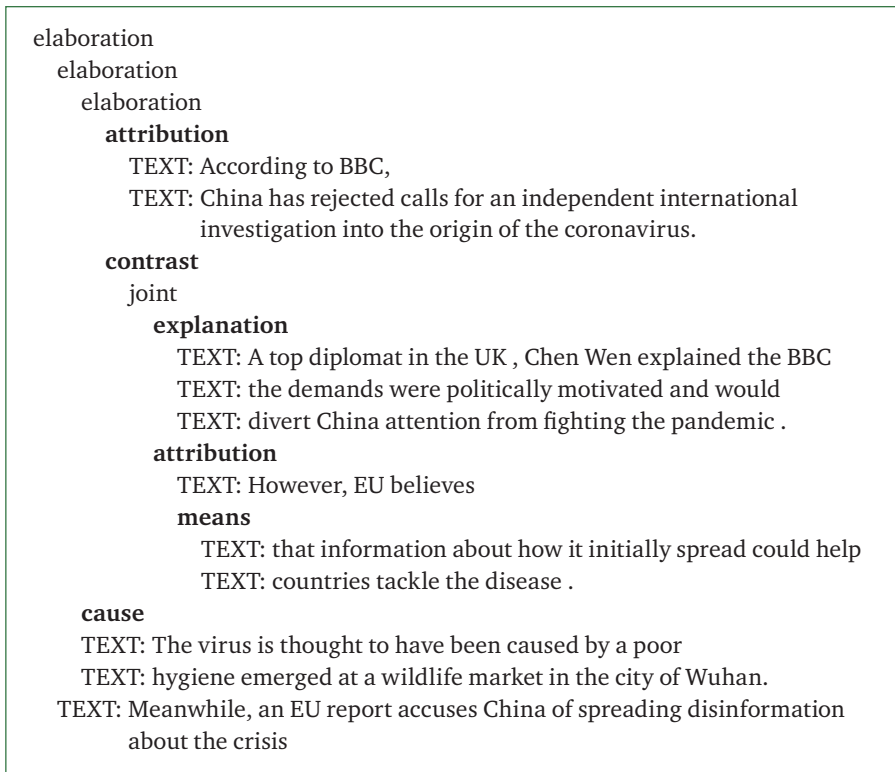
- 1) Always try to find an EDU matching the user query. Always give the closest EDU as an answer, even if it is very dissimilar to the question.
- 2) Try to find an EDU matching the user query. If a good match is not found then go to a foreign content. Then proceed with DT navigation. Give a foreign answer and then DT navigation EDU.

- 3) Try to find an EDU matching the user query. If a good match is not found then encourage a user to ask a question or make a comment which would match a rhetorical relation in a navigation path.
- 4) Try to search foreign content as much as possible.

3.2. Example of a Navigation using “Forcing” Approach

Initial Text: *According to BBC, China has rejected calls for an independent international investigation into the origin of the coronavirus. A top diplomat in the UK, Chen Wen explained the BBC the demands were politically motivated and would divert China attention from fighting the pandemic. However, EU believes that information about how it initially spread could help countries tackle the disease. The virus is thought to have been caused by a poor hygiene emerged at a wildlife market in the city of Wuhan.*¹

Discourse Tree for the text



¹ <https://www.bbc.com/news/world-asia-china-52420536>

Sample dialogue with the technical comments

Bot: “China has rejected calls for an independent international investigation into the origin of the coronavirus”

User: Why would they do it? Don’t believe that.

Bot: [doubt is detected] [(*Attribution* (“China has rejected...”) = According to BBC] *This is* “According to BBC”

User: Are politicians involved in this?

Bot: [Searching for ‘politi...’ & ‘involv...’ tokens in EDUs] “the demands were politically motivated”

User: Why is that?

Bot: [Detected explainability request. Looking for *Explanation* (“the demands were politically...”).] “A top diplomat in the UK, Chen Wen explained the BBC

User: Who is Chen Wen?

Bot: [question is about an entity from the answer but foreign to the thread. So search on the web and then continue navigate DT] Chen Wen is ... [Found information on the web]. *Notice also that* [*Elaboration, Joint, or other default rhetorical relation from the current node*] “and would divert China attention from fighting the pandemic.”

User: And how can countries tackle coronavirus?

Bot: [Searching for ‘countr...’ & ‘tackle...’ & ‘coronavirus’ tokens in EDUs & rhetorical relations related to ‘how’ such as *Explanation, Cause, Means* etc].

Selected fragment:

EU believes

means

TEXT: that information about how it initially spread could help

TEXT: countries tackle the disease.

“EU believes that information about how it initially spread could help countries tackle the disease”

User: Why is it EU business?

Bot: [question is about an entity from the answer but foreign to the thread.

So search on the web and then continue navigate DT]

[Found node with EDU: The virus is thought to have been caused ... Formed question for this EDU and the relation]

Do you know what causes “the virus”

User: Yes I know, it is caused by ...

Bot: [User confirmed. It does not matter what the user knows. Proceed to the next EDU & relation]

“It is caused by a poor hygiene emerged at a wildlife market in the city of Wuhan”

In *italic* we show the phrases added by the chat bot. In [square brackets] we show the results of chat bot analysis of utterances and a reply selection rule/result. EDUs from original text are shown in “double quotes”.

What we observe is that a dialogue is fairly plausible, although no data-driven method has been applied. It does not matter where the user deviates from the target text as long as the chat bot always takes her back to the EDU and rhetorical relation which is either relevant to what the user asked or claimed, or just follows the DT navigation flow from more important, closer to the root, to less important. If the user is asked a question outside of the scope of the target text, we provide an answer from the foreign source and then also switch topic and come back to the business of the target answer, proceeding with the DT navigation.

Finally, we show just the **readable sample dialogue**:

Bot: China has rejected calls for an independent international investigation into the origin of the coronavirus
User: Why would they do it? Don't believe that.
Bot: This is According to BBC
User: Are politicians involved in this?
Bot: The demands were politically motivated
User: Why is that?
Bot: A top diplomat in the UK, Chen Wen explained the BBC
User: Who is Chen Wen?
Bot:
 Chen Wen is ...
 and would divert China attention from fighting the pandemic
User: And how can countries tackle coronavirus?
Bot: EU believes that information about how it initially spread could help countries tackle the disease
User: Why is it EU business?
Bot: Do you know what causes the virus?
User: Yes I know, it is caused by ...
Bot: It is caused by a poor hygiene emerged at a wildlife market in the city of Wuhan

4. Preliminary Evaluation

We tried to estimate how reasonable is to “force” a user to request more knowledge using our approach. We evaluated the overall meaningfulness of the utterances produced by the bot and average “level of user satisfaction”.

The topics of the dialogues included product features, technical instructions, and latest news. Each discussion was based on the source text as it was shown in the example in [Section 3.2](#). Five users (author's colleagues) interacted with the chat bot on these topics, and stopped when they think that dialogue is not useful for them anymore. After each conversation users assessed the meaningfulness of each bot utterance (just “yes”/“no”) and then rated the whole “usefulness” of a dialogue using grades from 1 to 3 where 3 mean “dialogue was useful for me”.

Table 1. The overall meaningfulness of the dialogues with the bot

Domain & source	News	Instructions	Product features
Number of dialogues	40	40	45
Average number of utterances	9	6	10
Percentage of meaningful bot utterances	0.73	0.69	0.78
Average user satisfaction	2.2	1.9	2.3

5. Related Work

Discourse disentanglement (such as classification of links between portions of texts or documents) and dialogue/speech/communicative act tagging have been extensively studied [Wang et al., 2011]. Discourse disentanglement is the task of splitting a conversation [Elsner and Charniak, 2008] or documents [Wolf and Gibson, 2005] into a sequence of distinct portions of text (sub-discourses). The disentangled discourse is modeled via a tree structure [Grosz and Sidner 1986]; [Seo et al., 2009], an acyclic graph structure [Rose et al., 1995]; [Elsner and Charniak, 2008], or a cyclic chain graph structure [Wolf and Gibson, 2005]. Speech acts are used to describe the function or role of an utterance in a discourse, similarly to our CDT representation, and have been employed for the analysis of communication means including conversational speech instant messaging, security analysis of documents [Galitsky and Makowski, 2017], online forums [Kim et al., 2010], [Galitsky et al., 2017] and chats [Galitsky and Ilvovsky, 2017a]. Automated answer scoring benefits from semantic and discourse analyses as well [Wanas et al., 2008]. For a more complete review of models for discourse disentanglement and speech act tagging, we refer the reader to [Kim et al., 2010].

[Wang et al. 2011] presented the task of parsing user forum threads to determine the labeled dependencies between posts. Three methods, including a dependency parsing approach, are proposed to jointly classify the links (relationships) between posts and the dialogue act (type) of each link. The authors predicted not only the links between posts, but also showed the type of each link, in the form of the discourse structure of the thread.

6. Conclusions and Future Work

We presented a discourse-based dialogue management system for a chat bot with iterative content exploration that leads a user through a personalized knowledge acquisition session. The chat bot is focused on automated customer support or product recommendation agent that assists a user in learning product features, product usability, suitability, troubleshooting and other related tasks.

Although there has been a substantial advancement in document-level RST parsing, including the rich linguistic features-based of parsing models [Joty et al., 2014], document level discourse analysis has not found a broad range of applications such as search. The most valuable information from DT includes global discourse features and long range structural dependencies between DT constituents.

A number of studies including [Surdeanu et al., 2015] showed that discourse information is beneficial for search. Our paper is explicitly showing how discourse trees help to navigate search. Discourse trees and their extensions is a very promising subject of study for logical AI. Logical AI studies subjects such as logic forms and logic programs which are very limited in quantity in the real world. But logical AI tries to make sense of them: discourse trees are fairly interpretable structures. Discourse trees can be obtained in large quantity on one hand and they are adequate Logical AI subject on the other hand. That is why discourse trees and their extension is such an important subject of study for search engineering and chat bots.

In question answering the current version of chat bot relies only to the data extracted from text documents. Now we are working on complementing these data by the data from Linked Open Data cloud, including DBpedia [Lehmann et al., 2015]. As an interface between natural language user query and LOD datasets we would rely on the resource from the Linguistic Linked Open Data cloud [Cimiano et al., 2020], such as LLOD representation of WordNet [McCrae et al., 2014], BabelNet [Ehrmann et al., 2014], FrameNet [Rospocher et al., 2019] and RuThes [Kirillovich et al., 2017]; [Galieva et al., 2017]. We expect that exploitation of LOD cloud can improve user's satisfaction against the baseline obtained in this work.

7. Acknowledgements

The new results on discourse-based navigation proposed in this paper were received by Dmitry Ilvovsky during his work in KFU and was funded by Russian Science Foundation according to the research project no. 19-71-10056. These results are based on the research in discourse models that were by made by Dmitry Ilvovsky and Boris Galitsky in NRU HSE during last years.

References

1. Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). Linguistic Linked Open Data Cloud. In Cimiano, P., et al. Linguistic Linked Data: Representation, Generation and Applications, pages 29–41. Springer.
2. Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., and Navigli, R. (2014). Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In Calzolari, N., et al., Eds., Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pages 401–408. ELRA.
3. Elsner, M. and E. Charniak. (2008) You talking to me? a corpus and algorithm for conversation disentanglement. In Proceedings of the 46th Annual Meeting of the ACL: HLT (ACL 2008), pages 834–842, Columbus, USA.
4. Galieva, A., Kirillovich, A., Khakimov, B., Loukachevitch, N., Nevzorova, O., and Suleymanov, D. (2017). Toward Domain-Specific Russian-Tatar Thesaurus Construction. In Proceedings of the International Conference IMS-2017, pages. 120–124. ACM.

5. Galitsky B. and G. Makowski (2017) Document Classifier for a Data Loss Prevention System based on Learning Rhetoric Relations. CICLing 2017, April 17–23, Budapest, Hungary.
6. Galitsky B. and McKenna E. W. (2017) Sentiment Extraction from Consumer Reviews for Providing Product Recommendations. US Patent 9646078B2.
7. Galitsky B., Ilvovsky D. (2017) Chat bot with a Discourse Structure-Driven Dialogue Management, EACL Demo Program.
8. Galitsky B., Ilvovsky D. (2019b) On a Chatbot Conducting Virtual Dialogues. Proceedings of the 28th ACM International Conference on Information and Knowledge Management.
9. Galitsky B., Ilvovsky D., Goncharova, E. (2019a) On a Chatbot Conducting Dialogue-in-Dialogue. Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue.
10. Galitsky B. (2019) Developing Enterprise Chatbots – Springer International Publishing, 2019. Chapter 11. Discourse Level Dialogue Management.
11. Galitsky, B., M. P. González, C. I. Chesñevar (2009) A novel approach for classifying customer complaints through graphs similarities in argumentative dialogue. Decision Support Systems, Volume 46, Issue 3, pp. 717–729 (2009).
12. Grosz B. J. and C. L. Sidner (1986) Attention, intention and the structure of discourse. Computational Linguistics, 12(3):175–204.
13. Joty S. R. and A. Moschitti (2014) Discriminative Reranking of Discourse Parses Using Tree Kernels. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).
14. Kim, S. N., Li Wang, and Timothy Baldwin. (2010) Tagging and linking web forum posts. CoNLL-2010, pages 192–202, Uppsala, Sweden
15. Kirillovich, A., Nevzorova, O., Gimadiev, E., and Loukachevitch, N. (2017). RuThes Cloud: Towards a Multilevel Linguistic Linked Open Data Resource for Russian. In Różewski, P. and Lange, C., Eds. Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017). Communications in Computer and Information Science, vol. 786, pages. 38–52. Springer, Cham.
16. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia: A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 6(2):167–195.
17. Mann, W. and Thompson, S., *Rhetorical Structure Theory: Toward a functional theory of text organization, 1988*
18. McCrae, J. P., Fellbaum, C., and Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. In Chiarcos, C., et al., Eds., Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014), pages. 13–16. ELRA.
19. Rose CP, Barbara Di Eugenio, Lori S. Levin, and Carol Van Ess-Dykema (1995) Discourse processing of dialogues with multiple threads. ACL.
20. Rospocher, M., Corcoglioniti, F., and Palmero Aprosio, A. (2019). PreMON: LODifying linguistic predicate models. Language Resources and Evaluation, 53:499–524.

21. *Sakai, T.* (2007). Alternatives to Bpref. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, The Netherlands, ACM: 71–78.
22. *Seo, J., W. B. Croft, and D. A. Smith* (2009) Online community search using thread structure. CIKM 2009, pages 1907–1910, Hong Kong, China.
23. *Surdeanu M., T. Hicks, and M. A. Valenzuela-Escarcega.* Two Practical Rhetorical Structure Theory Parsers. NAACL HLT, 2015.
24. *Wanas, N., M. El-Saban, H. Ashour, and W. Ammar* (2008) Automatic scoring of online discussion posts. In Proceeding of the 2nd ACM workshop on Information credibility on the web (WICOW '08), pages 19–26, Napa Valley, USA.
25. *Wang, L., Lui, M., Kim, S. N., Nivre, J., Baldwin, T.:* Predicting thread discourse structure over technical web forums. EMNLP. pp. 13–25 (2011)
26. *Wolf, F. and E. Gibson* (2005) Representing discourse coherence: A corpus-based study. Computational Linguistics, 31(2):249–287.

КОЛИЧЕСТВЕННЫЙ МЕТОД АНАЛИЗА КОННЕКТОРОВ: «ПОРТРЕТ» РУССКОГО СОЮЗА *ИЛИ* В НАДКОРПУСНОЙ БАЗЕ ДАННЫХ КОННЕКТОРОВ

Инькова О. Ю. (Olga.Inkova@unige.ch)

ИПИ ФИЦ ИУ РАН, Москва, Россия; Женевский университет, Женева, Швейцария

DOI: 10.28995/2075-7182-2020-19-372-386

A QUANTITATIVE METHOD FOR ANALYSIS OF CONNECTIVES: THE “PORTRAIT” OF THE RUSSIAN CONJUNCTION *ИЛИ* (OR) IN THE SUPRACORPORA DATABASE OF CONNECTIVES

Inkova O. Yu. (Olga.Inkova@unige.ch)

Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland

The functional properties of the conjunction *или* ‘or’ are quite well studied and discussed in grammars and number of specific studies. However, they were not subjected to multivariate quantitative analysis. The paper proposes this kind of analysis, carried out according to six parameters: i) logical-semantic relation expressed by the conjunction, ii) syntactic structure of the text fragment introduced by it, iii) position of the conjunction in this text fragment, iv) order of the text fragments connected by it, v) status of *или* in the context (cf. its use as a particle mentioned in MAS), vi) disposition of the elements that make up the multiword connectives (cf. *или ... или prosto* ‘or ... or just’).

The analysis of the formal variants of *или*, carried out using the Supracorpora database of connectives, made it possible to formulate the conclusions that follow and to confirm them with quantitative data. i) Always occupying the initial position in the text fragment introduced by him, *или* is used mainly as a connective. ii) The most typical order of text fragments for formal variants with *или* is the p CNT q scheme. iii) By its syntactic characteristics *или* is directly opposite to *и* ‘and’: *или* is used in written texts in most cases for a non-predicative coordination. iv) Formal variants with *или* express the relation of alternative at all three semantic levels (propositional, illocutive, metalinguistic), as well as the relations of substitution, correction, and negative alternative, but with a clear predominance of propositional alternative. v) *или* forms multiword and two- and multicomponent connectives; their composition varies depending on the relation expressed by them.

Keywords: linguistics, semantics, quantitative analysis, connectives, Russian, *или*

1. Вводные замечания

Союз *или* не был обделен вниманием лингвистов: он упоминается во всех грамматиках и словарях, ему посвящены многочисленные исследования (см., например, [Ананьева 2005], [Богуславский 1996], [Гладкий 1979], [Инькова-Манзотти 2001], [Падучева 1964], [Перетрухин 1979], [Санников 1985, 1989], [Урысон 2004, 2010]). Однако почти все они выполнены в докорпусную эпоху: немногочисленные примеры часто сконструированы исследователями, полагающимися больше на свою языковую интуицию, чем на реальный языковой материал, который, если и присутствует, то не является предметом количественного анализа. В связи с созданием корпусов и лингвистических ресурсов у исследователей появилась возможность проверить свое языковое чутье на представительном языковом материале и квантифицировать функциональные особенности языковых единиц.

Одним из таких ресурсов является Надкорпусная база данных (НБД) коннекторов, разработанная для исследования языковых единиц, принадлежащих к этому функциональному классу. НБД функционирует на основе русско-французского и русско-итальянского параллельных подкорпусов Национального корпуса русского языка (НКРЯ), включающих наряду с художественными текстами 19–20-го вв., также научную литературу (М. Бахтин, Н. Бердяев, С. Булгаков, В. Соловьев, Ж. Женетт), публицистику (А. Политковская) и «Конституцию РФ». Общий объем текстов — около 9,2 млн словоупотреблений.

Система аннотирования в НБД включает набор параметров, отражающих формальную вариативность коннекторов и особенности их функционирования:

- выражаемое коннектором логико-семантическое отношение (ЛСО),
- синтаксическая структура вводимого коннектором фрагмента текста,
- позиция коннектора в этом фрагменте текста,
- порядок следования фрагментов текста, связанных коннектором,
- для полифункциональных единиц — их статус в данном контексте: коннектор или не коннектор (ср., например, *только* или *наконец*, которые далеко не во всех случаях своего употребления являются коннекторами),
- расположение элементов, входящих в состав неоднословного коннектора (контактное или дистантное)¹.

НБД дает возможность получать количественные данные по каждому из перечисленных параметров и их комбинациям; подробнее см. [Inkova, Popkova 2017]. Подчеркнем, что речь идет о многопараметровом анализе функционирования коннекторов и что корреляций между значениями каждого из параметров на сегодняшний день не выявлено. Сопоставление коннекторов целесообразно поэтому вести учитывая их значения по каждому из перечисленных параметров, разработанных с учетом функциональных особенностей единиц данного класса.

¹ Подробнее о концепции НБД, ее устройстве и ее функциональных возможностях см. [Инькова 2018a], [Инькова 2018b]. Представительный фрагмент НБД доступен по адресу: <http://a179.frccsc.ru/RSCF10004/main.aspx>

В данной статье мы предлагаем «нарисовать портрет» русского союза *или* с использованием информации, хранящейся в НБД. На 31.12.2019 методом сплошной выборки сформированы 1162 двуязычные аннотации с речевыми реализациями (РР) коннекторов², включающих в свой состав *или* (15,6% от общего числа контекстов с *или* в НБД): 523 аннотации в направлении перевода русский язык → французский язык, 226 в направлении французский язык → русский язык, 301 — в направлении русский → итальянский, 112 — итальянский → русский. Из этого количества были исключены аннотации с лексикализованными употреблениями *или* (*более или менее, рано или поздно, тот или иной* и др.). Общее количество аннотаций для всех направлений перевода составило, таким образом, 1141.

Для анализа синтаксических свойств были использованы только аннотации употребления РР *или* в русских оригинальных текстах (804: 507 в направлении русский → французский и 297 — русский → итальянский), чтобы избежать возможного влияния структуры иноязычного оригинала на структуру перевода, потенциально способного исказить реальную картину синтаксических свойств *или*. Переводы на русский язык используются для семантического описания: они позволяют высветить семантические особенности РР с *или*, которых на 31.12.2019 в НБД для всех направлений перевода зарегистрировано 92; наиболее частотная РР — одноэлементное *или* (70,1%).

2. «Портрет» *или* в НБД

2.1. Параметры «Позиция», «Статус», «Порядок» и «Расположение»

Начнем наш анализ с наиболее предсказуемых результатов. Из трех возможных для коннекторов позиций во вводимом им фрагменте текста (начальная, конечная, неначальная) *или* может употребляться только в начальной позиции, что неудивительно для сочинительного союза. Употребление, в котором *или*, не является коннектором (параметр «Статус»), в НБД не зафиксировано; речь идет о случаях типа (1), в котором *или* характеризуется словарями как частица, за исключением [МАС], считающего *или* в данном случае «вопросительным союзом».

- (1) Говорит один рыбак: Не поймать сома никак. *Или* снасти не крепки?
Или мы не рыбаки? [С. Михалков. Про сома; пример БАС]

Наибольший интерес представляют данные по параметру «Порядок» (Таблица 1).

² Речевой реализацией (РР) предлагается называть ту форму, в которой коннектор зафиксирован в конкретном контексте. Такой подход позволяет учесть формальное варьирование коннекторов и реальный состав связующих средств русского языка [Инькова 2020]. На 31.12.2019 в НБД зарегистрировано более 1300 РР.

Таблица 1. Статистика по параметру «Порядок» в НБД

Порядок	ру-фр всего аннотаций	ру-фр %	ру-ит всего аннотаций	ру-ит %	среднее для русских оригиналов (%)
p CNT q	422	83,2% ³	265	89,2%	86,2%
CNT p CNT q	55	10,8%	23	7,7%	9,3%
CNT p, CNT q, CNT r, ...	16	3,1%	7	2,4%	2,7%
CNT p [= CNT r, CNT s] CNT q	14	2,8%	3	1,0%	1,9%

Самой частотной из четырех зафиксированных в НБД структур, оформляемых РР с *или*, причем со значительным отрывом, является двухчленная структура, в которой коннектор вводит второй фрагмент текста: p CNT q.

- (2) Полномочия судьи могут быть прекращены *или* приостановлены не иначе как в порядке и по основаниям, установленным федеральным законом. [Конституция РФ (1993)]

Наиболее частотной РР и здесь оказывается одноэлементное *или*, на долю которого приходится 80,2%. Зафиксированы также многоэлементные РР, т.е. сочетания *или* с т.н. конкретизаторами, как правило, с контактными расположением элементов. Только две из 51 многоэлементной РР: *или|может быть*⁴ и *или|например*, характеризуются дистантным расположением входящих в их состав элементов (такую информацию позволяет получить параметр «Расположение»).

- (3) <...> Пушкин или Ломоносов хоть стихи и писали, а такую штуку никогда бы не устроили... *или* Надсон, *например*. [М. А. Булгаков. Белая гвардия (1924)]

Случаи, когда коннектор вводит последний член перечислительного бессоюзного ряда, также размечаются как имеющие порядок p CNT q.

- (4) В этом «грубо» нельзя жить биографической жизнью, — здесь, можно только переживать кризис, принимать последние решения, умирать *или* возрождаться [М. М. Бахтин. Проблемы поэтики Достоевского (1963)].

Следующей по частотности является также двухчленная конструкция, но в которой каждый из фрагментов текста маркируется компонентом РР: CNT p, CNT q. Это могут быть как симметричные РР (с повторяющимся *или*, в том числе в сочетании с *же*: *или||или, или||или же*), так и асимметричные, в которых компонент с *или* является вторым (*будь то||или же, ли||или, ли||или же, либо||или, то ли||или*):

³ Числа в таблицах приведены с округлением до десятой, поэтому их сумма в колонках или в строках может не равняться 100.

⁴ Одиночная вертикальная черта в форме РР означает, что входящие в ее состав элементы разделены текстом, двойная вертикальная черта — что разделяемые ей компоненты маркируют разные фрагменты текста.

- (5) И вот в зависимости от этого — дает вождь или не дает — армия *то ли* поддерживает этого конкретного вождя, *или* противодействует ему. [А. С. Политковская. Путинская Россия (2004)]

Примерно одинаковую частотность имеют остальные две структуры. Первая представляет собой открытый ряд, оформляемый симметричными, как в (6), так и асимметричными РР, как в (7). Максимально длинный ряд в НБД состоит из пяти повторяющихся *ли*, замыкаемых *или*.

- (6) А главный человек в розыске — свидетель, потому что в самом тайном делишке всегда отыщется человек, который *или* что-то видел, *или* слышал, *или* знает, *или* помнит, *или* догадывается. [А. и Г. Вайнеры. Эра милосердия (1975)]
- (7) Если б не эта тарелка, да не прислоненная к постели только что выкуренная трубка, *или* не сам хозяин, лежащий на ней, то можно было бы подумать, что тут никто не живет [И. А. Гончаров. Обломов (1848–1859)]

Вторая структура (порядок CNT p [= CNT r, CNT s] CNT q) представляет собой сложносочиненное предложение, оформляемое показателем ЛСО «вопреки ожидаемому», первая часть которого — сложносочиненное предложение с ЛСО альтернативы, выражаемым РР с *или*, одно- или двухкомпонентной. Ср. (8) с однокомпонентной РР с *или*:

- (8) Сильно трясло на ухабах заледеневшей мостовой *или* руки у меня так сильно тряслись, *но* свернуть сигарку никак не удавалось — все время табак просыпался. [А. и Г. Вайнеры. Эра милосердия (1975)]

ЛСО альтернативы, передающее элемент гипотетичности, является неотъемлемой частью общей семантики конструкции, значение которой можно сформулировать так: несмотря на то, что говорящему неизвестна некоторая информация о сообщаемом положении вещей (первая часть конструкции, оформляемая РР с *или*), он, тем не менее, делает некоторое утверждение (вторая часть конструкции, оформляемая РР со значением «вопреки ожидаемому»). Такие структуры являются не просто иерархическими, когда в правом или левом контексте коннектора есть еще один коннектор, не обязательный для реализации значения «главного» коннектора (ср. мириаду коннекторов в правом контексте *но* в примере (3) выше), а единую конструкцию.

2.2. Параметр «Структура»

О том, что *или* употребляется чаще всего в рамках предложения, простого или сложного, хорошо известно, но это утверждение, как правило, не опирается на количественный анализ. В Таблице 2 приводятся данные для одноэлементного *или* (318 аннотаций в направлении русский → французский и 235 — в направлении русский → итальянский), подтверждающие это утверждение.

Таблица 2. Статистика по параметру «Структура» для РР *или* в НБД⁵

	ру-фр	ру-ит	среднее (%)
Самостоятельное предложение	13 (4,1%)	9 (3,8%)	3,9%
Сложное предложение	2 (0,6%)	1 (0,4%)	0,5%
С предикацией	80 (25,1%)	38 (16,2%)	20,6%
Б/предикации (глагол)	19 (6,0%)	13 (5,5%)	5,7%
Б/предикации (не-глагол)	197 (61,7%)	138 (58,7%)	60,2%
Вставка	7 (2,2%)	36 (15,3%) ⁵	8,7%

Доля *или*, вводящего самостоятельное предложение (простое и сложное), составляет около 4%; в НБД зафиксированы случаи введения *или* повествовательных, побудительных и вопросительных, как в (9), предложений.

(9) — Вы бы, гражданин Жеглов, скинули пиджачок, а то вам не с руки играть-то будет. *Или* вы за пушку свою опасаетесь? — вежливо спросил Копченый. [А. и Г. Вайнеры. Эра милосердия (1975)]

Вводить сложное предложение в рамках сложного предложения — см. пример (10) — для *или* еще менее характерно (всего 3 случая в НБД).

(10) Разрушенное здание счастья погребло бы ее под развалинами, *или*, если б еще уцелели ее силы, она бы — искала... [И. А. Гончаров. Обломов (1848–1859)]

Более типичной (чуть более 20%) для фрагмента текста, вводимого *или*, является предикативная структура.

(11) Все имеет свое бытие лишь в понятии, *или* все есть бытие понятия. [Вл. Соловьев. Кризис западной философии (1874)]

Нолидируют здесь структуры с однородными сказуемыми: 61,2% в направлении русский → французский и 63,7% в направлении русский → итальянский.

(12) Вспоминаем друг о друге к концу жизни, когда кто тяжело заболит *или* помрет. [Ч. Айтматов. Прощай, Гульсары!]

Наиболее частотным для *или* является употребление для непредикативного сочинения (почти 66,0%), большинство (60,0%) составляют сочиненные неглагольные формы:

(13) Родители *или* лица, их заменяющие, обеспечивают получение детьми основного общего образования. [Конституция РФ (1993)]

Однородные глагольные формы (деепричастия и причастия) составляют 5,7%.

⁵ Значительная разница в количестве вставочных конструкций объясняется тем, что они характерны для оформления метаязыковой альтернативы (см. [раздел 2.3](#)), чаще встречающейся в научных текстах, которых в НБД больше в направлении русский → итальянский.

- (14) Одни закусывали, стоя *или* присев к столу <...>. [Л. Н. Толстой. Анна Каренина (1873–1877)]

Наконец, вставочные предложения различной степени сложности составляют в среднем около 9,0%:

- (15) Потом — *или* раньше? — я видела ее в Доме литераторов на вечере памяти Блока. [Л. Чуковская. Записки об Анне Ахматовой (1976)]

На основании этих данных можно отнести *или* к «жестким союзам» (в терминологии [Урысон 2013]), для которых характерно связывать простые предложения в составе сложного. Приведем для сравнения данные для союза *и* в письменном подкорпусе из [Подлеская 2019]: на предикативное сочинение приходится 70,5%, на непредикативное — 15,5% (оставшиеся 14,2% приходятся на несочинительные употребления *и* в лексикализованных употреблениях и в функции частицы), т. е. ситуация прямо противоположная, чем для *или*. Сближает оба союза тот факт, что в рамках предикативного сочинения в письменном подкорпусе оба отдают предпочтение сочинению глагольных групп: 58,6% для *и* и 62,2% для *или*. Данные по союзам *но* и *а* помогут установить, является ли это общей характеристикой сочинительных союзов.

2.3. Параметр «Отношение»

Количество выделяемых *у* *или* значений, а также их названия существенно разнятся в словарях и исследованиях. Достаточно сравнить описания *или* в основных словарях. Наиболее подробное описание значений *или* содержится в МАС: i) разделительный; см. пример (2) выше, ii) разделительно-перечислительный; см. (6) выше, iii) присоединительный, как в (4) выше, iv) противительный; см. (16), v) пояснительный; см. (11) выше, vi) вопросительный, как в (1) выше.

- (16) Прокурор его вызвал, — в двадцать четыре часа освободи площадь *или* пойдешь в лагерь на пять лет, детей в детдом заберем. [В. С. Гроссман. Жизнь и судьба (1959)]

В исследованиях высказываются разные точки зрения относительно значения *или*, часто диаметрально противоположные (нестрогая, инклюзивная дизъюнкция vs. строгая, эксклюзивная, дизъюнкция), а количество значений может доходить до девяти [Ананьева 2005]. При этом затемняется тот общий механизм, который лежит в их основе, а критерии, положенные в основу классификаций, делают несопоставимыми полученные результаты⁶. В НБД используется классификация ЛСО, последовательно различающая три уровня, на которых они могут быть установлены; подробнее см. [Инькова 2019]: i) пропозициональный уровень; ii) уровень высказывания (иллокутивные употребления); iii) метаязыковой уровень.

РР с *или* выражают, в подавляющем большинстве своих употреблений, ЛСО альтернативы (разделительное в терминах русской грамматики) на всех трех уровнях: оно установлено на пропозициональном уровне, связывая положения вещей, описанные в соединяемых фрагментах текста, например, в (12) выше; на уровне высказывания — в (9) выше, поскольку для его интерпретации необходимо принять во внимание коммуникативные характеристики связываемых фрагментов текста (вопрос, вводимый *или*, затрагивает здесь условия успешности речевого акта, осуществленного в предыдущем высказывании); на метаязыковом уровне, т. е. касается выбора словесной формы; см. (11) выше, где второй фрагмент текста переформулирует сказанное в первом.

Разметка по параметру «Отношения» в НБД не дает более дробной классификации значений, хотя фасетная классификация, лежащая в ее основе, позволяет при необходимости это сделать. Мы, в частности, отказались от разметки в терминах «инклюзивной» и «эксклюзивной» альтернативы, применяемой, например в [PDTV 3.0], поскольку эта оппозиция является трансверсальной и не позволяет построить иерархическую классификацию значений РР с *или*, предпочтительную при аннотировании корпуса: пропозициональная альтернатива может быть как инклюзивной, так и эксклюзивной; иллокутивная альтернатива всегда эксклюзивна, метаязыковая, напротив, инклюзивна. Кроме того, выбор инклюзивной и эксклюзивной интерпретации является в значительной степени контекстно или даже ситуативно обусловленным, особенно в случае одноэлементного *или* (на это обращают внимание и логики; см., в частности, [Кондаков 1975], [Gamut 1991]). Хорошо известный вопрос *Fromage ou dessert?*, задаваемый официантом в конце ужина и приводимый Б. Де Корнюлье [De Cornulier 1985] в качестве примера, иллюстрирующего это положение, понимается как эксклюзивная альтернатива только в данной ситуации и только если вы выбрали фиксированное меню. Но ничто, кроме состояния

⁶ Так, в [Санников 1989] классификация разделительных союзов основана на типах речевых актов, но в действительности ограничивается утверждениями; тогда как в [Ананьева 2005] этот критерий никак не учитывается, а предлагается классификация, основанная на степени абстрактности разделительного значения. Например, «альтернативно-перечислительное» значение является самым «конкретным» (самый высокий, пятый, уровень абстракции), являясь разновидностью «альтернативно-временного» значения, которое, в свою очередь, является видом «альтернативно-определенного», при котором говорящий уверен, что одна из ситуаций имеет место, которое, в свою очередь, является видом «собственно-альтернативного» значения.

ваших финансов и степени сытости, не мешает вам ответить на него: «И то, и другое» или «Ни то, ни другое». Ср. (17), где говорящий допускает осуществление обеих альтернатив.

- (17) Одеты она в старое ситцевое платье; руки у ней не то загорели, не то загубели от работы, от огня *или* от воды, **или от того и от другого**. [И. А. Гончаров. Обломов (1848–1859)]

Представление об альтернативе в естественном языке вообще имеет мало общего с логической связкой «V» (или «VV»), а основано на понятиях «выбора» между членами отношения альтернативы, который говорящий предлагает сделать [Богуславский 1996: 124–126], и «противопоставления» между ними, которое может быть более или менее сильным [Инькова-Манзотти 2001: 253–256]: оно минимально при т. н. перечислительном употреблении *или*, как в (4) выше, и максимально при двухкомпонентном повторяющемся *или*, как в (18).

- (18) «Нужно было *или* забыть его, *или* самой умереть». [М. А. Булгаков. Мастер и Маргарита (1929–1940)]

Однако чем длиннее цепочка повторяющихся *или*, тем слабее взаимоисключающий характер противопоставления. Не случайно французский переводчик примера (6) выше выбирает в качестве эквивалента РР *или||или||или||или||или* одноэлементное *ou* 'или', вводящее последний член ряда. Ср. также (19), где *наконец* перед последним *или*, а также вставочное предложение, следующее за ним, говорят о том, что мы имеем дело с перечислением равновозможных альтернатив, а не с их взаимоисключением:

- (19) Идея же была для него *или* пробным камнем для испытания человека в человеке, *или* формой его обнаружения, *или*, *наконец*, — это главное — тем «медиумом», тою средою, в которой раскрывается человеческое сознание в своей глубочайшей сущности. [М. М. Бахтин. Проблемы поэтики Достоевского (1929)]

Таблица 3. Распределение употреблений РР с *или* по ЛСО в НБД

	ру-фр	фр-ру	ру-ит	ит-ру
пропозициональная альтернатива	438 (86,4%)	195 (87,0%)	235 (79,1%)	100 (90,9%)
метаязыковая альтернатива	6 (1,2%)	12 (5,4%)	50 (16,8%)	5 (4,5%)
иллокутивная альтернатива	1 (0,2%)	1 (0,4%)	2 (0,7%)	0 (0,0%)
коррекция	4 (0,8%)	1 (0,4%)	0 (0,0%)	0 (0,0%)
замещение	54 (10,6%)	14 (6,2%)	9 (3,0%)	5 (4,5%)
отрицательная альтернатива	4 (0,8%)	1 (0,4%)	1 (0,3%)	0 (0,0%)
Всего	507	224	297	110

Объем статьи не позволяет нам подробно остановиться на семантике ЛСО альтернативы, а также на его разновидностях. Приведем лишь распределение аннотаций по трем уровням, на которых оно может быть установлено, а также состав РР, способных употребляться на каждом из них (пп. 3.2.1–3.2.2). Этот вопрос остается в русистике малоизученным, а количественных данных, освещающих этот аспект формальной вариативности коннекторов на сегодняшний день не существует. В [РГ-80: II, 627–628], например, упоминаются лишь три «лексических конкретизатора», которые сочетаются с союзом *или* в «несобственно разделительном» значении: *иначе говоря, вернее, точнее*. А «Словарь служебных частей речи» [Ефремова 2004], как правило, очень широко отражающий формальное варьирование коннекторов, не дает ни одного сочетания союза *или* с конкретизаторами.

2.3.1. ЛСО альтернативы

Наиболее частотное употребление РР с *или* — выражение ЛСО альтернативы на пропозициональном уровне (не менее 80%). Зафиксирована 71 РР с *или*, из них наиболее частотные (учитываются все направления переводов): *или* — 722 аннотации, *или||или* — 35, *ли||или* — 26, *или же* — 19, $\emptyset||или||или$ ⁷ — 18, *или даже* — 16. Частотны РР, включающие *хотя бы* и *просто*, которые могут сочетаться друг с другом: *или просто; или хотя бы; или хотя бы просто; или же просто, или же, может быть, просто*, а также РР, включающие в свой состав *напротив* и *наоборот*: *или, напротив; или, наоборот; или||или же, наоборот*, и модальный элемент *может (быть)*: *а, может быть||или; или, может; или, может быть; или, может быть||только; иль, может быть; может быть||или; может быть||или же; может||или; быть может||или*.

Вторым по частотности употреблением РР с *или* является метаязыковая альтернатива, связывающая эквивалентные наименования одного положения вещей или его составляющей (объекта, признака и т. д.). Этот вид альтернативы встречается преимущественно в научных текстах с поясняющей функцией. РР менее разнообразны, чем при пропозициональной альтернативе: 69 аннотаций из 73 приходятся на *или*; 4 аннотации — на *или* в сочетании с «конкретизаторами», в состав которых входят глаголы речи: *или, скажем иначе / так сказать / иначе говоря / говоря* + наречие. Редкость конкретизаторов можно объяснить тем, что говорящий, как правило, комментирует другими способами различия альтернативных наименований, как в (20)–(21):

- (20) <...> в порту был тифозный карантин, и вывоза, **или, как говорят по-тюремному**, «этапа», из него не было уже более месяца.
[В. Т. Шаламов. Колымские рассказы (1954–1962)]

⁷ Символ \emptyset означает, что первый компонент трехчленной альтернативной конструкции не маркирован; ср. пример (18) выше, где маркирован каждый фрагмент, и *Использовать \emptyset какой-то план, **или** обстоятельства, **или** условия, которые нам были известны и до нашей операции*. [А. и Г. Вайнеры. Эра милосердия (1975)]. Эти РР необходимо различать, поскольку они имеют разную семантику.

(21) Она есть его тень *или*, по Гегелю, диалектический антитезис.
[С. Н. Булгаков. У стен Херсониса (1922)]

Наименее частотным является ЛСО альтернативы, установленное на уровне высказывания; см. пример (9) выше. Этот вид альтернативы встречается, как правило, в разговорной речи, поэтому в нашем корпусе, учитывая характер загруженных в него текстов, на него приходится в среднем 0,5%. Зафиксированы две РР: *или* и *или|может быть*.

Наконец, РР с *или* могут выражать ЛСО отрицательной альтернативы, как в его разновидности «альтернативной мотивации» [РГ-80, II: 627], как в (16), так и — при поддержке *в противном случае* — для введения положения вещей, которое будет иметь место при неосуществлении положения вещей, описанного в первом фрагменте текста, как в (22):

(22) Буде же я, Обломов, пожелаю прежде времени съехать с квартиры, то обязан передать ее другому лицу на тех же условиях *или*, *в противном случае*, удовлетворить ее, Пшеницыну, сполна платою за весь год, по первое июня будущего года [И. А. Гончаров. Обломов (1848–1859)]

Данные в НБД о редкости (менее 1%) употребления РР с *или* в этом значении, также характерном для разговорной *или*, по крайней мере, диалогической речи, согласуются с данными О. А. Ананьевой [Ананьева 2005: 18], полученными на материале разножанровых текстов (1,8% в картотеке из 3000 примеров), и объясняется, на наш взгляд, тем, что в русском языке существуют как специализированные нейтральные показатели данного отношения (*иначе*), так и союз *а то*, также способный выражать данное ЛСО и имеющий разговорную окраску (в НБД из 177 аннотаций с альтернативной мотивацией в направлении русский → французский на него приходится 110).

2.3.2. Другие ЛСО, выражаемые РР с *или*

В рассмотренных выше случаях элементы, входящие в состав РР с *или*, можно квалифицировать как «конкретизаторы», которые уточняют значение союза *или* (*наконец* при перечислении), эксплицируют заложенную в нем сему потенциальности (*может быть*) или противопоставления (*напротив, наоборот*), вносят дополнительный оттенок градации (*даже*) и т. д. Их можно опустить без существенного изменения смысла высказывания. В других РР союз *или*, наоборот, является лишь синтаксической связкой, совместимой со значением основного показателя ЛСО, а именно — ЛСО замещения, при котором нарушается эквивалентность дизъюнктов: говорящий уже осуществил выбор между ними и замещает одно положение вещей *или* одно описание положения вещей другим (подробнее см. Инькова-Манзотти 2001; гл. VI). При этом, если опущение *или* часто невозможно по синтаксическим причинам, то опущение показателя замещения либо превращает замещение в альтернативу, как в (23), либо делает высказывание неприемлемым, как в (24).

(23) Он смотрел на это, как на естественное, иначе быть не могущее дело, *или, лучше сказать*, никак не смотрел, а поступал так, без всяких умозрений. [И. А. Гончаров. Обломов (1848–1859)]

(24) Основное, что определяло его лицо, это было, пожалуй, выражение добродушия, которое нарушали, впрочем, глаза, *или, вернее*, не глаза, а манера пришедшего глядеть на собеседника. [М. А. Булгаков. Мастер и Маргарита (1929–1940)]

Среди РР лидируют *или*, по крайней мере — 24 аннотации, *или, вернее* — 10, *или, точнее* — 9.

Наконец, ЛСО коррекции зафиксировано только в сочетании одноэлементного *или* с отрицанием *Нет!*, аннулирующем сказанное как несоответствующее действительности. Ср. пример (25), где первое *или*, благодаря *нет* воспринимается как корректирующее, а второе — как обычная альтернатива.

(25) Нет, *лучше* послезавтра... *или* нет, в пятницу *или* субботу, — отвечал он. [И. А. Гончаров. Обломов (1848–1859)]

3. Выводы

Проведенный с помощью НБД коннекторов анализ РР с *или* позволяет сформулировать следующие выводы и подтвердить их количественными данными.

- Занимая всегда начальную позицию во вводимом им фрагменте текста, *или* употребляется в основном как коннектор (менее 2,0% лексикализованных употреблений), сближаясь в этом отношении с *но* и отличаясь от *а* или *и*; см. [Подлеская 2019], [Зализняк, Микаэлян 2018].
- Наиболее характерным порядком следования фрагментов текста для РР с *или* является р CNT q (86,2%), тогда как, например, для *но* доля такой схемы в НБД — 66,0%.
- По своим синтаксическим характеристикам *или* прямо противоположен *и*, употребляясь в большинстве случаев для непредикативного сочинения (66,0%).
- РР с *или* выражают ЛСО альтернативы на всех трех семантических уровнях, а также ЛСО замещения, коррекции и отрицательной альтернативы, с явным преобладанием пропозициональной альтернативы (более 80,0%).
- Союз *или* способен образовывать многоэлементные, двух- и многокомпонентные РР, состав которых варьируется в зависимости от выражаемого ЛСО; наиболее частотной РР является одноэлементное *или* (70,1%); для сравнения: доля одноэлементной РР *но* составляет в НБД 50,9% (802 аннотации из 1575 РР с *но*).
- Тот факт, что многие неоднословные РР с *или* зафиксированы в НБД с единичными вхождениями, подтверждает гипотезу, согласно которой РР является единицей речи, а не языка, и часто создается говорящим *ex novo* с учетом его коммуникативного задания [Инькова 2018a].

Таким образом, информация, полученная в НБД в результате количественного многопараметрового анализа союза *или*, с одной стороны, подтвердила

лексикографические описания и высказанные гипотезы о его функционировании (в частности, его употребление исключительно в функции коннектора, его начальную позицию во вводимом им фрагменте текста, употребление преимущественно для связи частей предложения), а с другой — позволила получить новые данные, характеризующие особенности его функционирования: незначительное количество лексикализованных употреблений, преимущественное функционирование с порядком следования компонентов р CNT q для связи не-предикативных единиц, возможность устанавливать отношение альтернативы на всех трех семантических уровнях, а также выражать некоторые другие ЛСО, причем для выражения каждого из них и разновидностей ЛСО альтернативы одноэлементный союз *или* сочетается с определенным набором конкретизаторов, образуя многоэлементные и двух- и многокомпонентные РР.

Дальнейший количественный анализ аннотаций РР с сочинительными союзами позволит установить, какие функциональные свойства являются общими для всех сочинительных союзов и в какой мере они характерны для каждого из них. Качественный анализ состава РР позволит, в свою очередь, установить, каковы сочетаемостные возможности сочинительных союзов; ср., например, разную степень приемлемости и частотности сочетаний *и следовательно*, *а следовательно*, *но следовательно* (единичные употребления в НКРЯ) и **или следовательно*.

Литература

1. Anan'eva O. A. (2005) Polypredicative disjunctive sentences with a conjunction *ili* (or): structure and semantics [Polipredikativnye razdelitel'nye predlozhenija s sojuzom ili: stroenie i semantika]. PhD, Samara State Pedagogical University, Samara. 222 p.
2. Boguslavsky I. A. (1996) Scope of lexical units [Sfera deistviya leksicheskikh edinit], Yazyki slavyanskikh kul'tur, Moscow. 468 p.
3. De Cornulier B. (1985) Meaning effects [Effets de sens], Les éditions de Minuit, Paris. 212 p.
4. Efremova T. F. (2004), Explanatory dictionary of the Russian auxiliary parts of speech [Tolkovyj slovar' sluzhebnykh chastei rechi russkogo yazyka], Moscow, Astrel'-Ast. 815 p.
5. Gamut L. T.F. (1991) Logic, Language, and Meaning. Vol. I: Introduction to Logic, The University of Chicago Press Ltd., Chicago and London, 1991. 296 p.
6. Gladkii A. V. (1979) On the meaning of conjunction *ili* (or) [O znachenii soyuza *ili*], Semiotics and informatics [Semiotika i informatika], Vol. 13, pp. 196–214.
7. Inkova O. Yu. (2018a) Supracorpora database as an instrument of the study of the formal variability of connectives [Nadkorpornaja baza dannykh kak instrument izutcheniya formal'noj variativnosti konnektorov]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018” [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2018”]. Moscow, May 30–June 2, 2018. <http://www.dialog-21.ru/media/4299/inkovaoyu.pdf>.

8. *Inkova O. Yu.* (20186) The language-specificity of connectives: methods and parameters of description [Lingvospetsifitchnost' konnektorov: metody i parametry opisanija], Semantics of connectives: a contrastive study [Semantika konnektorov: kontrastivnoe issledovanie], O. Inkova (ed.), TORUS PRESS, Moscow, pp. 5–23.
9. *Inkova O. Yu.* (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], Izdatel'skii Dom YaSK, Moscow, pp. 11–98.
10. *Inkova O. Yu.* (2020) System of conjunctive tools in Russian: structure and description [Sistema svyazuyushchikh sredstv russkogo yazyka: sostav i metody opisanija], Proceeding of the V. V. Vinogradov Russian Language Institute, Vol. XXII: Materials of International Scientific Conference «Grammatical Processes and Systems in Synchrony and in Diachrony» (May 27–29, 2019) [Trudy Instituta russkogo yazyka im. V. V. Vinogradova. Vyp. 22. Materialy mezhdunarodnoi nauchnoi konferentsii «Grammaticheskie protsessy i sistemy v sinkhronii i diakhronii» (27–29 maya 2019)], Moscow, pp. 91–103.
11. *Inkova-Manzotti O. Yu.* (2001) Connectors of opposition in French and Russian: A comparative study [Konnektory protivopostavleniya vo frantsuzskom i russkom yazykakh: Sopostavitel'noe issledovanie], Informelektro, Moscow. 434 p.
12. *Inkova O., Popkova N.* (2017). Statistical data as information source for linguistic analysis of Russian connectors, Informatics and applications [Informatika i ee primeneniya], Vol. 11, No. 3, pp. 123–131.
13. *Kondakov N. I.* (1975) Logical reference dictionary [Logicheskii slovar'-spravochnik], Nauka, Moscow. 721p.
14. *MAS* — Dictionary of the Russian language: In 4 vol. [Slovar' russkogo yazyka], A. P. Evgen'eva (ed.), Russkii Yazyk, Moscow, 1981.
15. *Paducheva E. V.* (1964) Attempt of the logical analysis of the meaning of the conjunction ILI (OR) [Opyt logicheskogo analiza znacheniya soyuza ILI], Scientific reports of higher school, Philology [Nauchnye doklady vysshei shkoly, Filologicheskie nauki], No 6, pp. 145–148.
16. *Peretrukhin V. N.* (1979) Problems of syntax of equally functional sentence members in modern Russian [Problemy sintaksisa odnorodnykh chlenov predlozheniya v sovremennom russkom yazyke], Izdatel'stvo Voronezhskogo universiteta, Voronezh. 209 p.
17. *Podlesskaya V. I.* (2019) Prosody and grammar of clausal and vp coordination: the Russian conjunction *i* (*and*) viewed through the prism of prosodically annotated corpus data [Prosodiya i grammatika predikativnogo sochineniya: konstruktsii s soyuzom i po dannym prosodicheski razmechennogo korpusa], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2019"]. Moscow, May 29—June 1, 2019, <http://www.dialog-21.ru/media/4620/podlesskayavi-017.pdf>.
18. *RG-80* — Russian grammar [Russkaya grammatika], Shvedova N. Yu. (ed.), Moscow, Nauka, 1980.

19. *Sannikov V. Z.* (1985) Semantics and pragmatics of the conjunction *ili* (or) [Semantika i pragmatika soyuza *ili*], Semiotics and informatics [Semiotika i informatika], Vol. 24, pp. 117–141.
20. *Sannikov V. Z.* (1989) Russian coordinative structures. Semantics. Pragmatics. Syntax [Russkie sochinitel'nye konstruktsii. Semantika. Pragmatika. Sintaksis], Nauka, Moscow. 266 p.
21. *Uryson E. V.* (2004) Dictionary item “*Ili* (... *ili*), *libo* (...*libo*), *ne to... ne to*, to *li... to li* (Either (... or), or (... or), not ... not, either ..., or)” [Slovarnaya stat'ya «*Ili* (... *ili*), *libo* (...*libo*), *ne to... ne to*, to *li... to li*»], New Explanatory dictionary of synonyms of the Russian language [Novyi ob'yasnitel'nyi slovar' sinonimov russkogo yazyka], 2nd ed., rev. and add., Yazyki slavyanskikh kul'tur, Vienna Slavic almanac, Moscow — Vienna, pp. 437–440.
22. *Uryson E. V.* (2010) Compound conjunctions *a to* and *a ne to*: possibilities of semantic compositional analysis [Sostavnye soyuzy *a to* i *a ne to*: vozmozhnosti semanticheskogo kompozitsional'nogo analiza], Topics in the study of language [Voprosy yazykoznaniya], No 1, pp. 61–73.
23. *Uryson E. V.* (2013) Conjunctions and connectives syntax and valency theory [Sintaksis soyuzov i konnektorov i teoriya valentnostei], Topics in the study of language [Voprosy yazykoznaniya], No 3, pp. 3–24.
24. *Zaliznyak A. A., Mikaelyan I. L.* (2018) La conjunction *a* [Soyuz *a*], Semantics of connectives: a contrastive study [Semantika konnektorov: kontrastivnoe issledovanie], O. Inkova (ed.), TORUS PRESS, Moscow, pp. 24–79.

AN EXPERIMENTAL RULE-BASED PARSER FOR RUSSIAN EMPLOYING THE NLP RESOURCES OF THE ETAP SYSTEM¹

Inshakova E. S. (e.s.inshakova@gmail.com)

Laboratory of Computational Linguistics, A. A. Kharkevich
Institute for Information Transmission Problems;
Institute of Linguistics RAS, Moscow, Russia

Sizov V. G. (victor.sizov@gmail.com)

Laboratory of Computational Linguistics, A. A. Kharkevich
Institute for Information Transmission Problems, Moscow, Russia

This paper presents a rule-based dependency parser for Russian based on bottom-up approach. Its rules are partially rewritten ETAP syntagms, organized into groups that constitute a single pipeline. We demonstrate that such an organization enhances the performance of our parser relative to the ETAP system's and enables it to successfully process long phrases (more specifically, heavy nominal and prepositional phrases at the current experimental stage of our work).

Keywords: rule-based parser, dependency grammar, natural language processing

DOI: 10.28995/2075-7182-2020-19-387-399

ЭКСПЕРИМЕНТАЛЬНЫЙ ПРАВИЛОВЫЙ ПАРСЕР РУССКОГО ЯЗЫКА НА МАТЕРИАЛЕ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ СИСТЕМЫ ЭТАП

Иншакова Е. С. (e.s.inshakova@gmail.com)

Лаборатория компьютерной лингвистики,
Институт проблем передачи информации
им. А. А. Харкевича РАН;
Институт языкознания РАН, Москва, Россия

Сизов В. Г. (victor.sizov@gmail.com)

Лаборатория компьютерной лингвистики,
Институт проблем передачи информации
им. А. А. Харкевича РАН, Москва, Россия

¹ This paper was supported by the RSF grant No. 16-18-10422-P.

1. Introduction

In this paper, we present a pilot version of a Russian language parser that uses the parsing rules of the ETAP linguistic processor [Boguslavsky et al. 2008], [Iomdin et al. 2012] as its material. It creates similar dependency structures with the same tree link labels (syntactic relations), but is based on crucially different principles.

ETAP is one of the oldest existing knowledge-based NLP applications for Russian. Its Russian language processor had been being developed for over 20 years and accumulated a lot of important linguistic information, especially in Russian combinatorial dictionary (RCD). However, complex as it is, the ETAP parser is unable to meet some challenges. First of all, it cannot reliably work on long sentences because of inability to split such sentences into linguistically acceptable chunks and overgeneration of hypothetical links that causes combinatorial explosions [Iomdin et al. 2012], [Tsinman 2011]. This, actually, is sometimes true for not-so-long sentences and other types of phrases. In our paper, we show that ETAP's resources (rules and RCD) can be used to create a higher-performance parser. This is demonstrated through the example of heavy nominal and prepositional phrases that the ETAP parser cannot process.

The paper is structured as follows. In **Section 2**, we present our parser's architecture: in **subsection 2.1**, we list its major differences from the regular ETAP; in **subsection 2.2**, we describe in detail the parsing pipeline. In **Section 3**, we provide the results of evaluation of the experimental parser's performance and compare them to the ones of the current ETAP-4. In **Section 4**, we briefly discuss the advantages and drawbacks of our algorithm.

2. The parser's architecture

2.1. General principles and differences from the ETAP parser

1. *Tree-building technique.* The regular ETAP parser is neither bottom-up nor top-down. Its binary rules that connect head and dependent words (syntagms) first construct all the possible hypothetical links, and then other rules filter them [Iomdin et al. 2012]. Our parser uses the bottom-up technique to build dependency trees. As opposed to other relatively new rule-based parsers of Russian [Anisimovich et al. 2012], [Antonova and Misyurev 2012], [Boyarsky and Kanevsky 2015], [Moskvina et al. 2016] and other languages [Gamallo 2015], [Korzeniowski and Mazurkiewicz 2017], it consistently applies the linguistic principle of constituents' hierarchy. It starts from parsing subtrees that correspond to the smallest constituents, and then at each new stage proceeds to larger ones.
2. *Dictionary vs. grammar.* The regular ETAP is dictionary-oriented. It heavily relies on the words' features from the RCD entries, e.g. those that encode specific constructions a given lexeme can participate in. Such syntactic information as word order, projectivity and punctuation marks, though also very important for ETAP, often cannot be made the most of. This is partially because the primary matrix,

or network, of hypothetical links is often too noisy to check these parameters, partially because word/constituent order is not the main concern of the Meaning—Text theory that underlies the ETAP system. The filtering rules (INTER-SYNT) generally favour linear distance restrictions and discard long-distance links [Tsinman 2011], but the problem is that the linearly closest words are not necessarily the structurally closest ones.

On the contrary, the architecture of our parser that simulates bottom-up constituency tree derivation makes it more grammar-oriented. It parallels dependency (sub)trees to constituents, and relies on their order, projectivity and punctuation when merging those subtrees together.

3. *Ordering of the rules.* In the regular ETAP, the order syntagms are applied in is irrelevant. In our parser, all syntagms are strictly ordered. Every next rule makes use of the information about links built by previous rules.
4. *The continuity condition.* In our parser, each rule, while linking two words, presupposes that i) these words are heads of subtrees that are already built (a minimal subtree is defined as a single node); ii) those subtrees are contiguous, i.e. for each node x between the subtrees A (with head a) and B (with head b) it is true that at least one of the homonyms of x (variants of x 's morphological analysis) is connected via a dependency chain to either a or b . However, this condition must be weakened for parenthetical expressions between the heads a and b , because they can contain such syntactic strictures that haven't yet been parsed at a given stage (e.g. if a is an adjective, b is a noun, and a parenthetical expression between a and b contains a prepositional phrase—see the parsing pipeline in [subsection 2.2](#)). To be able to connect a and b in such cases, we provide that words that belong to parenthetical groups between a and b need not be connected to a or b via a dependency chain. We should also add to our definition that the heads a and b must belong to the same parenthetical group or to no parenthetical groups at all, except for one case. It is the JUXTAPOSE relation between a head word and the head of a clarifying expression in brackets (in this case, a and b belong to different parenthetical groups). The continuity condition with all these exceptions is implemented in a new binary predicate IS-CONTINUOUS(X,Y) of ETAP's formal language FORET.

As can be seen, the principle of continuity is based on the idea of projectivity. However, it can be also applied to non-projective strictures:

- i) to a displaced subtree A and its right neighbor B , as in (1);
 - ii) to parts A and B of comparative constructions as they are presented in ETAP (2).
- (1) [Kakuyu knigu]_A [vash prepodavatel']_B khotel, chtoby vy prochitali?
 '[What book]_A <did> [your teacher]_B want you to read?'
- (2) takie [oblasti znaniya]_A, [kak]_B istoriya filosofii
 'such [areas of knowledge]_A [as]_B history of philosophy' [in ETAP's syntax, the head *takie* 'such' governs the conjunction *kak* 'as']



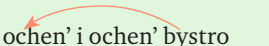
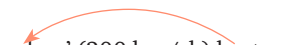

2.2. The process of parsing

2.2.1. Stages of parsing

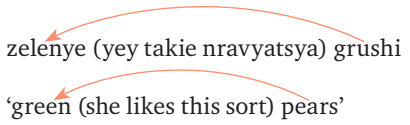
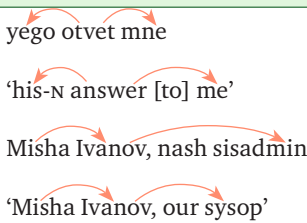
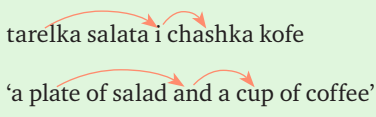
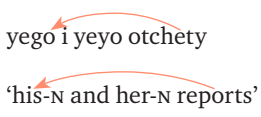
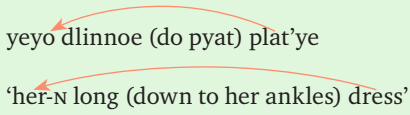
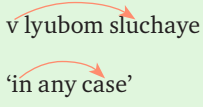
In our parser, rules are grouped into classes that correspond to the categories of constituents (= subtrees) they create, and are mostly named after these categories: Abbr[eviation], Add[itive], AdvP, AP, DP [determiner phrase], NP, PP, TP [tense phrase], TPFIn[ite], TSub[ordinate]. These big classes of rules are divided into subclasses: ∅ ‘bare’, -Attach, -Coord, -Postcoord, -With-Br[ackets] and -With-Par[enthesi]. Below we present the sequence of stages the experimental parser works in, and explain the functions of the groups of rules. Because our parser builds only NP/PPs so far, we will not give here the outline of full-sentence parsing.

Stage of parsing		Its output	Examples ¹
ABBR	ABBR	Linking parts of composite words	neftenepronitsaemyj ‘oil-tight’
	ABBR-COORD	Coordination links between parts of composite words	nefte- i gazonepronitsaemyj ‘oil- and gas-tight’
ADDP	ADDP	Linking parts of compound numerals, or parts of many idiomatic expressions	pyat’desyat tri ‘fifty-three’ drug druga ‘each other’
	ADDP-COORD	Linking numerals into coordinate chains	sorok dva ili tridtsat’ vosem’ ‘fourty-two or thirty-eight’

¹ In the examples, we show only the links that are built at the given stage of parsing, and do not show the links built at the preceding stages.

Stage of parsing	Its output	Examples ¹
ADVP	Building up subtrees headed by Adv (adverbial phrases)	
ADVP	Linking linearly closest dependent words to Adv	 ne ochen' bystro 'not very quickly'
ADVP-COORD	Linking AdvP's into coordinate and comparative constructions	 bystro ili medlenno 'quickly or slowly'
ADVP-POSTCOORD	Building links that bypass the ready AdvP coordinate chains	 ochen' i ochen' bystro lit. 'very and very quickly'
ADVP-WITH-BR	Creating long-distance links where the host (Adv) and the modifier are separated by a parenthetical expression of any kind	 ochen' (200 km/ch) bystro lit. 'very (200 km/h) quickly'
AP	Building up subtrees headed by Adj (adjective phrases)	
AP	Linking linearly closest dependent words to Adj	 ne ochen' bystryj 'not very quick'

Stage of parsing		Its output	Examples ¹
AP	AP-COORD	Linking AP's into coordinate and comparative constructions	bystryj ili medlennyj 'quick or slow'
	AP-POSTCOORD	Building links that bypass the ready AP coordinate chains	samyj ili ne samyj bystryj lit. 'the most or not the most quick'
	AP-WITH-BR	Creating long-distance links where the host (Adj) and the modifier are separated by a parenthetical expression of any kind	naskol'ko široko (u nas) izvestnyj lit. 'how widely (over here) known'
DP	Building up subtrees headed by N, without complements (≈ determiner phrases)		
	DP	Linking linearly closest dependent words to N	bol'shaya gonochnaya mashina 'big racing car' dve ili tri mashiny 'two or three cars'
	DP-COORD	Linking DP's into coordinate and comparative constructions	tvoya mashina ili moy velosiped 'your car or my bicycle'
	DP-POSTCOORD	Building links that use or bypass the ready DP coordinate chains	zelenye shapka i sharf 'green-PL hat and scarf'

Stage of parsing		Its output	Examples ¹
DP	DP-WITH-BR	Creating long-distance links where the host (N) and the modifier are separated by a parenthetical expression of any kind	 <p>zelenye (yey takie nrayvatsya) grushi 'green (she likes this sort) pears'</p>
NP	Building up subtrees headed by N, with complements and nominal modifiers (nominal phrases)		
	NP1-4	Attaching complements and nominal modifiers to N's	 <p>yego otvet mne 'his-N answer [to] me' Misha Ivanov, nash sisadmin 'Misha Ivanov, our sysop'</p>
	NP-COORD	Linking NP's into coordinate and comparative constructions	 <p>tarelka salata i chashka kofe 'a plate of salad and a cup of coffee'</p>
	NP-POSTCOORD	Building links that bypass the ready DP coordinate chains	 <p>yego i yeyo otchetu 'his-N and her-N reports'</p>
	NP-WITH-BR	Creating long-distance links where the host (N) and the modifier are separated by a parenthetical expression of any kind	 <p>yeyo dlinnoe (do pyat) plat'ye 'her-N long (down to her ankles) dress'</p>
PP	Building up subtrees headed by Prep (prepositional phrases)		
	PP	Building links from prepositions to N/A/Num's	 <p>v lyubom sluchaye 'in any case'</p>

Stage of parsing		Its output	Examples ¹
PP	PP-WITH-BR0	Building links from prepositions that bypass parenthetical expressions	nad (ili pod) chastnoy zemley 'above (or under) private land'
	PP-ATTACH1-5	Attaching prepositional complements and modifiers to N/A/Adv/Num/Pr	yego uverenost' v pobede nad sopernikom 'his confidence in his win against his rival' v chastnosti o probleme vybora 'in particular about the problem of choice'
	PP-COORD	Linking PP's into coordinate and comparative constructions	v devyat' utra ili v sem' vechera 'at nine a.m. or at seven p.m.' temno, kak v pogrebe 'as dark as in a vault'
	PP-POSTCOORD	Building links that bypass the already attached PP's	interesnye dlya nas predlozheniya 'offers interesting for us' (lit. 'interesting for us offers')
	PP-WITH-BR1-5	Creating long-distance PP-internal links where the host and the modifier are separated by a parenthetical expression of any kind	vstrecha s nim (chtoby obsudit' eto) v chetyre chasa 'meeting with him (to discuss this) at four o'clock'
	PP-WITH-PAR	Drawing links to parenthetical expressions	na dva (ili na tri) chasa ran'she 'two (or three) hours earlier'

2.2.2. Attaching multiple dependents

The above-mentioned rules, like in the regular ETAP, are binary: they connect two subtrees into a bigger (sub)tree. To enable the parser to process such groups where a head word has multiple dependent words, we subdivided the NP, PP-Attach, TP, TPFin stages into NP1-4, PP-Attach1-5, TP1-5, TPFin1-5 respectively. The result of such a subdivision partially resembles that of slot-filling parsing grammars [McCord et al. 2012], [Anisimovich et al. 2012], where each lexeme or word class has a fixed number of slots for complements, adjuncts etc. around the head word. But our way of linking multiple dependents to heads can be rather called ‘layer-by-layer’ attachment. Each ‘layer’ is such a subtree (or a pair of subtrees) to the right and/or left of the head that constitutes a projective tree with this head. The closer a layer is to the core word, the earlier it is added.

There are basically two ways of layering dependents around the head word (an N in our experiment). The first one is creating recursive rules that attach adjectives or particles to N (from the 1st to the nth one):

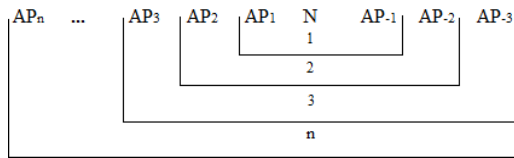


Fig. 1. ‘Layers’ of AP’s around a nominal head

These rules apply to the same head N until the most distant AP/particle is attached to it.

The second way is multiplying and ordering the rules that attach dependent NP’s, PP’s, AdvP’s and nominalized AP/NumP’s to core words. The number of such copied rules (up to 4 so far) mirrors the maximal number of ‘layers’ around the core N.

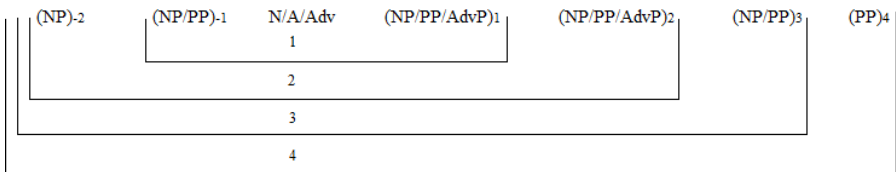


Fig. 2. The maximal number of ‘layers’ of dependent subtrees around a nominal head

Because dependents of the two types can intersperse, we multiplied the rules of the first type: they first apply to a bare head word, and then to the same word after its arguments have been attached. There is also another cause to multiply rules: we need links that ‘circumvent’ different types of subtrees. For instance, the syntagm `MODIF.01` that connects adjectives to head N’s should be repeated at DP, NP, and PP stages to ‘circumvent’ first AP’s, then DP’s, NP’s and finally PP’s (*eto ves’ma interesnoe predlozhenie* ‘this very interesting offer’; *eto yego predlozhenie* ‘this offer of his’, lit. ‘this his-N offer’; *interesnoe dlya nas predlozhenie* ‘an offer interesting for us’, lit. ‘interesting for us offer’; *nashe s Petey predlozhenie* ‘me and Petya’s offer’, lit. ‘our with Petya offer’, etc.).

3. Evaluation of the parser’s performance

At the current stage of our parser’s development, we restricted ourselves to constructions that are characterized by:

- POS repertory of the nodes: N, A, Adv, Particle, Conj, Prep;
- syntactic relations between the nodes: 1–5-COMPL (argument relations); QUASIAGENT (a relation between an action nominal and its genitive argument that denotes an agent or possessor); ATTRIB (it links head N’s with non-argument nominal and prepositional modifiers); MODIF (a relation between N’s and their adjectival modifiers); PREPOS (a relation between a preposition and a word that it governs); RESTR (it links restrictive particles or PP’s to words of any POS); LOCUT (a relation between parts of highly idiomatic expressions); COORDIN (a relation between heads of two coordinate phrases or between the head of a conjunct and a coordinate conjunction); COORD-CONJ (a relation between a coordinate conjunction and head of the last conjunct); COMPAR (a relation between a head word and a comparative conjunction or a noun in the Genitive of comparison); COMP-CONJ (a relation between a comparative conjunction and the head of the phrase denoting the standard of comparison); JUXTAPOSE (a relation between a head word and the head of a clarifying expression in brackets)².
- length: NP’s and PP’s can be as heavy as possible (in our corpus, the maximal length is 45 words).

We ran two evaluation experiments. In the first one, we used a very limited corpus that was 100 phrases long (the average phrase length was 16.6 words). The selection principle was the regular ETAP’s inability to correctly parse a given NP or PP. We selected almost ½ of the phrases (49) on the basis of a certain feature—long-distance links that ‘circumvent’ parenthetical expressions in brackets and present a challenge for the regular ETAP parser. Other sources of errors (in 62 sentences) are very diverse: e.g. unknown words, long phrase length, incorrect PP-attachment or conjunct attachment, very long linear distance between the head and the dependent word, and constructions undescribed in ETAP rules. The phrases were taken from the Russian National Corpus (some of them had been originally sentences that we nominalized).

² See the full tagset of the ETAP parser in <http://ruscorpora.ru/new/instruction-syntax.html>.

Table 1. Results of the 1st evaluation experiment (100 phrases)

Parameters	The regular ETAP parser	The experimental parser
Number/percentage of correct heads (UAS)	1,377 (82%)	1,528 (92%)
Number/percentage of correct dependency labels	1,427 (85%)	1,568 (94%)
Number/percentage of correct heads and dependency labels (LAS)	1,327 (79%)	1,498 (90%)
Number/percentage of correct syntactic structures (without dependency labels)	3 (3%)	37 (37%)
Number/percentage of correct syntactic structures (with dependency labels)	0	29 (29%)
Processing time	64 sec.	41 sec.

For the second evaluation experiment, we extracted from the SynTagRus tree-bank all subtrees: 1) headed by N or Prep, 2) featuring only the POS and syntactic relations from the list given above, 3) being ≥ 10 words long (2,838 phrases in total). 2,500 phrases comprised the testing corpus, and 338 were used as a development corpus. The evaluation results for our parser and the regular ETAP are presented in [Table 2](#).

Table 2. Results of the 2nd evaluation experiment (2,500 phrases)

Parameters	The regular ETAP	The experimental parser
UAS	91.21	93.45
LAS	88.68	90.94
Number/percentage of correct structures (without dependency labels)	1185 (47.40)	1354 (54.16)
Number/percentage of correct structures (with dependency labels)	922 (36.88)	1048 (41.92)
Processing time	948 sec.	738 sec.

4. Conclusion

The evaluation results show that at the current stage our experimental parser outperforms the regular ETAP parser. However, as can be seen from the difference between its performance in the 1st and in the 2nd experiments, our parser is still somehow ‘tuned’ to cope with a certain type of errors, which is frequent in the 1st testing corpus and infrequent in the 2nd one. Besides that, the experimental parser faces some other problems. Firstly, its pipeline architecture is fraught with the risk of accumulation of errors with every step; secondly, the multiplication of rules (by approximately

10 times for NP's / PP's so far) can potentially increase the processing time of whole sentences. Thirdly, some issues listed in [Iomdin et al. 2012] still hold for the experimental parser, e.g. it cannot yet correctly parse structures with gapping, phrases with orthographical errors and other kinds of non-standard spelling.

One may also note that the regular ETAP and our parser often make different types of errors. For instance, if ETAP fails to find a correct host for some word, it is likely to attach this word to some of the linear closest ones (e.g. via the ATTRIB relation), whereas our parser attaches 'hostless' words to the root of the phrase. Consequently, if the experimental parser is given a whole sentence, it will pick out all the NP/PP/AP's and correctly parse them (this capability makes it possible to use it as an NP/PP extractor).

References

1. *Anisimovich K. V. et al.* (2012) Syntactic and Semantic Parser Based on ABBYY Comprendo Linguistic Technologies. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue 11 (18). P. 90–103.
2. *Antonova A. A., Misyurev A. V.* (2012) Russian dependency parser SyntAutom at the Dialogue-2012 parser evaluation task. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue 11 (18). P. 104–118.
3. *Boguslavsky I. M. et al.* (2008) Parser of the ETAP system and its evaluation with the aid of a deeply annotated corpus of Russian texts [Sintaksicheskiy analizator sistemy ETAP i yego otsenka s pomoshchyu gluboko razmechennogo korpusa russkikh tekstov]. In: Proceedings of the international conference 'Corpus linguistics—2008' (St. Petersburg). P. 56–74.
4. *Boyarisky K., Kanevsky E.* (2015) SemSin semantic and syntactic parser [Semantiko-sintaksicheskiy parser SimSin]. In: Scientific and Technical Journal of Information Technologies, Mechanics and Optics, vol. 15, no. 5, pp. 869–876.
5. *Gamallo P.* (2015) Dependency Parsing with Compression Rules. In: The 14th International Conference on Parsing Technologies (IWPT-2015). P. 107–117.
6. *Iomdin L. L. et al.* (2012) ETAP parser: state of the art. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue 11 (18). P. 119–131.
7. *Korzeniowski M., Mazurkiewicz J.* (2017) Rule Based Dependency Parser for Polish Language. In: Rutkowski L., Korytkowski M., Scherer R., Tadeusiewicz R., Zadeh L., Zurada J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2017. Lecture Notes in Computer Science, vol. 10246. P. 498–508. Springer.
8. *McCord M.C., Murdock J. W., Boguraev B. K.* (2012) Deep parsing in Watson. In: IBM Journal of Research and Development, vol. 56 3.4: IBM, pp. 3–1.
9. *Moskvina A. D. et al.* (2016) Development of the Core for Syntactic Parser for Russian based on NLTK libraries [Razrabotka yadra sintaksicheskogo analizatora dlya russkogo yazyka na osnove bibliotek NLTK]. In: Computer linguistics and computational ontologies. Proceedings of the XIX joint scientific conference.

10. *Tsinman L. L.* (2011) Ranking syntactic hypotheses in the syntactic parser of ETAP-3 linguistic processor [Ranzhirovanie sintaksicheskikh gipotez v sintaksicheskom analizatore lingvisticheskogo protsessora ETAP-3]. In: *Word and Language [Slovo i yazyk]*. Festschrift for Yu. D. Apresyan's 80th anniversary. Moscow, 2011. P. 573–587.

ВАЛЕНТНАЯ СТРУКТУРА НЕКОТОРЫХ РЕЧЕВЫХ ПРЕДИКАТНЫХ СЛОВ: НОВЫЕ НАХОДКИ¹

Иомдин Б. Л. (iomdin@ruslang.ru),

Иомдин Л. Л. (iomdin@iitp.ru)

ИРЯ РАН им. В. В. Виноградова РАН,

ИППИ РАН им. А. А. Харкевича

В статье рассматриваются валентные рамки ряда русских глагольных предикатов, в значение которых входит речевой акт, а также, на некоторой стадии семантического разложения, отрицание — такие как *возражать*, *возмущаться*, *извиняться* и др. Высказывается предположение, что валентные рамки таких предикатов включают в себя пару пропозициональных валентностей, отчетливо противопоставленных друг другу: (1) валентность стимула, которая выражает положение дел, и (2) валентность реакции, которая вводит речевой акт, совершаемый субъектом в качестве отклика на это положение дел и предлагающий его объяснение. Например, в предложении *Иван извинился, что не пришел на мой день рождения* клауза, вводимая союзом *что*, выражает положение дел, а в предложении *Иван извинился, что плохо себя чувствовал* такая клауза передает речевую реакцию Ивана на положение дел (например, отсутствие на моем дне рождения), стимулирующее его дать объяснение этому отсутствию. Показано, что эти валентности нельзя адекватно описать в рамках единой семантической роли содержания. Авторы также предлагают обобщение этого явления, сравнивая его с другими типами валентных пар, и выдвигают гипотезу о существовании предикатов, имеющих два валентных центра.

Ключевые слова: речевые предикаты, валентности предикатных слов, валентность содержания, валентность стимула, валентность реакции, неоднозначное заполнение валентностей

DOI: 10.28995/2075-7182-2020-19-400-415

¹ Авторы благодарят РФФИ за частичное финансирование этого исследования (грант № 19-012-00291 «Подготовка четвертого выпуска Активного словаря русского языка», рук. Б. Л. Иомдин, и грант № 19-07-00842 «Разработка корпуса русских текстов, снабженного морфосинтаксической, лексико-функциональной, анафорической и темпоральной разметкой», рук. И. М. Богуславский).

VALENCY STRUCTURE OF CERTAIN PREDICATES OF SPEECH: NEW FINDINGS

Iomdin B. L. (iomdin@ruslang.ru),

Iomdin L. L. (iomdin@iitp.ru)

V. V. Vinogradov Institute of Russian Language, RAS;

A. A. Kharkevich Institute for Information Transmission Problems, RAS

The paper discusses valency frames of a number of Russian verbal predicates whose semantics includes speech acts and, at a certain step of semantic decomposition, the negation, like *vozražat'* 'object, retort', *vozmuščit'sja* 'resent, be indignant' or *izvinjat'sja* 'apologize'. It is hypothesized that the frames of such predicates include a pair of propositional valencies distinctly opposed to each other: (1) the valency of stimulus that expresses the state of events and (2) the valency of response that introduces a speech act performed by the subject as a reaction to this state of event and offering an explanation. For example, in the sentence *Ivan izvinilsja, čto ne prišel na moj den' roždenija* 'Ivan apologized that he did not come to my birthday party' the clause starting with *čto* 'that' represents the state of events, whilst in the sentence *Ivan izvinilsja, čto ploxo sebja čuvstvoval* 'Ivan apologized that he was not feeling well' the *čto*-clause introduces Ivan's response to the stimulus (e. g. of not coming to the birthday party). It is shown that these valencies cannot be adequately described with a single semantic role of content. The authors also give a generalization of this phenomenon, comparing it to other instances of valency pairs, and suggest the existence of predicates having two valency centers.

Keywords: speech predicates, valencies of predicate words, valency of content, valency of stimulus, valency of response, valency filling ambiguity

1. Постановка задачи

Предикаты, в значение которых входят речевой акт, отрицание и оценка, могут помимо валентности субъекта иметь еще несколько сложно устроенных валентностей, каждая из которых может выражаться пропозициональным актантом. Цель настоящей работы — предложить семантическое описание этих валентностей и выдвинуть гипотезу, в каких случаях они возникают.

Для начала коротко рассмотрим два аналитических толкования из словарных статей Активного словаря русского языка (2014), с которыми мы в целом согласны: это толкования глаголов *возмущаться* (автор — В. Ю. Апресян) и *возражать* (автор — О. Ю. Богуславская).

ВОЗМУЩАТЬСЯ:

A1 возмущается A2 ‘Человек A1 испытывает сильное и глубокое неприятное чувство, вызванное ситуацией A2, которую он отрицательно оценивает как противоречащую этическим нормам, причем это чувство может выражаться в критических высказываниях A3 по поводу A2’;

ВОЗРАЖАТЬ:

A1 возражает A2 на A3, что A4 ‘Человек A1 отвечает на высказывание A3 человека A2 высказыванием A4, в котором выражает свое несогласие с A3’.

Обратим внимание на следующие два обстоятельства: (1) в обоих случаях глаголы содержат в себе указание на смысловой компонент ‘высказывание’ (т.е. мы имеем дело с речевыми предикатами) и (2) оба глагола содержат по две валентности, характеризующие непредметные (пропозициональные) ситуации: для *возмущаться* это валентности A2 и A3 (критическое высказывание A3 по поводу A2), а для *возражать* — A3 и A4 (высказывание A4 в ответ на высказывание A3).

Например, в предложении

- (1) *Брюллов возмущался тем, что Пушкина не выпускают за границу* (А. Костин, Тайна болезни и смерти Пушкина)

при глаголе *возмущаться* придаточным, вводимым словом *тем*, выражена валентность A2, а в предложении

- (2) *«Ну что вы такое говорите, никакого Бога нет!» — возмутилась она* (М. Козаков, Актерская книга)

прямой речью выражена валентность A3. Аналогичным образом, в предложении (3) придаточным, вводимым союзом *чтобы*, выражена валентность A3 глагола *возражать*:

- (3) *Коллектив категорически возражал, чтобы в «Ленком» пришел человек со стороны* (М. Варшавер, Человек на своем месте),

а в (4) с помощью придаточного, вводимого союзом *что*, выражена валентность A4 этого глагола:

- (4) *Наш знаменитый дирижер возразил, что фотография передает только внешность человека, а живописный портрет отражает его внутренний мир* (А. Рыбаков, Тяжелый песок).

Иногда в предложении с такими глаголами бывают насыщены обе пропозициональные валентности. Ср.:

- (5) *Я знаю, что Христос велел всех любить, а она возмущается христианством: «Как будто можно отделить в человеке одно от другого — тело от Бога», — говорит она.* (М. Шишкин, Венерин волос).

Здесь слово в творительном падеже (*христианством*) выражает валентность A2 глагола *возмущаться*, а прямая речь выражает валентность A3 этого глагола.

В примере (6) предложная группа, вводимая предлогом *на*, заполняет валентность А3 глагола *возражать*, а однородные придаточные с союзом *что* — валентность А4:

- (6) *Марфа Игнатьевна на горькие, хотя и справедливые, упреки барина возражала, что курица и без того была уже очень старая, а что сама она в поварах не училась* (Ф. М. Достоевский, Братья Карамазовы).²

В настоящей работе мы ставим перед собой цель установить, каким семантическим ролям соответствуют приведенные выше валентности предикатных слов.

2. Валентность содержания и ее разновидности

Как представляется, во всех приведенных выше примерах отмеченные нами валентности, во всяком случае в первом приближении, предстают как валентности **содержания**. В подтверждение такой интерпретации можно привести практику Московской семантической школы; так, Ю. Д. Апресян усматривает валентность содержания (наряду с валентностью темы) у фактивных и путативных глаголов типа *знать* и *считать* [Апресян 2001]. Очевидно, что в примерах, приведенных выше, смыслы 'знать' или 'считать' входят — возможно, даже двукратно — в семантическое представление рассматриваемых глаголов *возмущаться* и *возражать*. Действительно, в (1), например, речь идет о **знании** Брюллова — субъекта возмущения, в (2) — о **мнении** субъекта возмущения, в (4) — о **мнении** того, кто возражает, а в (6) — о **знании** возражающего субъекта — Марфы Игнатьевны. Естественно, таким образом, ожидать, что глаголы *возмущаться* и *возражать* должны, по крайней мере частично, наследовать валентную структуру глаголов *знать* и/или *считать*.

Однако ограничиться утверждением, что пропозициональные валентности рассматриваемых глаголов представляют собой валентность содержания, разумеется, нельзя. Этому противоречили бы как минимум два простых обстоятельства:

1) поскольку оба глагола допускают одновременное выражение двух пропозициональных валентностей, как в примерах (5)–(6), получилось бы, что эти глаголы имеют по две одинаковые сильные валентности, что семантически антиинтуитивно³;

2) коль скоро во многих случаях обе валентности могут выражаться одинаковыми средствами, это неизбежно влечет возможность неоднозначной интерпретации высказывания, и вряд ли можно такую неоднозначность объяснить в рамках единой семантической роли. Например, предложение

² Здесь и далее большинство примеров заимствовано из разных подкорпусов НКРЯ. Если примеры сконструированы авторами, в них не дается ссылки на литературный источник.

³ Напомним в этой связи, что в лингвистической теории И. А. Мельчука «Смысл ↔ Текст» актантные синтаксические отношения (т. е. отношения, связывающие предикатное слово в качестве вершины с его актантами в качестве зависимых) считаются неповторимыми, ср. в частности, [Мельчук 1977].

- (7) *Я возражал, что украинцев 40 миллионов* (В. Н. Гельфанд, Дневники 1941–1943 гг.)

можно понять (i) так, что говорящий не соглашается с чьим-то утверждением о количестве украинцев (тогда придаточное со *что* заполняет валентность A3 глагола *возражать*), либо (ii) так, что говорящий не соглашается с некоторым неизвестным нам утверждением и в поддержку своего несогласия сообщает информацию о числе украинцев (в этом случае придаточное заполняет валентность A4 этого глагола).

Разрешение такой неоднозначности возможно либо на основании широкого контекста; в частности, предложение (7) следует понимать способом (ii), как явствует из содержания развернутого фрагмента «Дневников» В. Н. Гельфанда:

- (7а) *Спорил со мной, доказывал, что большинство украинцев с немцами, на стороне немцев... Я возражал, что украинцев 40 миллионов, и если бы даже треть нас предала, то были бы многие миллионы предателей...*,

либо на основании энциклопедических знаний; ср.

- (8) *Коперник возражал, что Земля вращается вокруг Солнца.*

Неоднозначность того же типа обнаруживается и в примерах (9) и (10):

- (9) *Президент компании возражал, что единственной причиной гибели парохода был сильный шторм у побережья Флориды* ('не соглашался, что единственной причиной был шторм' vs. 'утверждал, что единственной причиной был шторм (а не, скажем, халатность)';
- (10) *Сын возмущился, что он будет ездить в школу один* ('не соглашался с чьим-то мнением, что он будет ездить в школу один' vs. 'не соглашался с запретом ехать одному,').

В первой интерпретации имеет место насыщение валентности A2, а во второй — насыщение валентности A3. В примере

- (11) *Прочитав письмо, Иван возмущился: «Я больше не сотрудник института!»*

также можно наблюдать валентную неоднозначность ('Иван был возмущен тем, что перестал быть сотрудником института' vs. 'Иван был возмущен чьим-то мнением, что он по-прежнему является сотрудником института'). В (11), в отличие от (10), валентности A2 (в первой интерпретации) и A3 (во второй интерпретации) выражаются не придаточным с союзом *что*, а прямой речью⁴.

На наш взгляд, приведенных примеров и рассуждений достаточно, чтобы сделать вывод о неприемлемости квалификации указанных валентностей глаголов как единой валентности содержания. Какое же решение можно тут предложить?

⁴ Тот факт, что в (10) и (11) одна из интерпретаций (с помощью валентности A3) является более вероятной, чем другая, не имеет принципиального значения для наших рассуждений.

Как известно, уже довольно давно в ряде семантических и лексикографических исследований, в первую очередь в тех, что выполнены в русле Московской лингвистической школы, валентность содержания принято расщеплять на две семантические роли — собственно валентность содержания и валентность темы (ср., в частности, [Апресян 2001], [Кустова 2015], [Левонтина 2016]). Такие валентности свойственны многим предикатам сообщения, а также фактивным и путативным глаголам.

В номенклатуре семантических ролей фундаментальной классификации предикатов Ю. Д. Апресяна [Апресян и др. 2010: 376] роли собственно содержания и темы определяются следующим образом: содержание — это «то основное, к чему сводится существо нашего знания, мысли или высказывания или смысл нашего действия», а тема — «то, что составляет предмет высказывания, мысли или информационного объекта».

Тем самым тема несет гораздо менее детальную информацию о ситуации, чем содержание. Например, в предложении

(12а) *Я слышал, что Маша и Петя собираются пожениться*

при глаголе *слышать* выражена валентность содержания, а в предложении

(12б) *Я слышал о Маше и Пете*

выражена валентность темы. Часто валентность темы и содержания выражаются при одном предикате одновременно, как в предложении

(12в) *Я слышал о Маше и Пете, что они собираются пожениться.*

Еще одна особенность валентностей темы и содержания состоит в том, что они могут выражаться одинаковыми средствами, в первую очередь предложными группами, ср.

(13а) *Мы знаем о вашем намерении*

(валентность темы; какое именно намерение имеет место, мы не знаем) и

(13б) *Мы знаем о вашем намерении поступить в университет*

(валентность содержания: это предложение можно перефразировать с помощью пропозиции как *Мы знаем, что вы намереваетесь поступить в университет*).

Очевидно, что автоматический выбор интерпретации такой предложной группы при предикате представляет собой исключительно трудную задачу, едва ли имеющую сколько-нибудь общее решение.

Можно ли оппозицию валентностей содержания и темы использовать для адекватного представления валентных рамок глаголов *возмущаться* и *выражать*, которые были рассмотрены выше? Определенно нет. Нетрудно убедиться в том, что интересующие нас валентности А2 и А3 глагола *возмущаться* и валентности А3 и А4 глагола *возражать* несут соизмеримый объем информации, и ни одна из них не может расцениваться как валентность темы. Добавим, кроме того, что валентности темы и содержания, выраженные при

одном предикате, находятся в отношении частичной кореференции в смысле [Падучева 2004]: невозможно представить себе высказывание типа

(12r) **Я слышал о Маше и Пете, что Леша и Наташа собираются пожениться*

(если только из широкого контекста или экстралингвистической информации не следует существование связи Маши и Пети с Лешей и Наташей). Никаких ограничений такого рода на рассматриваемые здесь валентности не накладывается.

В недавней статье [Иомдин и Иомдин 2020] авторы предложили описывать эти валентности как **валентность факта** (она предназначена для представления состояния дел, с которым не согласен субъект соответствующего речевого действия — это валентность А2 у *возмущаться* и А3 у *возражать*) и **валентность аргументации** (выражающую непосредственно содержание речи этого субъекта — таковы валентность А3 у *возмущаться* и А4 у *возражать*).

Для рассматриваемых в настоящей статье случаев термин «валентность факта» может быть не вполне подходящим: он ассоциируется с фактивными глаголами, которые вводят факты реальной действительности, между тем как в выражениях типа *возмущаться некоторым Р* или *возражать против некоторого Р* пропозиция Р не обязана быть фактом: она может быть идеей, гипотетическим событием или вообще ложью.

По существу, пропозиция Р, выражаемая валентностью А2 у *возмущаться* и валентностью А3 у *возражать*, отражает повод для возмущения или возражения, мотивацию, **стимул** для речевого действия. Именно на этот стимул субъект **реагирует** при помощи второй из рассматриваемых валентностей — т. е. валентности, соответствующей содержанию его речи (для *возмущаться* это А3, для *возражать* — А4).

3. Свойства предикатных слов с валентностями стимула и реакции

Посмотрим теперь, какими свойствами обладают предикатные слова, валентная структура которых содержит валентности стимула и реакции.

Мы начнем со способов выражения обоих типов валентностей. Поскольку и та, и другая валентности являются пропозициональными, естественно ожидать, что они выражаются в первую очередь посредством союзов (точнее, придаточных, вводимых союзами), а также бессоюзными придаточными.

Примеры союзного и бессоюзного оформления этих валентностей мы приводили выше, в **разделах 1 и 2**. Рассмотрим еще несколько примеров, характеризующих разные аспекты заполнения этих валентностей.

Придаточные, реализующие при предикатах валентность реакции, часто содержат показатели косвенной речи: *мол, дескать, де, якобы* и т. д.; ср. (14) и (15):

(14) *Никанор Иванович в некотором недоумении возразил, что, мол, иностранцам полагается жить в «Метрополе», а вовсе не на частных квартирах* (М. Булгаков, Мастер и Маргарита);

- (15) *Вот собираются они там и начинают считать: на исследование кольчатых червей мы бросим сто мегаватт, на проталкивание такого-то проекта — семьдесят пять гигаватт, а на запрещение Малянова хватит и десятки. А кто-нибудь там возражает: десятки-де мало* (А. и Б. Стругацкие, За миллиард лет до конца света).

Здесь же относится употребление в клаузах, выражающих валентность реакции, характерных элементов разговорной речи, таких как *отродясь* в (16):

- (16) *Я надеваю новое платье, а он возмущается, что таких женщин у нас в роду отродясь не было* (Е. Завершнева, Высотка)

Любопытный семантический эффект можно наблюдать в примере (17):

- (17) *Помню, как я возражала, чтобы не брали с собой эти «тряпки», не брали лишних вещей. Хорошо, что мама не послушала меня: именно эти «тряпки» помогли нам выжить в эвакуации — мы меняли их на продукты* (И. Архипова, Музыка жизни).

Несомненно, здесь мы имеем дело с заполнением валентности реакции: очевидно, что позиция субъекта состояла именно в том, что лишние вещи **не следует брать с собой**, т. е. субъект возражал против того, чтобы их брали. Речевое же высказывание оказывается по смыслу полностью противоположным содержанию возражения: возражала, чтобы брали, говоря «не берите». В результате получается, что фраза

- (17а) *Возражала, чтобы не брали лишних вещей*

может быть полностью синонимична фразе

- (17б) *Возражала, чтобы брали лишние вещи.*

Объясняется это тем, что в (17а) и (17б) заполнены разные валентности — в (17а) валентность реакции, а в (17б) — валентность стимула. Очень часто реакция на стимул состоит в отрицании того, что содержится в стимуле. Этот эффект не обязательно присутствует в высказываниях с предикатами, имеющими валентности стимула и реакции, но достаточно характерен для них. Добавим, что в (17) мы имеем дело с нетипичным заполнителем валентности реакции: она почти никогда не выражается союзом *чтобы*.

Обратимся теперь к вопросу о том, каков класс предикатных слов, которые обладают набором валентностей, содержащих валентности стимула и реакции. Разумеется, рассмотренным выше списком из двух глаголов этот класс не исчерпывается, хотя он довольно ограничен. Насколько удалось обнаружить авторам, этот класс состоит из предикатов, смысл которых содержит, во-первых, отрицание некоторой пропозиции (на каком-либо уровне семантического разложения), а во вторых, речевое высказывание.

Приведем несколько примеров в дополнение к рассмотренным выше. Очень похожим образом ведет себя глагол *извиняться*. Ср. предложения

- (18) *Он извинился, что заставил себя ждать, и спросил, чем может служить* (Н. С. Лесков, На ножах),

в котором выражена валентность стимула (субъект извинения считает плохим, что заставил себя ждать, и выражает сожаление об этом), и

(19) *Он извинился, что ему надо домой, и стал прощаться* (Л. Н. Толстой, Воскресение),

в котором имеет место валентность реакции (извинился не за то, что ему надо домой, а за нечто эксплицитно не названное, скорее всего, за то, что уходит): обратим внимание, что в отличие от (18), в этом предложении неуместно добавить *за то*, хотя предлог *за* — стандартный способ реализации валентности стимула при *извиняться*.⁵

Аналогичное актантное поведение демонстрирует глагол *простить* (выступающий, прежде всего в форме императива, в качестве формулы извинения); ср. (20), где в первом предложении у *простить* выражена валентность стимула, и (21), где налицо валентность реакции этого глагола:

(20) *Простите, что я вас расстроила. Я не ожидала такой реакции* (И. Грекова, Скрипка Ротшильда)

(21) *Простите, я не хотел вас обидеть* (говорящий извиняется за нечто неназванное).

При глаголе *простить* в этом значении валентность реакции, по-видимому, не допускает выражения с союзом *что* и выражается лишь бессоюзно.

Похожим образом ведет себя в отношении рассматриваемых валентностей и перформативное выражение *прошу прощения*:

(22) *Еще раз прошу прощения, что обеспокоил Вас длинным, ненужным письмом* (А. Куприн, Гранатовый браслет) [валентность стимула] vs.

(23) *А теперь прошу прощения, меня ждет народ* (В. Липатов, И это все о нем) [валентность реакции].

Очень близкой валентной структурой обладает и квазисинонимичный глаголу *извиняться* глагол *оправдываться*, допускающий как валентность стимула, ср. (24)–(26):

(24) *Я должен был оправдываться перед невежественными людьми, что занимаюсь увлекающими меня вопросами* (А. Чижевский, Вся жизнь);

(25) *Не хожу... И не буду оправдываться, что не хожу на политические шоу* (С. Алексиевич, Время second-hand);

(26) *И мне бывает больно оправдываться за то, что я не умер* (Ф. Искандер, Сандро из Чегема) —

⁵ Отметим, что до XX века глагол *извиняться* мог управлять именной группой в творительном падеже, что в МАС помечено как устар., а в БАС еще дано без всякой пометы: 'Привести что-л. в качестве смягчающего вину обстоятельства, в качестве оправдания', ср. *Я отказался решительно от декламации, извинившись тем, что ничего припомнить не могу* (С. П. Жихарев, Записки современника). Эта именная группа заполняет валентность реакции.

так и валентность реакции:

(27) *Этот лживый барабанщик, которого давно уже вычеркнули из списков четвертого отряда, вероятно, будет плакать и оправдываться, что все вышло как-то нечаянно* (А. Гайдар, Судьба барабанщика);

(28) *Иван Поликарпович стал неловко оправдываться, что, мол, врачи иногда ошибаются в долгосрочных «прогнозах», ничего не поделаешь, живые люди, а не автоматы* (Ф. Незнанский, Черный амулет).

Есть валентности стимула и реакции у глаголов смежного лексического класса *повиниться* и *каяться/покаяться*. Так в (29)

(29) *Еще я должен повиниться, / Что в постный день я у девицы, / У этой Лиззи смуглицей, / Гостил тайком. / Но я в тот день, как говорится, / Был под хмельком* (Р. Бёрнс, Молитва святоши Вилли, пер. С. Маршака)

выражена валентность стимула, а в (30)

(30) *Повиниться надо, что сгоряча! Повиниться!* (Б. Васильев, Дом, который построил Дед) —

валентность реакции: говорящий считает, что собеседнику надо повиниться (в каком-то неназванном проступке или преступлении), и побуждает его прореагировать утверждением, что это было сделано сгоряча. Аналогично, в (31)–(32) выражена валентность стимула глагола *(по)каяться*, а в (33)–(34) — валентность реакции.

(31) *Должен покаяться, что и меня в то время захватила идея крупных сельских поселений* (А. Яковлев, Омут памяти);

(32) *Впоследствии он всю жизнь каялся, что так долго откладывал эту поездку.*

(33) *И вот однажды она бросилась мне в ноги и со слезами покаялась, что Андрей Яковлевич соблазнил ее, обещая взять к себе в дом, когда женится на барышне, т. е. на мне* (А. В. Амфитеатров, Отравленная совесть):

понятно, что кающаяся героиня повести аргументирует свое поведение, а не сообщает, в чем состоит ее грех. Тот факт, что он легко вычисляется, не имеет отношения к квалификации валентностей.

(34) *Дьякон, произносивший «о блаженном успении вечный покой», поперхнулся смехом на этом месте, потом каялся, что-де только дошел до высокой ноты, посмотрел на протоиерея и сразу в глазах, как живой, козел зеленый явился, как дьявол смешнувший...* (К. С. Петров-Водкин, Моя повесть).

Примечательно, что глагол *(по)каяться* имеет валентности стимула и реакции, а *раскаяться*, во всяком случае, в современных употреблениях — не имеет: в его смысл не входит обязательное речевое действие, и весьма сомнительно высказывание типа **Она раскаялась, что Андрей Яковлевич соблазнил ее*. У *раскаяться* есть лишь валентность стимула, которую в отсутствие оппозиции скорее следует считать валентностью содержания, как в (35)–(36):

(35) *Я тотчас же раскаялся, что напугал ее, заставил считать часы, и проклял себя за припадок злости* (Ф. М. Достоевский, Белые ночи);

(36) *Товарищ Бездомный, помилуйте, — ответило лицо, краснея, пятась и уже раскаиваясь, что ввязалось в это дело* (М. Булгаков, Мастер и Маргарита).

Еще один интересный лексический класс предикатов, имеющих валентности стимула и реакции — это несколько речевых глаголов со значением отказа: *отнекиваться* (см. примеры (46)–(47) ниже), *отпираться* и *отбрехиваться*. Приведем несколько примеров этих глаголов без подробных комментариев.

(37) *Вы ко мне писали, / Не отпирайтесь* (А. С. Пушкин, Евгений Онегин):

здесь первое предложение бесспорно реализует валентность стимула при глаголе *отпираться* (= не отпирайтесь, что вы ко мне писали);

(38) *Я был среди тех кружковцев, кто высказывался и веселился. Отпирался, что пишу стихи, и впервые прочел их только осенью 1939 года на третьем году хождения к Брику* (Б. Слуцкий, О других и о себе);

(39) *Он этого не знал и готовился по-прежнему отпираться, что-де бунта не заводил и на государево здоровье не умышлялся...* (А. Н. Толстой, Петр Первый).

В (38) тоже выражена валентность стимула, а в (39) — валентность реакции при глаголе *отпираться*.

(40) *Я так и не смог отбрехаться, что прогулял работу* (валентность стимула у отбрехиваться);

(41) *Пробую отбрехаться, что с работой я не знаком* (В. Лазарев, Сборник воспоминаний) — валентность реакции.

Интересно, что более ходовые слова этого класса — *отказываться* и *отрицать* — порой рассматриваемых валентностей этого класса, как кажется, не обладают, по-видимому, поскольку они в меньшей степени могут рассматриваться как речевые акты.

Приведем еще один класс предикатов, имеющих эту пару валентностей: это глаголы, обозначающие высокую степень возмущения, сопровождающегося резкой речевой реакцией: *вспылить* и *взвиться*. В предложениях (42)–(43) выражена валентность стимула этих глаголов, а в предложениях (44)–(45) — валентность реакции).

(42) *Однажды будущий крестный отец вспылил, что его заставляют слушать «всякую ерунду за его же деньги»* (сайт «Дело Модульбанка»);

(43) *Индир взвился, что ему не хватит сил одновременно залечить и обезболить. Придется выбрать что-то одно* (А. Волкова, Гордость черного дракона);

- (44) *Акира вспыллил, что он имеет право делать то, что хочет* (интернет-форум Slash World);
- (45) *Когда я сказала ему об этом, он взвился, что я его преследую, и забанил меня везде* (интернет-форум Cosmo).

Отметим в заключение этого раздела, что все предикатные слова, имеющие валентности стимула и реакции, устроены семантически похожим образом: они вводят некоторую ситуацию, которая с какой-то точки зрения является неудовлетворительной, и эта ситуация стимулирует речевую реакцию, которая данную ситуацию объясняет. Если хотя бы один из этих двух элементов — исходная неудовлетворительная ситуация или реакция на нее, в семантике слова отсутствует, появление таких валентностей исключено. Поэтому такие речевые глаголы, как *сообщать, информировать, обрадовать*, равно как и *сврать, солгать, обмануть* этой парой валентностей не обладают. Еще один характерный пример — внешне похоже устроенные глаголы *отнекиваться* и *поддакивать*, образованные от реплик *не(т)* и *да*, соответственно. Глагол *отнекиваться* имеет валентности стимула и реакции (описанные в [Июмдин и Июмдин 2020] как валентности факта и аргументации), ср.

- (46) *Так вот почему она отнекивалась, что ест по ночам* (П. Дрожжинова. Внутренние миры) —

факт, стимул vs.

- (47) *Лецев попутно уговаривал меня выпить портвейну, а я отнекивался, что и так слишком много пью* (Е. Попов. Подлинная история «Зеленых музыкантов») —

аргументация, реакция. Глагол же *поддакивать*, в семантику которого не входит отрицание, имеет только вторую из этих валентностей — реакции, которую можно просто называть валентностью содержания:

- (48) *Я поддакивал, что это, дескать, точно, пустяки* (Н. Г. Чернышевский, Что делать).

Отметим, что в принципе реакция может быть и невербальной, и тогда валентность реакции заполняется не придаточным предложением с союзом, а деепричастным оборотом или конструкцией с союзом *и* и глаголом в личной форме.⁶

- (49) *Третьи возмущаются, выискивая в новой моде что-то вызывающее, потом ее все равно принимают, но с большим опозданием* (Н. Аршавская, Искусство одеваться);

⁶ Напомним, что такие неканонические способы заполнения валентности глаголов известны и описаны в литературе; таковы, например глаголы *спохватиться, изловчиться* [Богуславский 1996: 31] или интерпретационные глаголы [Апресян 2004].

- (50) *До конца приема работ осталось всего четыре дня, а самому господу богу для сотворения мира потребовалась целая неделя. Но Степан возразил, показав в ответ шесть пальцев — к четырем дням он прибавил две ночи* (К. Абрамов, Степан Эрзя);
- (51) *Примерно через квартал поравнялся с ним, он мне бикнул и через стекло извинился, показав жестами, что не увидел меня* (автомобильный интернет-форум);
- (52) *Потому рокер и вспылил, нанеся «пошутившему» фанату внушительный удар* («Комсомольская правда», 2013.03.16);
- (53) *Когда Фомин ударил его по лицу, Валько вспылил и сбил полицейского с ног* (А. Фадеев, Молодая гвардия).

Интересно, что в таком случае валентность реакции может выражаться и у глаголов *отказаться* и *соглашаться*:

- (54) *Кивком предложила выпить управляющему — тот отказался, покачав головой и похлопав рукой по правому боку: печень* (М. Нестеров, Агент ливийского полковника);
- (55) — *Может, прорвемся с боем, все-таки Аллах с нами? — сказал кто-то. Я отказался и покачал головой* (А. Самари, Русская жена эмира);
- (56) *Рыцарь молча соглашался, кивая головой* (М. Шелков, Проект «Калевала»).

Конструкция с деепричастием может называть и речевую реакцию (но не стимул):

- (57) *Наташа Лаптева попробовала было возразить, сказав, что нельзя бить человека по лицу, ногами, нельзя!* (С. Соловейчик, Ватага «Семь ветров»).
- (58) *Я очень возмутился, говоря, что служу уже шесть лет и все питаюсь ихними завтраками* (Е. Попов, Мыслящий тростник).

Итак, предикаты, в значение которых входят речевой акт, отрицание и оценка, имеют валентности стимула и реакции, если они вводят ситуацию, которую не одобряет субъект, и эта ситуация стимулирует речевую реакцию, которая ее объясняет. При этом обе валентности могут заполняться пропозициональным актантом, что в некоторых случаях затрудняет интерпретацию соответствующих предложений вне широкого контекста.

4. Слова с двумя валентными центрами

Завершая рассмотрение предикатов с валентностями стимула и реакции, попытаемся сделать некоторое семантическое обобщение.

По нашему убеждению, в русском языке есть несколько классов предикатов с нетривиальной семантикой, которые характеризуются наличием связанных друг с другом семантических валентностей — таких, которые у других предикатов не различаются и объединены в одну валентность. Два таких

класса были рассмотрены выше — это класс слов с валентностями содержания и темы и класс слов с валентностями стимула и реакции. Нам хотелось присокупить к ним еще один класс слов, в которых фигурируют связанные валентности. Мы имеем в виду русские глаголы с приставкой *до*, которые описывают немонотонные действия достижения предела: это, в частности, глаголы движения и положения в пространстве типа *добежать*, *доплыть*, *доехать*, *дойти*, *досидеть*, *долежать*, а также глаголы завершения действия типа *доучиться*, *доработать*, *досмотреть*. Как было показано в [Апресян 2006] и уточнено в [Iomdin and Iomdin 2013], эти глаголы приобретают в дополнение к основной (в первую очередь объектной) валентности производящего глагола (*бежать*, *плыть*, *учиться*, *смотреть* и т. п.) валентность недостающего количества. Эта валентность в большинстве ситуаций выражается у глаголов с отрицанием (ср. *не добежал всего десять метров до финиша*, *не доучился один семестр в университете и пошел в армию*), но допустима и в отсутствие отрицания (*добежал последние десять метров*, *доучился оставшийся семестр в университете*) и т. п.

Таким образом, рассмотренные нами предикаты не уникальны и пополняют арсенал языковых единиц, имеющих два семантических центра. Эти единицы, конечно, находятся на периферии семантической системы языка, но занимают в ней заметное место.

Литература

1. *Активный словарь* (2014): Активный словарь современного русского языка. / В. Ю. Апресян, Ю. Д. Апресян, Е. Э. Бабаева, О. Ю. Богуславская, И. В. Галактионова, М. Я. Гловинская, Б. Л. Иомдин, Т. В. Крылова, И. Б. Левонтина, А. А. Лопухина, А. В. Птенцова, А. В. Санников, Е. В. Урысон; Отв. ред. Ю. Д. Апресян. — М.: Языки славянской культуры, (2014). — Т. 2. ISBN 978-5-(9906)039-2-9.
2. *Апресян Ю. Д.* (2001). Системообразующие смыслы «знать» и «считать» в русском языке. // *Русский язык в научном освещении*. — № 1. — М. — С. 5–26.
3. *Апресян Ю. Д.* (2004). Интерпретационные глаголы: семантическая структура и свойства. // *Русский язык в научном освещении*. — № 1 (7). — М. — С. 5–22.
4. *Апресян Ю. Д.* (2006). Типы соответствия семантических и синтаксических актантов // *Проблемы типологии и общей лингвистики*. — СПб., — С. 15–27
5. *Апресян и др.* (2010). Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин, В. З. Санников. Теоретические проблемы русского синтаксиса: взаимодействие грамматики и словаря [Studia philologica]. М.: Языки славянских культур.
6. *Богуславский И. М.* (1996). Сфера действия лексических единиц. М.: Школа «Языки русской культуры».
7. *Иомдин Б. Л., Иомдин Л. Л.* (2020). Валентность факта и валентность аргументации. // *От семантических кварков до вселенной в алфавитном порядке*. Сборник статей к 90-летию академика Юрия Дерениковича Апресяна. Труды ИРЯ РАН им. В. В. Виноградова, вып. XXIV (в п

8. Кустова Г. И. (2015). Валентности абстрактных существительных: Редукция vs. Конкретизация // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог (2015)». Вып. 14 (21): М.: РГГУ.
9. Левонтина И. Б. (2016). О чем речь? М., Corpus.
10. Мельчук И. А. (1977). О типах поверхностно-синтаксических отношений (три критерия различения). // *Russian Linguistics*, Vol. 3, No. 3/4, pp. 245–270.
11. Падучева Е. В. (2004). Динамические модели в семантике лексики. М., Языки славянской культуры.
12. Iomdin B., Iomdin L. (2013). Negation and Valencies of Russian Predicates. // *Meaning Text Theory: Current Developments. Wiener Slawistischer Almanach*, issue 85. München.

References

1. *Active Dictionary* (2014), Active dictionary of the Russian language [Aktivnyy slovar' russkogo yazyka]. V. 2. / V. Yu. Apresjan, Yu. D. Apresjan, E. E. Babaeva, O. Yu. Boguslavskaya, I. V. Galaktionova, M. Ya. Glovinskaya, B. L. Iomdin, T. V. Krylova, I. B. Levontina, A. A. Lopukhina, A. V. Ptentsova, A. V. Sannikov, E. V. Uryson. Editor-in-Chief: Yu. D. Apresjan. — Moscow, Jazyki russoj kultury. — 736 p.
2. Apresjan Yu. D. (2001), Systemic meanings «know» and «believe» in Russian. [Sistemoobrazujushchie smysly «znat'» i «schitat'»]. // *Russkij jazyk v nauchnom osveshchenii*. No. 1. P. 5–26.
3. Apresjan Yu. D. (2004), Interpretation words: semantic structure and properties. [Interpretacionnye glagoly: semanticheskaja struktura i svojstva]. // *Russkij jazyk v nauchnom osveshchenii*. No. 1 (7). P. 5–22.
4. Apresjan Yu. D. (2006), Types of relations between semantic and syntactic actants [Tipy sootvetstvija semanticheskix i sintaksicheskix aktantov]. // *Problemy tipologii i obshchej lingvistiki*. St. Petersburg. P. 15–27.
5. Apresjan Yu. D. et al. (2010), Yu. D. Apresjan, I. M. Boguslavsky, L. L. Iomdin, V. Z. Sannikov. Theoretical Issues of Russian Syntax: Interaction of the Grammar and the Lexicon. [Teoreticheskie problemy russkogo sintaksisa: vzaimodejstvie grammatiki i slovarja.] Moscow, Jazyki slavyanskix kul'tur. [Studia philologica].
6. Boguslavsky I. M. (1996), The sphere of action of lexical units. [Sfera dejstvija leksicheskix edinic]. Moscow, Shkola “Jazyki russoj kultury”.
7. Iomdin B., Iomdin L. (2013), Negation and Valencies of Russian Predicates. // *Meaning Text Theory: Current Developments. Wiener Slawistischer Almanach*, issue 85. München.
8. Iomdin B. L., Iomdin L. L. (2020), Valency of Fact and Valency of Argumentation [Valentnost' fakta i valentnost' argumentatsii. // Валентность факта и валентность аргументации. // *Ot semanticheskix kvarkov do vselej v alfavitnom porjadke. Sbornik statej k 90-letiju akademika Jurija Derenikovicha Apresjana*. Trudy IRYA RAN im. V. V. Vinogradova, vyp. XXIV (in print).

9. *Kustova G. I.* (2015), Valencies of Abstract Nouns. Reduction vs. Specification. [Valentnosti abstraktnyx sushchestvitel'nyx. Reduktsija vs. Konkretizatsija. // Валентности абстрактных существительных: Редукция vs. Конкретизация // Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conferences «Dialogue-2015». Issue 14(21). Moscow, RSUH.
10. *Levontina I. B.* (2016), What is it about? [O chem rech'?). Moscow, Corpus.
11. *Mel'čuk I. A.* (1977), On the types of surface-syntactic relations (three criteria for distinguishing). [O tipax poverxnostno-sintaksicheskix otnoshenij (tri kriterija razlichenija)]. // Russian Linguistics, Vol. 3, No. 3/4, pp. 245–270.
12. *Paducheva E. V.* (2004), Dynamic Models in the Semantics of the Vocabulary [Dinamicheskie modeli v semantike leksiki]. Moscow, Jazyki slavyanskoj kul'tury.

RUREBUS-2020 SHARED TASK: RUSSIAN RELATION EXTRACTION FOR BUSINESS

Ivanin V. A. (vitalii.ivanin@abbyy.com)^{1,2},
Artemova E. L. (Echernyak@hse.ru)³,
Batura T. V. (tatiana.v.batura@gmail.com)^{4,7},
Ivanov V. V. (nomemm@gmail.com)^{5,7},
Sarkisyan V. V. (vsarkisyan@hse.ru)³,
Tutubalina E. V. (elvtutubalina@kpfu.ru)^{6,7},
Smurov I. M. (ivan.smurov@abbyy.com)^{1,2}

¹ABBYY, ²Moscow Institute of Physics and Technology,

³National Research University Higher School of Economics,

⁴Novosibirsk State University, ⁵Innopolis University,

⁶Kazan Federal University,

⁷Lomonosov Moscow State University

In this paper, we present a shared task on core information extraction problems, named entity recognition and relation extraction. In contrast to popular shared tasks on related problems, we try to move away from strictly academic rigor and rather model a business case. As a source for textual data we choose the corpus of Russian strategic documents, which we annotated according to our own annotation scheme. To speed up the annotation process, we exploit various active learning techniques. In total we ended up with more than two hundred annotated documents. Thus we managed to create a high-quality data set in short time. The shared task consisted of three tracks, devoted to 1) named entity recognition, 2) relation extraction and 3) joint named entity recognition and relation extraction. We provided with the annotated texts as well as a set of unannotated texts, which could of been used in any way to improve solutions. In the paper we overview and compare solutions, submitted by the shared task participants. We release both raw and annotated corpora along with annotation guidelines, evaluation scripts and results at <https://github.com/dialogue-evaluation/RuREBus>.

Keywords: named entity recognition, relation extraction, shared task, Russian fine-tuning, BERT

DOI: 10.28995/2075-7182-2020-19-416-431

RUREBUS-2020: СОРЕВНОВАНИЕ ПО ИЗВЛЕЧЕНИЮ ОТНОШЕНИЙ В БИЗНЕС-ПОСТАНОВКЕ

Иванин В. А. (vitalii.ivanin@abbyy.com)^{1,2},
Артемova Е. Л. (echernyak@hse.ru)³,
Батура Т. В. (tatiana.v.batura@gmail.com)^{4,7},
Иванов В. В. (nomemm@gmail.com)^{5,7},
Саркисян В. В. (vsarkisyan@hse.ru)³,
Тутубалина Е. В. (elvtutubalina@kpfu.ru)^{6,7},
Смуrow И. М. (ivan.smurov@abbyy.com)^{1,2}

¹АВВУУ, ²Московский Физико-технический Институт

³Национальный исследовательский университет
Высшая школа экономики

⁴Новосибирский государственный университет

⁵Иннополис, ⁶Казанский федеральный университет

⁷МГУ им. М. В. Ломоносова

В статье представлены результаты соревнования по распознаванию именованных сущностей и извлечению отношений. Целью соревнования является сравнение методов извлечения сущностей и отношений на русском языке в постановке, приближенной к индустриальным задачам. В качестве исходной коллекции текстов использовался корпус Минэкономразвития РФ, содержащий программы стратегического развития. Корпус был размечен в соответствии с инструкцией, разработанной авторами статьи. В процессе разметки использовались различные методы активного обучения, что позволило за короткое время создать качественный набор данных. Всего было размечено более двухсот документов. Соревнование проводилось по трем задачам (дорожкам): 1) распознавание именованных сущностей, 2) извлечение отношений и 3) совместное распознавание именованных сущностей и извлечение отношений. Вместе с коллекцией размеченных текстов участникам также были предоставлены неразмеченные тексты, которые могли быть использованы для улучшения решений. В статье дается обзор и сравниваются результаты участников соревнования. Детальное описание соревнования, текстовые коллекции, инструкция по разметке и скрипты для оценки качества доступны по ссылке: <https://github.com/dialogue-evaluation/RuREBus>.

Ключевые слова: распознавание именованных сущностей, извлечение отношений, соревнование, русский, дообучение, BERT

1. Introduction

Structuring unstructured information is one of the most popular industrial application of natural language processing. Standard approaches to it require named entity recognition (NER) and/or relation extraction (RE). NER and RE are classical NLP tasks, formulated as early as mid-1990s [39]. There exist a number of well-studied academic corpora (see next section for multiple examples of such corpora). Scores obtained on these corpora are typically high. Taking recent advances in NER in account one can even assume that it is a largely solved task.

However, business applications seldom do enjoy the high scores reported in academia. In our opinion the main reason for that is the fact that both text sources and entities in industry and academia present with several noticeable differences.

Firstly, business case texts are usually domain-specific (e. g. legal) texts that can contain less than perfect language or other irregularities (ponderous sentences with complicated syntactic structure, slang etc.). Academic baselines, on the other hand, typically consist of well-written news or biography (or scientific in case of BioNLP) texts without any irregularities of this kind.

Secondly, while entities in academia are usually compact and well-defined, industry sometimes has to deal with something much more loose, spanning for many words and with less than clear borders.

Our main motivation for conducting this work was to attempt to bridge the gap between academic NLP and less-than-ideal business scenarios. In order to do so, we have collected and marked up a corpus of governmental documents, produced by the Ministry of Economic Development of the Russian Federation and organized a shared task on it, which are two main contributions of this paper.

2. Related work

In this section we number related research areas:

1. general domain named entity recognition and relation extraction
2. methods for named entity recognition and relation extraction
3. named entity recognition and relation extraction for the Russian language

2.1. General domain named entity recognition and relation extraction

The entity recognition task is a necessary stage of extracting information from texts. Today there are quite a lot of datasets for the task in different languages for the general domain, such as CONLL 2003 [39], MUC-6 [17], OntoNotes 5 [19], etc.

These datasets usually poses a few types of named entities, such as *persons*, *organizations*, *locations* and casual relations, such as *being born in*, *have position at*, etc.

To perform semantic analysis, it is also important to extract relations that link named entities. This requires building datasets for solving the problem of relation extraction. The relation extraction problem goes further than named entity recognition, as it requires greater understanding of language semantics.

Therefore, there are fewer datasets available both for named entity recognition and relation extraction. The most used datasets are CONLL04 [10], ACE 2005 [47], TACRED [44], SemEval 2010 Task 8 [18].

2.2. Methods for named entity recognition and relation extraction

At the core of the majority of current methods both for named entity recognition and relation extraction are pre-trained language models, such as ELMo [31], BERT [14] and their descents. Using pre-trained language models does not require training a model from scratch, but rather fine-tuning the model for the task under consideration. An example of BERT fine-tuning is presented in [37]. To achieve relation representation by fine-tuning BERT with a large scale “matching the blanks” pre-training entity linked texts are used. This method performs well on the SemEval 2010 Task 8 dataset (F1-measure of 89.5%) and outperforms previous methods on TACRED (F1-measure of 71.5%). For the entity recognition task, the BERT-MRC model [25] achieves the best results on ACE 2005 (F1 score of 86.88%).

The state-of-the-art approach to relation extraction is an entity pair graph-based neural network (EPGNN) model, relying on a graph convolutional network [45]. EPGNN combines sentence semantic features generated by a pre-trained BERT model with graph topological features for relation classification. It shows a macro-F1 score of 90.2% on the SemEval 2010 Task 8 dataset and a micro-F1 score of 77.1% on ACE 2005.

2.3. Named entity recognition and relation extraction for the Russian language

To the best of our knowledge, several datasets for named entity recognition in the Russian language are available: the dataset, developed by Gareev et al. [16], Persons 1000 and Collection 5 [30], [40], [42], FactRuEval 2016 [7], the Russian subset of the BSNLP Shared Task [33].

Prior to deep and even machine learning methods, rule-based approaches dominated the information extraction systems [11], [15]. Most of the early works for the Russian language NER describe systems based on linguistic resources: dictionaries, templates, and rules [9], [12], [35]. Popov et al. described the adaptation of the vocabulary approach for the Russian language [35]. Craidlin introduced the TagLite program, which aims to distinguish named groups consist of three types of proper names: persons, organizations and geographical objects [12]. The system includes the following dictionaries: proper names, generic concepts of investigated entity types (director, river, office) and other auxiliary words that can be part of target noun groups. In order to resolve the ambiguity and process words that are not encountered in dictionaries, the rule-based “predictor” module is applied. The authors evaluated the quality of the system on their own annotated corpus. TagLite obtained 85.8% of F-measure for all categories of named groups. Brykina et al. proposed a system that recognizes named entities based on lists of terms from the input ontology and resolving polysemy with a set of manually developed rules and dictionaries of context words [9]. The authors evaluated the efficiency of a system on their own corpus, considering only entities

included in the ontology. The system obtained F-measure varying from 91% to 98% for different types of entities. Both systems were evaluated on closed corpora, which makes it difficult to conduct a comparative analysis of the achievements in this area.

Starting from 2013, studies about Russian NER [4, 16, 34] started to apply Conditional Random Fields (CRF) [23]. Antonova et al. applied a CRF model to their own annotated corpora consisting of news feed texts [4]. There were five types of annotations: person names, geographical objects, organization, products, and events. The authors also evaluated different types of optimizers for CRF. The highest F-measure obtained by this approach was 87.18%. Podobryaev applied CRF model to person recognition and used information from ontology as one of the features [34]. The quality of the proposed approach was evaluated on a manually annotated corpora. Gareev et al. developed an annotated corpus of Russian-language texts for evaluating NER methods and compared the effectiveness of two approaches [16]. The first approach is based on dictionaries of names and rules, which analyze the context of a named entity and compare the set of references to the same entity in a document. The second is based on the CRF model with various features. The developed corpus is publicly available and contains two types of annotations: persons and organizations. The results of experiments showed that CRF-based approach outperformed knowledge-based approach on 13% of F-measure. Mozharova and Loukachevitch investigated the knowledge and context features for the CRF model in the NER task [30].

Recent works on Russian NER focused on neural network models. Anh et al. [2] investigated a combination of bidirectional Long Short Term Memory (LSTM) with CRF and word embeddings presented in [20], [24], [26]. This model showed the best results on three Russian language data sets Gareev's [16], Person-1000 [40] and FactRuEval 2016 [7]. In the work [28], authors showed close to the state of the art results at the time, exploiting only neural model trained on the small dataset without pre-training. There is another aspect of robustness of LSTM-CRF in NER task addressed in the paper [27]. The experiments were conducted in three datasets, Persons-1000 for Russian language, CAp'2017 for French, and CoNLL'03 for English. Remarkably a proprietary system, which combines rule-based approach to statistics analysis, achieves state of the art [7] for the FactRuEval dataset. For fair comparison it should be noted, that proprietary systems are being developed for longer period of time, while the majority of shared task participants train the models from scratch without possessing rule and code base.

The BSNLP 2019 Shared Task [32] introduced a new multilingual dataset, annotated with named entities for four Slavic languages, Bulgarian, Czech, Polish, Russian. The named entities considered were persons, locations, organizations, events, products. The majority of the systems which participated in the shared tag exploited BERT-based solutions by either fine-tuning multilingual BERT [41] or by training the BERT model from scratch for the target languages [5].

Much less attention is paid to relation extraction for the Russian language compared to named entity recognition. To the best of our knowledge, three datasets are annotated with relations. The aforementioned FactRuEval dataset [7] provides with two layers of annotations, the first layer being named entities and the second layer being relations between them. As FactRuEval covers news domain the relations annotated express attribution and occupation properties as well as some facts, such as meetings and deals. Existing datasets [22], [42] are much smaller and are not widely used for experiments.

3. Data

In the RuREBus Shared Task, we proposed a corpus with annotations of named entities and relations. The novelty of the task is in its focus on methods for the extraction of business-relevant entities and relations from corporate documents. For obvious reasons, it is hard to find such documents online and make a shared collection of business-related documents. Therefore, we use a similar collection of corporate documents shared by the Ministry of Economic Development of Russia.

3.1. General corpus description

The collection shared by the Ministry of Economic Development of Russia contains strategic planning programs of development for Russian regions. The corpus was studied in [1]. It has 85,501 documents, 298,809,024 tokens overall. Key features of the corpus include:

- uniformity of texts: documents have the same domain, purpose, very similar style and size;
- shared scope: documents mention various types of economic and social entities and relations at different levels of management;
- fixed modalities: a fixed list of modalities in documents that cover current state of the economy or society (problems), as well as plans for future (actions, tasks, etc.)

For the purpose of the RuREBus Task we selected and annotated 218 documents. The annotation guidelines and results of manual annotation are presented below and available at <https://github.com/dialogue-evaluation/RuREBus>.

3.2. Annotation Guidelines

To support consistency of markup we developed an instruction for entity and relation identification. We also use Brat Rapid Annotation Tool (BRAT) [38] to provide an annotation interface for assigning entities and relations. Preview may be seen in Figure 1.

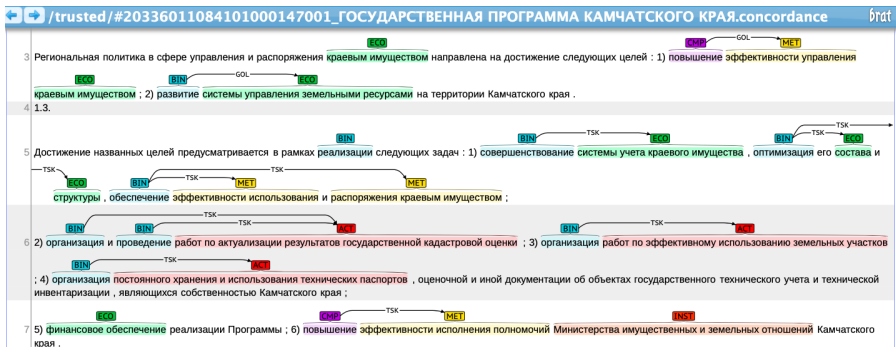


Figure 1: Annotation interface for assigning entities and relations

We partly employ double annotation. For each annotator we compare several documents with another independent annotation by the verified annotator. This mirroring helps the moderator to resolve arguable cases. After moderation we consider annotator as experienced enough and approve markup without doubling. However, we moderate each document manually even for experienced annotators.

The Cohen's kappa measured on the documents that were marked up twice (not taking into account moderators) is equal to 0.698.

3.3. Entity Descriptions

We developed eight entity types for annotation. Entities are described in [Table 1](#).

Table 1: Entity descriptions

Entity	Entity description	Examples (English)	Examples (Russian)
MET (metric)	indicator or object on which the comparison operation is defined	students' education level total length of roads birth rate economic growth	уровень образования студентов общая протяженность дорог уровень рождаемости экономический рост
ECO (economics)	economic entity (excluding MET) or infrastructure object	private business PJSC Gazprom fuel and energy complex library and museum funds	частный бизнес ПАО Газпром топливно-энергетический комплекс библиотечные и музейные фонды
INST (institution)	institutions, structures and organizations	Youth Employment Center Family and Child Support Organizations metro stations road system	Центр занятости молодёжи Организации поддержки семьи и детства станции метрополитена система дорог
BIN (binary)	binary characteristics or single action	modernization rendering is functioning absence	модернизация оказание функционирует отсутствие
CMP (compare)	comparative characteristic	increase saturation excess of level negative dynamics	рост насыщение превышение уровня негативная динамика
QUA (qualitative)	quality characteristic	effective stable safe poorly developed	эффективный стабильный безопасный плохо развитый
SOC (social)	social object	scientific and educational potential leisure activities folk art the youth	научный и образовательный потенциал досуг народное творчество молодежь
ACT (activity)	activities, events or measures taken by the authorities; these entities are often combined with BIN, e.g., <BIN> developed </BIN> <ACT>an educational project for rural schools </ACT>	restoration work educational project "University 2020" orphan prevention weekend fair	реставрационные работы образовательный проект «Университет 2020» профилактика сиротства ярмарка выходного дня

3.4. Relations Description

We define two relations to describe plans and goals, and nine to describe state of affairs. These relation types could be useful in specific practical applications [6]. **GOL** relation represents abstract goals and aims of the program, e.g., *birth rate increase*. These goals are some objectives that must be achieved as a result of programs' actions.

TSK relation corresponds to concrete tasks and actions taken to achieve some goals, e.g., *opening of new metro stations*.

The other nine relations can be grouped by two criteria: time component (past **P**, present **N**, future **F**) and estimation component (negative **NG**, neutral **NT**, positive **PS**). Past negative, neutral and positive relations (**PNG**, **PNT**, **PPS** respectively) denote implemented changes, present relations (**NNG**, **NNT**, **NPS**) describe the current state of affair, and future relations (**FNG**, **FNT**, **FPS**) present plans and forecasts.

The examples of annotated relations are shown in **Table 2**.

Table 2: Relation examples

Relation	Example (English)	Example (Russian)
GOL	<CMP> increasing </CMP> <MET> accessibility of transport services </MET>	<CMP> увеличение </CMP> <MET> доступности транспортных сервисов </MET>
TSK	hospital <ACT> overhaul </ACT>	<ACT> капитальный ремонт </ACT> </ACT> больницы
PPS	<ACT> road works </ACT> <BIN> are completed </BIN>	<ACT> дорожные работы </ACT> <BIN> завершены </BIN>
FNG	<ECO> ruble exchange rate </ECO> <CMP> is expected to drop </CMP>	<CMP> ожидается снижение </CMP> </CMP> <ECO> курса обмена рубля </ECO>
NNT	<ECO> salary level </ECO> <BIN> stabilized </BIN>	<ECO> уровень заработной платы </ECO> <BIN> стабилизировался </BIN>

3.5. Active learning

We also use an active learning technique [36] to help the annotators and speed up their work. Firstly we obtained a subset of the corpus marked with defined named entities and relations. Next we trained the NER model and employ it further to markup unlabeled documents. Then documents were marked up by annotators. The annotations were verified by moderators. After obtaining new parts of the final corpus model were retrained with this part added to training set. Full pipeline could be seen in **Figure 2**.

In this work we employ a basic NER model, namely char-CNN-BiLSTM-CRF (proposed by Lample et al [24] and further developed by Ma and Hovy [26]). This architecture is widely used as a robust baseline in sequence tagging tasks. We use FastText [8] embeddings trained by RusVectors [21]. We also employ morphological, syntactical and semantical features, obtained from Comprono [3], [46] and some hand-made features, such as capitalization templates and dependency tree distance between relation members.

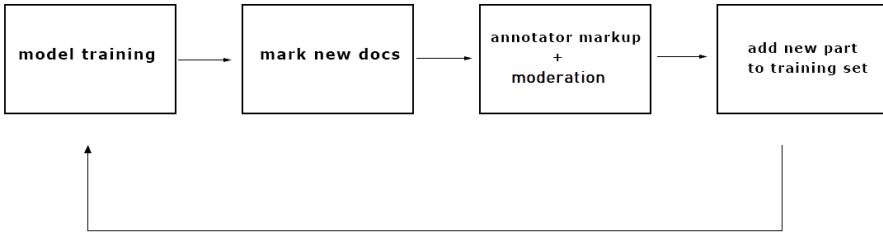


Figure 2: Active learning pipeline

3.6. Basic statistics

Table 3: Statistics of annotated entities

	total	mean len (std)
BIN	30,201	1.05 (0.28)
MET	14,161	4.23 (3.50)
QUA	7,719	1.14 (0.52)
CMP	9,288	1.16 (0.78)
SOC	10,834	2.77 (2.31)
INST	7,903	3.69 (2.81)
ECO	24,853	2.78 (2.19)
ACT	12,274	4.74 (4.57)

We computed descriptive statistics based on annotated documents. Each document was divided into parts by 150 sentences cutoff. In the training set there are 188 annotated parts, where the average number of named entities in a file is 289 and there are 67 relations in average. The mean file length is 1,787 tokens. In the test set there are 30 files, 287 entities and 67 relations on average, the mean length is 1,967 tokens. Tokenization was performed by [razdel tokenizer](#).¹

4. Shared Task Set-Up

The participants were offered 3 different NER and RE tasks:

1. Named entity recognition.
2. Relation extraction. In this task, the participants were provided with manually annotated named entities. The task was to extract relations between them.
3. End-to-end relation extraction. The participants were to extract both named entities and the relations between them.

¹ <https://github.com/natasha/razdel>

All three tasks were evaluated with micro-averaged F-measure (evaluation script is available at <https://github.com/dialogue-evaluation/RuREBus>).

Since Task 2 requires gold standard NER labels, evaluation was organized in two phases. During phase one participants had raw texts of the test set without any markup and were able to solve Tasks 1 and 3. After phase one completion, gold standard labels for all test set texts were provided and evaluation on Task 2 commenced.

During both phases “true” test set was mixed within 514 unannotated texts in order to deny participants the possibility of identifying the exact texts used for evaluation. For phase two, NER markup for these additional texts was obtained with the model used for active learning.

5. Results and analysis

5.1. Results

We have received several submissions after the Shared Task baseline (but before the gold-standard test markup was published). While these results are not considered being part of official Shared Task evaluation, it is prudent to provide this numbers. Participant *davletov-aa* was able to achieve f-measure of 0.132 on Task 3, while *bondarenko* got 0.498 on Task 1.

Table 4: Results of the competition (Micro F1-score). Table is sorted by scores on the NER task, but all 3 tasks are equally important.

Team	NER	RE with NEs	End-to-end RE
davletov-aa	.561	.394	—
Sdernal	.464	.441	—
ksmith	.463	.152	.062
viby	.417	.218	—
dimsolo	.400	—	—
bond005	.338	.045	—
Student2020	.253	—	—

Table 5: F1-score performance measure on the NER task by NE class

Team	ACT	BIN	CMP	ECO	INST	MET	QUA	SOC
davletov-aa	0.33	0.62	0.79	0.52	0.52	0.57	0.56	0.45
Sdernal	0.22	0.60	0.77	0.41	0.34	0.43	0.49	0.29
ksmith	0.13	0.52	0.82	0.36	0.34	0.41	0.54	0.33
viby	0.21	0.49	0.77	0.29	0.33	0.39	0.52	0.23
dimsolo	0.12	0.50	0.70	0.30	0.37	0.31	0.40	0.29
bond005	0.08	0.56	0.74	0.28	0.16	0.28	0.39	0.14
Student2020	0.10	0.20	0.28	0.28	0.31	0.34	0.34	0.22
average	0.17	0.50	0.70	0.35	0.34	0.39	0.46	0.28

Table 6: F1-score performance measure on the RE task by RE class

Team	NNG	NNT	NPS	FNG	FNT	FPS	PNG	PNT	PPS	GOL	TSK
bond005	0.02	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.03	0.08	0.14
davletov-aa	0.63	0.16	0.28	0.34	0.23	0.45	0.52	0.30	0.48	0.32	0.42
ksmith	0.25	0.00	0.13	0.05	0.00	0.11	0.00	0.00	0.07	0.27	0.13
Sdernal	0.62	0.19	0.24	0.44	0.00	0.39	0.23	0.10	0.58	0.43	0.47
viby	0.43	0.01	0.05	0.06	0.05	0.08	0.06	0.07	0.12	0.29	0.23
average	0.39	0.08	0.14	0.18	0.06	0.21	0.16	0.10	0.26	0.28	0.28

One can easily notice that the scores obtained for all tasks are incomparable to the ones usually reported on most widespread academic corpora such as CoNLL-03. In our opinion this fact cannot be attributed to the methods used by participants since most popular approaches were tested by them (as is shown in the next section). However, these less than perfect scores are much closer to the scores often obtained in industry.

Another comparison we can draw is with SemEval-2020 Task 11 [13], a shared task on detecting propaganda spans from text (where various linguistic structures were considered propaganda, such as “Loaded Language”, “Whataboutism, Straw Men, Red Herring” or “Flag-Waving”). Both Propaganda Detection and RuREBus required identifying non-trivial entities, often spanning for many words, and on both winning solution is within 0.5–0.6 f-measure range. While not all business applications require entities of this type, such scenarios exist one shouldn’t expect CoNLL-2003 scores on such corpora.

In order to get a better understanding of the nature of our participants errors we decided to compute additional metrics: char-level F1-score (as opposed to span-level score reported previously). We observe that the most illustrative statistics is the difference between char-level F1-score and span-level F1-score. The average difference for top 3 participants is provided in Table 7 along with the mean length of each entity in chars.

Table 7: Differences in char-based f-measure and span-based

Metrics	ACT	BIN	CMP	ECO	INST	MET	QUA	SOC
Average F1 diff	0.28	0.03	0.00	0.23	0.21	0.27	0.00	0.19
Mean char length	34	12	10	24	27	31	12	21

One can easily notice that the difference is marginal for short entities and increases with the length of entity (moreover even the ordering of average F1 difference and mean char length is same with only one exception).

One possible explanation is that models have more difficulties with determining the exact borders of entities rather than detecting the entities themselves. With short entities there is little ground for border mistakes and the scores obtained are reasonably high. With longer entities the borders become less defined and thus the performance drops.

5.2. Methods

Since one of our main goals was to replicate a business scenario, we decided not to limit the participants in their choice of methods. They were at liberty to use any available methods, including proprietary models, as well as were allowed to create additional markup in order to train their model on a larger training set (participants were asked to send the organization committee any data they annotated themselves). All top participants, however, used exclusively open-source solutions and did not create any additional training data.

We have additionally published full unannotated corpus described in [Section 3.1](#) for the purpose of fine-tuning language models on it. To our best knowledge, however, no participant attempted it.

The methods used by most participants relied on academic standards.

For NER most participants started with popular CharCNN-BLSTM-CRF baseline [\[24\]](#) and attempted to improve it mainly with the help of contextualized word embeddings such as ELMo [\[31\]](#). Two top systems are designed in essentially the same way: BERT [\[14\]](#) followed by MLP. The difference in scores between the two systems can be attributed to different BERTs used (multilingual BERT for the winner and RuBERT for the runner up) and different learning strategies.

Relation extraction allowed for better diversity of models. Several approaches were tested from simple heuristics and classical BLSTM-based approach [\[29\]](#) to once more BERT-derived pipelines. Unsurprisingly, the top two systems are both representatives of the latter category, however, unlike with NER the two systems have noticeable differences. The winner used R-BERT-inspired model [\[43\]](#). Since R-BERT reimplementation is currently SOTA on SemEval-2010 Task 8, it is a small wonder, that its adaptation works well for Russian. The runner up has successfully reduced relation extraction to sequence-labelling task and employed multi-task learning simultaneously training on both NER and RE tasks.

6. Conclusion

In this paper we have presented RuREBus corpus and shared task.

Our main goal is to bridge the gap between academic corpora and real-world scenarios. Keeping it in mind, we have obtained a corpus of governmental texts, produced by the Ministry of Economic Development of the Russian Federation and developed a markup instruction for eight entity types and eleven relation types. We also provide a large (300 million tokens) corpus of unmarked texts of the same source, intended for language model training and fine-tuning.

Our corpus consists of texts with specific and non-trivial domain (i. e. governmental texts), containing non-perfect language and other irregularities. Our entities and relations are non-balanced and their spans can often be rather long. Thus in our opinion this corpus is well-suited for being test-case "worst-case" industrial application.

We have further organized a shared task on our corpus, thus establishing a reasonable baseline for it. The participating systems (8 for NER and 5 for RE) used methods, close to SOTA on academic baselines and yet were able to score rather unimpressive 0.56 for

NER task and 0.44 for RE. Given that simultaneously happening SemEval-2020 task 11 demonstrated comparable results, we can claim that this is the current performance on "worst-case" business scenarios. Indeed, often industrial application can contain both classical entities such as persons and entities similar to the ones present in our corpus, thus providing the scores in between traditional corpora and the ones recently developed.

Thus in our opinion we created a useful testing ground for applications of NER and RE in industry. We hope that it will be useful for NLP community in general and Russian NLP in particular.

Future work directions include but are not limited to developing more advanced machine learning methods and analytical solutions, better usage of linguistic features, ensembling of different approaches and open source tools.

7. Acknowledgements

Work on maintenance of the annotation system, discussions of results, and manuscript preparation was carried out by Elena Tutubalina, Vladimir Ivanov, Tatiana Batura and supported by the Russian Science Foundation grant no. 20-11-20166. Ekaterina Artemova and Veronika Sarkisyan worked on text annotation, discussions of results, and manuscript. Their work was supported by the framework of the HSE University Basic Research Program and Russian Academic Excellence Project "5-100".

We would also like to thank students, participated in corpus annotation process, especially Anna Golenkovskaya, Polina Demina, Anastasia ..., Karolina ..., Ekaterina Sidorenko, Alexandra Pavlova.

References

1. *Alekseychuk, N. et al.*: Processing and analysis of russian strategic planning programs. In: International conference on digital transformation and global society. pp. 68–81 Springer (2019).
2. *Anh, L. T. et al.*: Application of a hybrid bi-lstm-crf model to the task of russian named entity recognition. Communications in Computer; Information Science book series—CCIS, volume 789 (2017).
3. *Anisimovich, K. et al.*: Syntactic and semantic parser based on abbyy compreno linguistic technologies. In: Computational linguistics and intellectual technologies: Proceedings of the international conference "dialog" [komp'uternaia lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii "dialog"]. pp. 90–103, Bekasovo, Russia (2012).
4. *Antonova, A. Y., Soloviev, A. N.*: Conditional random field models for the processing of russian. Communications of the ACM 56(6) (2013).
5. *Arkipov, M. et al.*: Tuning multilingual transformers for named entity recognition on slavic languages. BSNLP'2019. 89 (2019).
6. *Artemova, E. et al.*: So what's the plan? Mining strategic planning documents. In: Digital transformation and global society: Proceedings of the 5th international conference (dtgs 2020)., St. Petersburg, Russia (2020).

7. *AS, S. et al.*: FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian.
8. *Bojanowski, P. et al.*: Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*. 5, 135–146 (2017).
9. *Brykina, M. M. et al.*: Dictionary-based ambiguity resolution in Russian named entities recognition. *International Workshop on Computational Linguistics; its Applications*, ed. A. Narin'yani, v.1 (2013).
10. *Carreras, X., Màrquez, L.*: Introduction to the conll-2004 shared task: Semantic role labeling. In: *In proceedings of conll2004*. Association for Computational Linguistics, Boston, MA (2004).
11. *Chiticariu, L. et al.*: Rule-based information extraction is dead! Long live rule-based information extraction systems! In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. pp. 827–832 (2013).
12. *Craidlin, L.*: Program of allocation of Russian individualized nominal groups taglite. *Computational linguistics; intellectual technologies Dialog* (2005).
13. *Da San Martino, G. et al.*: SemEval-2020 task 11: Detection of propaganda techniques in news articles. In: *Proceedings of the 14th international workshop on semantic evaluation.*, Barcelona, Spain (2020).
14. *Devlin, J. et al.*: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)*. pp. 4171–4186 (2019).
15. *Feldman, R., Rosenfeld, B.*: Boosting unsupervised relation extraction by using ner. In: *Proceedings of the 2006 conference on empirical methods in natural language processing*. pp. 473–481 (2006).
16. *Gareev, R. et al.*: Introducing baselines for Russian named entity recognition. *Computational Linguistics; Intelligent Text Processing* (2013).
17. *Grishman, R., Sundheim, B.*: Design of the MUC-6 evaluation. In: *Proceedings of the 6th conference on message understanding*. pp. 1–11 Association for Computational Linguistics, USA (1995).
18. *Hendrickx, I. et al.*: SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: *Proceedings of the 5th international workshop on semantic evaluation*. pp. 33–38 Association for Computational Linguistics, Uppsala, Sweden (2010).
19. *Hovy, E. et al.*: OntoNotes: The 90. In: *Proceedings of the human language technology conference of the NAACL, companion volume: Short papers*. pp. 57–60 Association for Computational Linguistics, USA (2006).
20. *Huang, Z. et al.*: Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*. (2015).
21. *Kutuzov, A., Kuzmenko, E.*: WebVectors: A toolkit for building web interfaces for vector semantic models. In: *Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, AIST 2016, Yekaterinburg, Russia, April 7–9, 2016, revised selected papers*. pp. 155–161 Springer International Publishing, Cham (2017).

22. *Kuznetsov, A. et al.*: Family matters: Company relations extraction from wikipedia. In: International conference on knowledge engineering and the semantic web. pp. 81–92 Springer (2016).
23. *Lafferty, J. et al.*: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning (2001).
24. *Lample, G. et al.*: Neural architectures for named entity recognition. In: Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies. pp. 260–270 Association for Computational Linguistics, San Diego, California (2016).
25. *Li, X. et al.*: A unified mrc framework for named entity recognition. ArXiv. abs/1910.11476, (2019).
26. *Ma, X., Hovy, E.*: End-to-end sequence labeling via bi-directional lstm-cnns-crf, (2016).
27. *Malykh, V., Lyalin, V.*: Named entity recognition in noisy domains. In: 2018 international conference on artificial intelligence applications and innovations (ic-ai). pp. 60–65 IEEE.
28. *Malykh, V., Ozerin, A.*: Reproducing russian ner baseline quality without additional data. In: CDUD@ cla. pp. 54–59 (2016).
29. *Miwa, M., Bansal, M.*: End-to-end relation extraction using lstms on sequences and tree structures. 2016. arXiv preprint arXiv:1601.00770.
30. *Mozharova, V., Loukachevitch, N.*: Two-stage approach in russian named entity recognition. In: Intelligence, social media and web (ismw fruct), 2016 international fruct conference on. pp. 1–6 IEEE (2016).
31. *Peters, M. E. et al.*: Deep contextualized word representations. In: Proceedings of naacl-hlt. pp. 2227–2237 (2018).
32. *Piskorski, J. et al.*: The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In: Proceedings of the 7th workshop on balto-slavic natural language processing. pp. 63–74 Association for Computational Linguistics, Florence, Italy (2019).
33. *Piskorski, J. et al.*: The first cross-lingual challenge on recognition, normalization, and matching of named entities in slavic languages. In: Proceedings of the 6th workshop on balto-slavic natural language processing. pp. 76–85 (2017).
34. *Podobryaev, A. V.*: Searching for person memories in news texts with the use of a model of conditional random fields. RCDL (2013).
35. *Popov, B. et al.*: KIM—a semantic platform for information extraction and retrieval. Journal of Natural Language Engineering 10 (2004).
36. *Shen, Y. et al.*: Deep active learning for named entity recognition. CoRR. abs/1707.05928, (2017).
37. *Soares, L. B. et al.*: Matching the blanks: Distributional similarity for relation learning. In: ACL. (2019).
38. *Stenetorp, P. et al.*: BRAT: A web-based tool for nlp-assisted text annotation. In: Proceedings of the demonstrations at the 13th conference of the european chapter of the association for computational linguistics. pp. 102–107 Association for Computational Linguistics (2012).

39. *Tjong Kim Sang, E. F., De Meulder, F.*: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the seventh conference on natural language learning at hlt-naacl 2003—volume 4. pp. 142–147 Association for Computational Linguistics, USA (2003).
40. *Trofimov, I.*: Identification of personal names in news texts on collections persons-1000/1111-f (in russian). Proceedings of RCDL-2014. 217–221 (2014).
41. *Tsygankova, T. et al.*: BSNLP2019 shared task submission: Multisource neural ner transfer. In: Proceedings of the 7th workshop on balto-slavic natural language processing. pp. 75–82 (2019).
42. *Vlasova, N. et al.*: Report on russian corpus for personal name retrieval. Proceedings of computational; cognitive linguistics TEL (2014).
43. *Wu, S., He, Y.*: Enriching pre-trained language model with entity information for relation classification. In: Proceedings of the 28th acm international conference on information and knowledge management. pp. 2361–2364 (2019).
44. *Zhang, Y. et al.*: Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 conference on empirical methods in natural language processing (emnlp 2017). pp. 35–45 (2017).
45. *Zhao, Y. et al.*: Improving relation classification by entity pair graph. In: ACML. (2019).
46. *Zuev K. A., J. M. V., Indenbom M. E.*: Statistical machine translation with linguistic language model. In: Computational linguistics and intellectual technologies: Proceedings of the international conference “dialog” [komp’iuternaia lingvistika i intellektual’nye tehnologii: Trudy mezhdunarodnoj konferentsii “dialog”]. pp. 164–172, Bekasovo, Russia (2013).
47. *The ace 2005 (ace05) evaluation plan.* (2005).

COMPARATIVE ANALYSIS OF RHETORICAL AND ARGUMENTATIVE STRUCTURES IN THE STUDY OF POPULAR SCIENCE DISCOURSE

Kononenko I. S. (irina_k@cn.ru),
Sidorova E. A. (Isidorova@iis.nsk.su),
Akhmadeeva I. R. (ah.irishka@gmail.com)

A. P. Ershov Institute of Informatics Systems SB RAS,
Novosibirsk, Russia

The proposed work is performed as a part of an on-going research project aimed at creation of discourse annotated corpus of popular science texts written in Russian. Annotation is carried out within the framework of a multi-level model of discourse, which considers the text from the perspective of genre, rhetorical and argumentative organization. We conduct a comparative study of the rhetorical and argument annotations, discuss their similarities and differences on the segment and structural levels and show them on the examples of standard schemes of reasoning described in D. Walton's theory of structured argumentation: "Argument from Expert Opinion", "Argument from Example", and "Argument from Cause to Effect". Special attention is paid to discourse markers registered during annotation as key indicators of discourse structure. We report the results of the experiment with argument indicator patterns, based on the list of rhetorical markers, and aimed at the extraction of "from Expert Opinion" arguments.

Key words: argument mining, argument indicator, indicator pattern, argumentation scheme, discourse annotation, corpus linguistics

DOI: 10.28995/2075-7182-2020-19-432-444

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РИТОРИЧЕСКИХ И АРГУМЕНТАТИВНЫХ СТРУКТУР ПРИ ИССЛЕДОВАНИИ НАУЧНО-ПОПУЛЯРНОГО ДИСКУРСА

Кононенко И. С. (irina_k@cn.ru),
Сидорова Е. А. (Isidorova@iis.nsk.su),
Ахмадеева И. Р. (ah.irishka@gmail.com)

Институт систем информатики им. А. П. Ершова
ИСИ СО РАН, Новосибирск, Россия

В статье обсуждаются промежуточные результаты работы по созданию дискурсивно аннотированного корпуса научно-популярных текстов. Разметка осуществляется на основе многоуровневой модели представления дискурса, которая рассматривает текст с позиции жанровой, риторической и аргументативной организации. Проводится сравнительное исследование риторической и аргументативной аннотаций на примере фрагментов, представляющих типовые схемы рассуждения, основанные на теории Д. Уолтона. В рамках проводимого эксперимента на основе списка риторических маркеров разработан словарь шаблонов для извлечения схем аргументации «от эксперта» и проведено тестирование индикаторного подхода для анализа аргументации.

Ключевые слова: извлечение аргументации, индикатор аргументации, шаблон индикатора, схема аргумента, аннотирование дискурса, корпусная лингвистика

1. Introduction

The study of discourse involves a description of its structure in the form of related discourse units. One of the most famous models applied to this task is the Rhetorical Structure Theory (RST) and its modifications [6], [7], [8].

A particular role for discourse study is played by argumentation presented by the speaker or the author to convince the audience in her position. The theory of argumentation is growing in the framework of logic and philosophy, history and sociology, linguistics and psychology, computer science and artificial intelligence [4]. Over the past decade argument mining has been actively developing to solve the task of automatic extracting from text a sequence of statements (premises) that lead to a certain conclusion (thesis).

When analyzing argumentation presented in the text, it is necessary not only to extract arguments and chains of arguments that confirm or refute a certain thesis, but also to study the structure of each argument, its role and significance for the whole argumentation. The internal relations of the argument often coincide or can be built on the basis of the rhetorical relations of the RST model. To describe the various methods of reasoning, models or schemes of arguments are used [13], [19]. The most famous description of structured argumentation is the model of D. Walton [21], which defines about 60 basic argumentation schemes, each being the description of a specific reasoning pattern (inference form expressing the relations of premises and conclusion). The model has been used in a number of applications and tools for argument analysis and corpora annotation (OVA [5], Carneades [3], ArgDF).

So far, there exist only a few resources with annotated argumentative structures mainly over monologue texts in English. The best known is AIFdb, the former Araucaria corpus [14], which includes news articles, records of online debates. Resources are created in German: University of Darmstadt Corpus includes subcorpora of student essays [15], news texts and scientific articles; the Potsdam corpus contains a small set of microtexts on a given topic [10]. There exist projects for some other languages (Italian, Greek, Chinese). As for the Russian language, such resources, as far as we know,

do not yet exist. In most cases, corpus annotation includes text segmentation with highlighting of argument units, markup of roles (premise, conclusion) and relations (support, attack), without matching the argumentation schemes on which the reasoning is based. An exception is Araucaria, where annotation of argumentative structure is related to particular argumentation scheme based on the theory of Walton [22].

In the past few years, studies have emerged in the field of automatic argument analysis that consider the potential use of existing text corpora with annotated discourse structure to facilitate the annotation of components and argumentative relations. The idea is to create a corpus of texts with multi-level annotation: along with the existing markup of rhetorical units and relations, texts are marked up with respect to the argument structure. In the presence of such a corpus, the task is to establish the relationship between rhetorical and argumentative structure of the text, the correlation of the components and relations between them in order to use existing resources and tools of discourse analysis to extract arguments. So, [9] describe the development and use in experiments of a two-level corpus of 112 short texts written in the genre of argumentative essays, and in [1], the material for annotation are scientific articles from the field of computer linguistics.

An important linguistic aspect of the process of discourse annotation is registration of discourse markers (discourse connectives)—language tools for structuring discourse, which play a key role in the process of understanding. Thus, argument indicators constitute keystones in the discourse, facilitating the identification and reconstruction of argumentative moves that are made in argumentative discussions and texts (see [20]). Argument indicators are language means (words, constructions) that serve as discourse clues in identifying the structure of argumentation: they help determine the presence of arguments in a given segment of text, reconstruct the connections between statements, relate the argument to a specific reasoning scheme.

The aim of this work is to study the correlations between rhetorical and argumentative structures in popular science discourse. The main research tools are annotated text corpus and dictionaries of indicators of rhetorical and argumentative relations.

2. Annotation model

The proposed work was performed as part of an on-going research project aimed at creation of discourse annotated corpus of popular science texts written in Russian. It includes 68 articles on linguistics and 11 texts taken from the open corpus of RuRSTreebank [11], [17], [18]. Popular science discourse is defined as a way of transmitting scientific knowledge or innovation projects by the author-scientist (or a journalist as an intermediary) for their understanding by a mass audience.

Three levels are distinguished in the structure of the discourse, of which the first two correspond to the superstructure and relational structure in [6], [8].

1. **Genre structure** is responsible for the compositional and semantic organization of text at the highest level and depends on the text genre membership. This level involves a breakdown into meaningful compositional parts, such as chapters and paragraphs.

2. **Rhetorical structure** is responsible for organizing the text itself, transforming it from a simple sequence of formal segments into a single whole. It reflects the functional relationships existing between segments, called rhetorical relations.
3. **Argumentative structure:** represents the text as a means of reproducing the process of argumentation, highlighting the components of the argument field and the relationships between them (controversial thesis, arguments for or against). Argumentation is putting forward arguments in order to change or form some belief (position) of the other side [4].

To model and mark up the arguments, we used standard schemes (models) of reasoning from the Walton's compendium [22]. Consider a typical scheme of reasoning "Argument from Expert Opinion":

Major Premise: Source E is an expert in subject domain S containing proposition A
Minor Premise: E asserts that proposition A is true (false)
Conclusion: A is true (false)

Subject-independent concepts in this scheme are the source of information E (it can be either a person, or, for example, an article in a journal), subject area S and statement A. They are related by subject-independent relations *be_expert* (*E, S, true*), *belong_to* (*A, S, true*) and *assert* (*E, A, true*), *be_true* (*A, true / false*).

In addition to the above three types of annotation (genre, rhetorical and argumentative), another one is important—that concerning indicators of rhetorical relations or aspects of argumentation. When annotating arguments, the indicators are marked up at the same time, which allows us to automatically build a dictionary of indicators.

3. Study of correlations between rhetorical and argument annotation

For the comparative study, a subcorpus of 11 rhetorically annotated texts was taken from the Ru-RSTreebank and provided with argument annotation.

3.1. Segment-level comparison

The basis of both rhetorical and argument annotation is segmentation into elementary discourse units (EDUs). These units are sentences, clauses or minimal text spans which have propositional content including nominalized propositions and prepositional phrases with the meaning of cause, effect, concession, contrast (for example EDU 32 in Fig. 1b).

In the rhetorical annotation process EDUs are combined into larger DUs which can be used in rhetorical relations alongside with EDUs. This iterative process provides the construction of RST trees.

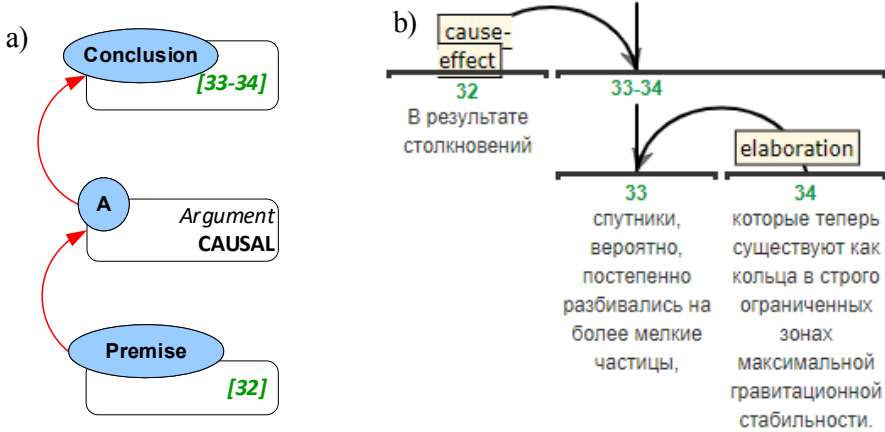
The argument annotation process is different.

Firstly, there is no unit enlargement procedure since the semantics of the generated graph chains is associated with meaningful aspects of the argumentative relations, and not with the structural ones, as it happens in the process of segment merging in rhetorical structures.

Secondly, despite the fact that a definition of argumentation units (ADUs) is based on discourse ones, i.e. ADU is an EDU representing a separate (independent) component of the argument, however, there is no one-to-one correspondence:

1. EDUs not relevant to the argumentation are ignored. For example, spans of text containing introductory or factual information not used as a premise.

2. Adjacent EDUs are combined into one ADU if they do not represent independent components of the argument: in Fig. 1 segments 33 and 34 connected by *Elaboration* relation are combined.



Text: As a result of collisions [32], the satellites were likely to gradually break into smaller particles [33], which now exist as rings in strictly limited zones of maximum gravitational stability [34].

Fig. 1. Comparison of segmentation with argument (a) and rhetorical (b) annotation

A similar merge is possible for segments connected by *Joint* relation (segments 102–103 in Fig. 2).

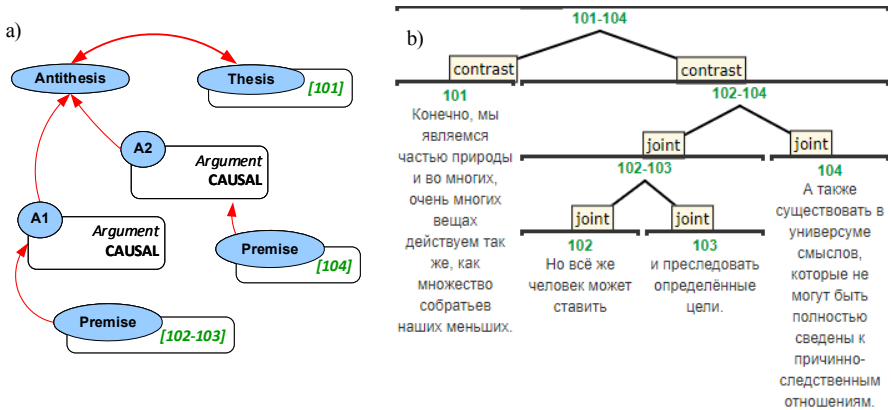


Fig. 2. Comparison of segmentation with argument (a) and rhetorical (b) annotation

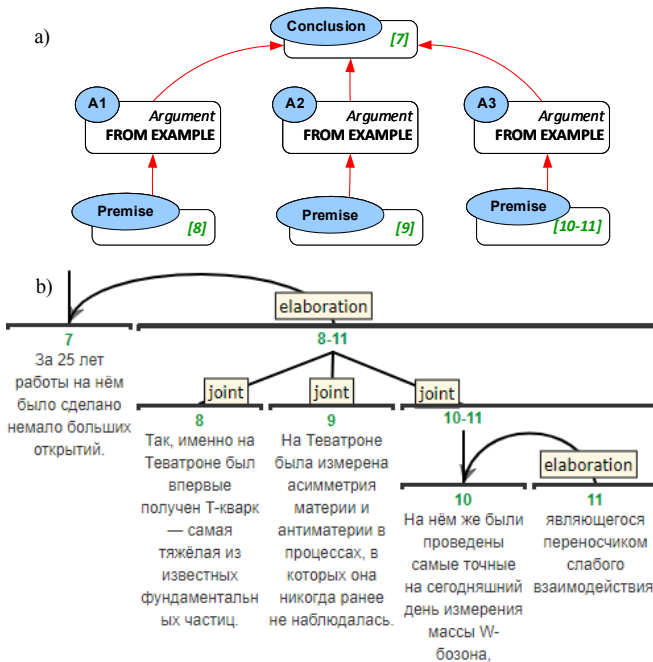
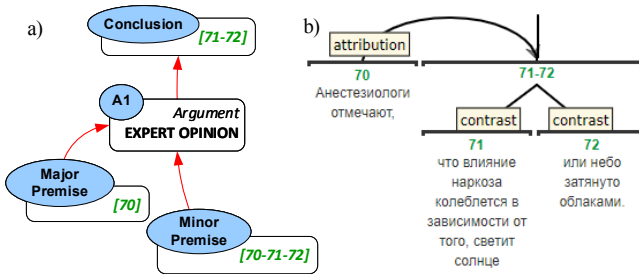


Fig. 3. Comparison of segmentation with argument (a) and rhetorical (b) annotation

In contrast to the latter, segments 8–11 connected by *Joint* relation in Fig. 3 are not united, since each member of the *Joint* corresponds to an independent argument that implements the argumentation scheme “From the Example”.

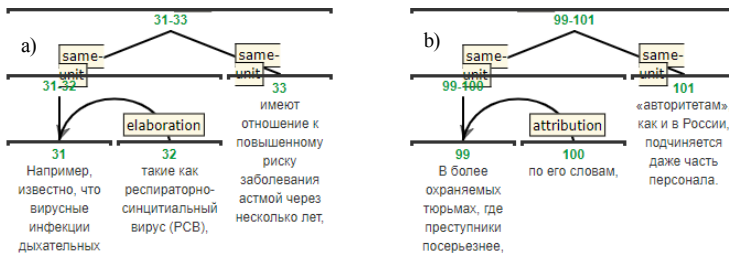
3. The correspondence between rhetorical and argumentation units is also broken in case of arguments appealing to the Knower, in particular, to the Expert (see Fig. 4). In the rhetorical structure there are two adjacent EDUs connected by the *Attribution* relation. In the argumentative structure the thesis segment is nested in the minor premise segment and this premise is the result of merge of adjacent EDUs—according to the “Argument from Expert Opinion” scheme presented in section 1.



Text: Anesthetists note [70] that the effect of anesthesia varies depending on whether the sun is shining [71] or the sky is cloudy [72].

Fig. 4. Comparison of segmentation with argument (a) and rhetorical (b) annotation

4. Two EDUs connected by the *Same_unit* relation are combined in one ADU to represent a discontinuous segment: segments 31, 33 (Fig. 5a) and segments 99, 101 (Fig. 5b).



Text: For example, viral infections of the respiratory tract in young children [31], such as the human respiratory syncytial virus (HRSV)[32], are known to be associated with an increased risk of asthma after a few years [33]

Text: In more secure prisons, where criminals are more serious [99], according to him [100], even part of the staff obeys the «authorities», as in Russia [101].

Fig. 5. Merging segments into one ADU to represent discontinuous statements

5. An EDU containing homogeneous groups is split up if the recovered propositions are considered to correspond to independent arguments (segments 102 and 103 in Fig. 2).

And finally, the treatment of indicators of discourse relations is different. For example, in rhetorical annotation indicators are always included in the DUs, while most of argument indicators (*так как* 'because'; *то, что* 'the fact that'; *следовательно* 'therefore', etc.) are excluded from ADU as they do not belong to the propositional content of ADU that appears in the role of premise or conclusion of the argument. Thus, the boundaries of text fragments acting as discourse units are different.

3.2. Comparison of rhetorical and argument indicators

In the Penn Discourse Treebank (PDTB) discourse connectives are considered as binary predicates that express certain semantic relationships between sentences, events and states [12]. In the RST, subject matter relations are opposed to presentational relations whose intended effect is to increase some inclination in the reader, such as the desire to act or the degree of positive regard for, belief in, or acceptance of the nucleus of the rhetorical relation. Considering the linguistic theory of argumentation of Ancombe and Ducro [2], in which two classes of argument markers (connectors and operators) are distinguished, as well as the concept of presentational rhetorical relation, it can be assumed that in argumentation (as a special case of discourse sequence) a representative subset of rhetorical discourse markers is used as argument indicators.

Heterogeneous set of discourse connectives is usually considered to be indicators of discourse structure. It traditionally includes non-significant lexical units, or functional words and phrases: subordinating conjunctions, coordinating conjunctions, adverbs, prepositional phrases, parenthetical words, particles, etc. Recent studies of discourse indicators [16], [17] take into account not only traditional discourse connectives, but also grammatical time, punctuation marks and their combinations, as well as content words and constructions based on them [18].

Connectors that express presentational rhetorical relations signal semantic-ontological relation between statements and indicate the following pragmatic aspects of argumentation:

- strength of the argument (*по-видимому* 'seemingly');
- inference relation between two statements (*следовательно* 'therefore');
- role of the statement in the inference relation: Premise (*поскольку* 'since') vs. Conclusion (*поэтому* 'for that reason');
- type of argumentative relation: Support (*например* 'for example') vs. Attack (*однако* 'however');
- structural type of argumentation: Multiple (*тем более что* 'all the more so that') vs. Serial argumentation (*в подтверждение этому* 'in confirmation of this').

As for operators (particles such as *именно* 'just', *только* 'only', *еще* 'yet', *почти* 'almost', *по крайней мере* 'at least', etc.), they usually don't signal about rhetorical relations (as there is no relation of two explicit propositions), but clearly give an argumentative orientation to the statement:

- (1) [*Быстрее всего обновляется лексический фонд*], так как **только** [он непосредственно связан с материальным миром]. Диалектная фонетика, грамматика и синтаксис сохраняются очень хорошо.

[The lexical fund is updated most quickly], since **only** [it is directly connected with the material world]. Dialectal phonetics, grammar and syntax persist very well.

Annotators of the Ru-RSTreebank took account of rhetorical markers and compiled their lists, which include not only functional words, but also content words and constructions with them: *причина стать* ‘reason to become’; *являться результатом* ‘be the result’; *как стать известно* ‘as (it) become known’; *намекать на то, что* ‘hint at the fact that’. However, words and phrases in the list set too general contexts. For example, there are markers formed on the basis of the speech verb *говорить* ‘speak/tell/ say’: *Х говорит о Y* ‘X talks/speaks about Y’; *говори* ‘speak/tell/say’; *говорить о том, что* ‘to show/suggest that/ be indicative of’; *тот факт, что ... говорит о том* ‘the fact that...shows that’; *иначе говоря* ‘in other words’; *не говоря уже о* ‘to say nothing of’; *точнее говоря* ‘to be more exact’. As analysis shows, the consideration of speech verbs as indicators of the *Attribution* rhetorical relation gives in case of the verb *говорить* and some other verbs (*подтверждать* ‘confirm’, *признавать* ‘recognize’) up to 70 percent of noise:

- (2) *По-разному говорят представители различных профессий, люди разных возрастов и интересов, жители разных городов.*
Representatives of various professions, people of different ages and interests, residents of different cities speak differently.

The specification of the indicator in the form of a pattern with the core speech verb and a sentential actant marker *что* ‘that’ or a citation marker for direct speech allows to increase the precision in identifying the *Attribution* relation:

- (3) *Гениальный дипломат Шарль Морис Талейран говорил, что язык нам дан для того, чтобы скрывать наши мысли.*
The brilliant diplomat Charles Maurice Talleyrand said that language was given to us in order to hide our thoughts.

To recognize such constructions, the following pattern is appropriate:

speech_activity_that = [<Sem:speech_activity>, s/, , *что* ‘that’]

A more precise definition of this pattern with addition of the lexical class *Expert* as a source of information allows us to present the indicator of the “from Expert Opinion” argument scheme:

[begin: <Sem:expert>, end: speech_activity_that]

3.3. Experimental study of argument extraction by indicator patterns

An experimental study of indicator patterns was carried out using training and test collections. The training collection consisted of 68 popular science articles on linguistics (more than 64 thousand word uses). Based on a dictionary of rhetorical markers and annotated corpus, patterns for extracting arguments were developed. The final dictionary of indicators included constructions for cases of expressing an authoritative opinion (“Argument from Expert Opinion”), presented as literal quotation (direct speech) or indirectly (using speech or mental predicates with sentential actants). The

lexical dictionary contained 114 units (90 nouns and 24 noun phrases) of the expert class, as well as 78 speech, 14 mental and 9 predicates of intellectual activity.

The evaluation of the quality of extracting arguments by the indicator method was carried out on the test collection, which included 877 popular science articles taken from open sources (about 2,880 thousand word uses). Table 1 presents experiment results.

Table 1. Evaluation of the quality of argument extraction

	Training collection	Test collection
Total	275	2,694
Incorrect (0)	10	509
No arguments (1)	58	414
Conditionally correct (2)	36	528
Correct (3)	168	1,243
Precision	74.18%	65.73%

In the evaluation, the following scale was used:

- 0 — the extracted text fragment does not match target construction;
- 1 — fragment contains no argument;
- 2 — target construction is matched but the presence of argument requires clarification in a wider context;
- 3 — target construction and argument are extracted.

When evaluating precision, answers scored 2 and 3 were considered correct.

The analysis of the experiment results revealed the following reasons for incorrect identification of arguments:

- a) lack of structural correspondence due to syntactic ambiguity or punctuation marks usage (dash, quotes);
 - b) lexical ambiguity: actualization of the non-speech meaning of the predicate (the speech one is often non-primary, e.g. *продолжить* 'to continue', *говорить* 'to indicate'), homonymy of the type "science-specialist" (*статистика-статистик* 'statistics-statistician');
 - c) no proposition in the actant of the predicate (actant position filled by the object)
- (4) *Исследователь жизни поэта учился в московском Университете дружбы народов и написал книгу о Пушкине*
Researcher of the poet's life studied at the Peoples' Friendship University in Moscow and wrote a book about Pushkin

or incomplete proposition in the actant (indirect question)

- (5) *О том, как удержать баланс между изменчивостью и сохранением идентичности, и кто играет, рассказал Михаил Штудинер, кандидат фил. наук,*
How to maintain a balance between variability and preservation of identity, and who plays, said Mikhail Shtudiner, Ph.D.;

d) lack of reference annotation for events and authoritative persons (experts) in case they are referred to by name or by anaphoric pronoun

- (6) “*Ngram Viewer позволяет получить статистику любому желающему и проводить собственные исследования*”,—*рассказывает Валерий Соловьев.* “*Ngram Viewer allows you to get statistics for anyone and conduct their own research*” says Valery Soloviev.

4. Conclusion

In the paper we presented an approach to modeling and studying the arguments found in popular science literature. The approach is based on the comparative analysis of the rhetorical and argumentative discourse structures. A large number of correlations makes it possible to apply recent developments in the field of rhetorical structures to the analysis of argumentation, in particular, to the construction of a dictionary of argument indicators.

In the course of the experiment based on the list of rhetorical markers, a dictionary of indicator patterns was developed to extract typical reasoning schemes “from Expert Opinion”, and an indicator approach was tested for analysis of argumentation.

Further research focuses on the comparison of argumentative and rhetorical structures with the view of revealing correspondences between argumentative chains and rhetorical trees.

5. Acknowledgment

The research has been supported by Russian Foundation for Basic Research (Grant No. 18-00-01376 (18-00-00889)).

References

1. *Accuosto P., Saggion H.* (2019), Discourse-driven argument mining in scientific abstracts, 24th International Conference on Applications of Natural Language to Information Systems NLDB 2019, Salford, UK, pp. 182–194.
2. *Anscombre J.-C., Ducrot O.* (1983), *Argumentation in the language [L'argumentation dans la langue]*, Brussels.
3. *Gordon T. F., Walton D.* (2006), The Carneades argumentation framework—using presumptions and exceptions to model critical questions, Proc. of the 2006 Conf. on Computational Models of Argument: Proceedings of COMMA 2006, Amsterdam, Vol. 6, pp.195–207.
4. *Ivin A. A.* (2017), *Argumentation in the processes of communication. Pro et contra [Argumentatsiya v protsessakh kommunikatsii. Pro et contra]*, Prospect, Moscow.
5. *Janier M., Lawrence J., Reed C.* (2014), OVA+: An argument analysis interface, Computational Models of Argument: Proc. of COMMA 2014, Amsterdam, Vol. 266, pp. 463–464.

6. *Litvinenko A. L.* (2001), Description of the structure of discourse in the context of the theory of Rhetorical Structure: application in Russian material [Opisanie struktury diskursa v ramkakh Teorii Ritoricheskoy Struktury: primeneniye na russkom materiale], Computational Linguistics and its Applications: Proceedings of the International Seminar “Dialog 2001” [Komp’yuternaya Lingvistika i yeye prilozheniya: Trudy Mezhdunarodnogo Seminara “Dialog 2001”], Aksakovo, pp. 159–168.
7. *Mann W. C., Thompson S. A.* (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text, 1988, 8(3).
8. *Mann W., Matthiessen C., Thompson S. A.* (1992) Rhetorical structure theory and text analysis, Discourse Description, Amsterdam, Benjamins, pp. 39–78.
9. *Musi E., Alhindi T., Stede M., Kriese L., Muresan S., Rocci A.* (2018), A Multi-layer Annotated Corpus of Argumentative Text: From Argument Schemes to Discourse Relations, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC’2018), Miyazaki, Japan, pp.1629–1636.
10. *Peldszus A., Stede M.* (2016), An annotated corpus of argumentative microtexts, Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, London, College Publications, Vol. 2, pp. 801–816.
11. *Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A.* (2017), Towards building a Discourse-annotated corpus of Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017”, Moscow, pp. 194–204.
12. *Prasad R., Miltsakaki E., Dinesh N., Lee A., Joshi A., Robaldo L., Webber B.* (2007), The Penn Discourse Treebank 2.0 Annotation Manual, Technical Report 203, Institute for Research in Cognitive Science, University of Pennsylvania.
13. *Reed C., Walton D.* (2003), Argumentation schemes in argument-as-process and argument-as-product, Proc. of Conference Celebrating Informal Logic, OSSA Conference Archive 75, University of Windsor, Windsor, Ontario.
14. *Reed C., Rowe G.* (2004), Araucaria: Software for argument analysis, diagramming and representation, International Journal on Artificial Intelligence Tools, Vol. 13(4), pp. 961–979.
15. *Stab C., Gurevych I.* (2014), Identifying Argumentative Discourse Structures in Persuasive Essay, Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 46–56.
16. *Taboada M., Das D.* (2013), Annotation upon Annotation: Adding Signalling Information to a Corpus of Discourse Relations, Dialogue and Discourse, 4(2), pp. 249–281.
17. *Toldova S., Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A.* (2017), Rhetorical relations markers in Russian RST Treebank, Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms. Santiago de Compostela, Spain, pp. 29–33.
18. *Toldova S., Pisarevskaya D., Vasilyeva M., Kobozeva M.* (2018), The cues for rhetorical relations in Russian: “Cause-Effect” relation in Russian Rhetorical Structure Treebank, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, pp. 747–761.

19. *Toulmin S.* (2003) *The Uses of Argument*, Cambridge University Press, Cambridge.
20. *Van Eemeren F. H., Houtlosser P., and F. Snoeck Henkemans* (2007), *Argumentative Indicators in Discourse: A Pragma-Dialectical Study*, Dordrecht, Springer.
21. *Walton D.* (2009) *Argumentation theory: A very short introduction*, *Argumentation in Artificial Intelligence*, Springer, Boston, pp. 1–22.
22. *Walton D., Reed C., Macagno F.* (2008) *Argumentation schemes*, Cambridge University Press.

EXPLORING THE BERT CROSS-LINGUAL TRANSFER FOR READING COMPREHENSION

Kononov V. P. (vaskoncv@phystech.edu)[†],
Gulyaev P. A. (guliaev.pa@phystech.edu)[†],
Sorokin A. A. (alexey.sorokin@list.ru)^{†,‡},
Kuratov Y. M. (yurii.kuratov@phystech.edu)[†],
Burtsev M. S. (burtcev.ms@mipt.ru)[†]

[†]Moscow Institute of Physics and Technology, Dolgoprudny, Russia

[‡]Moscow State University, Moscow, Russia

Multilingual BERT has been shown to generalize well in a zero-shot cross-lingual setting. This generalization was measured on POS and NER tasks. We explore the multilingual BERT cross-language transferability on the reading comprehension task. We compare different modes of training of question-answering model for a non-English language using both English and language-specific data. We demonstrate that the model based on multilingual BERT is slightly behind the monolingual BERT-based on Russian data, however, it achieves comparable results with the language-specific variant on Chinese. We also show that training jointly on English data and additional 10,000 monolingual samples allows it to reach the performance comparable to the one trained on monolingual data only.

Keywords: BERT, SQuAD, Question Answering, QA, DeepPavlov

DOI: 10.28995/2075-7182-2020-19-445-453

ИССЛЕДОВАНИЕ КРОСС-ЯЗЫКОВОГО ПЕРЕНОСА МОДЕЛЕЙ ДЛЯ ЗАДАЧИ ОТВЕТА НА ВОПРОСЫ ПО ТЕКСТУ

Коновалов В. П. (vaskoncv@phystech.edu)[†],
Гуляев П. А. (guliaev.pa@phystech.edu)[†],
Сорокин А. А. (alexey.sorokin@list.ru)^{†,‡},
Куратов Ю. М. (yurii.kuratov@phystech.edu)[†],
Бурцев М. С. (burtcev.ms@mipt.ru)[†]

[†]Московский Физико-Технический Институт,
Долгопрудный, Россия

[‡]Московский Государственный Университет
им. М. В. Ломоносова, Москва, Россия

В работе исследуются разные способы выбора данных и модели в задаче обучения вопросно-ответных систем для языка, отличного от английского. Мы показываем, что моноязычный и мультязычный энкодер приводят к сравнимым результатам при обучении только на данных этого языка. При этом более выгодной стратегией является предварительное обучение мультязычного энкодера на англоязычных данных с последующей настройкой на данных конкретного языка, поскольку в этом случае требуется значительно меньше обучающих примеров для достижения сопоставимого качества.

Ключевые слова: BERT, SQuAD, Question Answering, QA, DeepPavlov

1. Introduction

Bidirectional Encoder Representations from Transformers (BERT) are the language representation model. Unlike the other models, BERT is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, including text classification, sequence classification, and question answering, without significant task-specific architecture modifications [4]. Google AI released multiple versions of BERT, such as multilingual BERT (M-BERT), English BERT, and Chinese BERT [6]. M-BERT provides zero-shot cross-lingual model transferability, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. The experiments on NER (Named Entity Recognition) and POS (Part-of-speech Tagging) tasks show that while the high lexical overlap between languages improves transfer, M-BERT is also able to transfer between languages having almost no lexical overlap indicating that it captures multilingual representations [10].

In this paper, we measure the cross-lingual model transferability of M-BERT on the reading comprehension task (SQuAD-like datasets) for three languages with small or zero lexical overlap: English, Russian, Chinese. Reading Comprehension is an important task for language understanding, also, it is less susceptible to annotation artifacts found in other datasets [7].

Pre-training BERT for an additional language is a fairly expensive process (four days on 4 to 16 Cloud TPUs) [6]. We show that at least for the Question Answering task this stage is not always necessary, since the obtained performance gain over M-BERT is only marginal, if any. Collecting SQuAD-like datasets takes tremendous efforts and significant funding. Consequently, such datasets rarely exist in languages other than English, making training QA systems in other languages challenging [9]. An alternative to building large monolingual training datasets is to develop cross-lingual systems that can transfer to a target language without requiring training data in that language. We found that M-BERT based model trained jointly on widely available English data and a number of the language-specific monolingual training samples achieves promising results and falls behind the model trained entirely on the language-specific data by a small margin.

Our contributions in this paper are as follows: (i) we show that M-BERT transferability allows us to achieve promising results almost on par with language-specific BERT while both are trained on the language-specific datasets; (ii) even further we show that M-BERT based model jointly trained on English SQuAD with a number of language-specific monolingual samples falls behind the model trained entirely on the language-specific data by a small margin.

We use the **DeepPavlov** framework as our evaluation testbed. **DeepPavlov** is an open-source library for deep learning end-to-end dialog systems and chatbots [2].

2. Related Work

The M-BERT model was introduced in the original BERT paper [4]. It was trained on Wikipedia for 104 languages and uses a shared cross-language BPE vocabulary. Its cross-lingual transferability was demonstrated on sequential tasks like NER and POS. The NER model performance was examined on two publicly available datasets CoNLL-2002 and CoNLL-2003, containing Dutch, Spanish, English, and German. The results showed that the M-BERT performance drops insignificantly when tested on the language that did not participated in the train. The POS experiments were performed by using Universal Dependencies (UD) data for 41 languages. It was shown that M-BERT generalizes well across languages, achieving over 80% accuracy for all pairs. Also, the authors claim that M-BERT's performance is flat for a wide range of overlaps, and even for language pairs with almost no lexical overlap, however, they tested it only on sequential tasks [10].

Moreover, M-BERT was found to be competitive with the state-of-the-art methods for zero-shot cross-lingual transfer on five NLP tasks: natural language inference (XNLI) [3], document classification, NER, POS tagging, and dependency parsing for 39 languages from various language families. Across all five tasks, M-BERT achieved high zero-shot cross-lingual performance without any cross-lingual signal. It outperforms cross-lingual embeddings in four tasks except for XNLI [14].

The previous works came to a consensus that the cross-lingual generalization ability of M-BERT is based on three factors: (i) shared vocabulary items; (ii) joint training across multiple languages (iii) deep cross-lingual representations that generalize across languages and tasks. However, it was shown that the state-of-the-art multilingual representation learning models and a monolingual model that is transferred to new languages at the lexical level perform comparably while tested on standard benchmarks: natural language inference (XNLI dataset), document classification (MLDoc)[12], paraphrase classification (PAWS-X)[15] and question answering (XQuAD)[1]. The authors also showed that a monolingual model trained on a particular language learns some semantic abstractions that are generalizable to other languages[1].

3. Datasets

We measure M-BERT cross-lingual transferability on Reading Comprehension task for three lexical non-overlapping languages: English, Russian, Chinese.

English The Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage (context). SQuAD contains 107,785 question-answer pairs on 536 articles, and is almost two orders of magnitude larger than the previous manually labeled RC dataset [11].

Russia SberQuAD is a Russian reading comprehension dataset analogous to Stanford SQuAD. SberQuAD was collected strictly following the SQuAD’s annotations guidelines, which resulted in the high lexical overlap between questions and sentences with answers. We use the same train/test split as in the original paper (45,328/5,036) [5].

Chinese DRCD (Delta Reading Comprehension Dataset) is an open domain traditional Chinese machine reading comprehension dataset. It is based on 2,108 Wikipedia articles [13].

4. Experimental Setup

Our implementation is based on the original model architecture described in [4] and implemented in the **DeepPavlov** framework [2], depicted in **Figure 1**.

We compare QA models’ performance based on multilingual BERT (M-BERT) and language-specific BERT (Russian BERT, Chinese BERT). For our models, we use only the **Base** model configuration with 12 hidden layers of 768 units and 12 self-attention heads for input text. The English, Chinese, and multilingual pre-trained BERT were taken from Google’s BERT repository. Russian BERT was trained on the Russian part of Wikipedia and news data. Pretraining RuBERT took 250 thousand steps on Tesla P100 x 8 [8]. First, we want to show whether spending expensive computational resources to pre-train language-specific BERT is justified or already available M-BERT can provide us with comparable results. To answer this question we compare language-specific BERT-based models with a multilingual BERT-based model while trained on the same language-specific monolingual training data. We perform comparison in two modes: **fixed mode** (3 epoch, learning_rate=5e-05, batch_size=8, dropout=0.1) and **validated mode**, where the training step is going until validation patience is achieved. Validation patience (equals to 10) defines how many steps to continue the training process while the performance on the validation set is not improving.

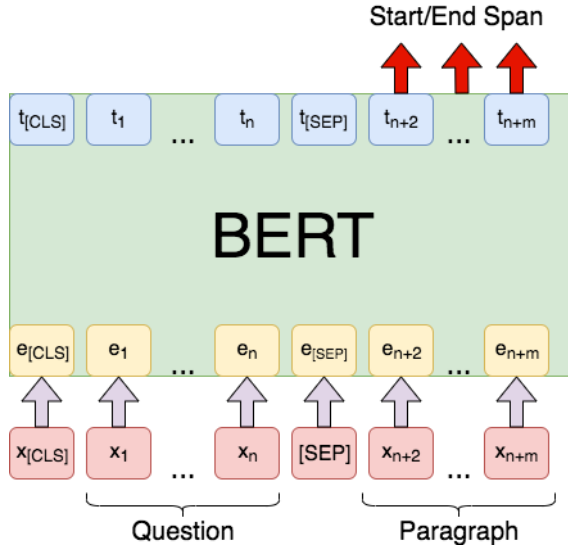


Figure 1: Architecture of the BERT model for SQuAD task

To identify to what extent the language-specific monolingual dataset contributes to the performance of the M-BERT based model, we build learning curves by comparing M-BERT based models in **validated mode**.

4.1. Evaluation Metrics

To measure the model's performance we use two metrics Exact Match (EM) and F1-score.

- **Exact match** measures the percentage of predictions that match any one of the ground truth answers exactly.
- **Macro-averaged F1** score measures the average overlap between the prediction and ground truth answers. The prediction and ground truth are bags of tokens.

5. Experimental Results

As expected, validating the models during the training step results in better performance as shown in the Tables.

Table 1: The models' performance (F1/EM) on SQuAD test set (English)

Model Settings (training set)	Fixed mode		Validated mode	
	F1	EM	F1	EM
M-BERT(SQuAD)	80.95	71.68	88.88	81.91
EnBERT(SQuAD)	82.87	74.48	88.67	81.32

The M-BERT based model trained on a language-specific training data substantially outperforms the same model trained on the English training data, this can be observed for both tested languages: **M-BERT(SberQuAD) >> M-BERT(SQuAD)** while tested on the SberQuAD test set and **M-BERT(DRCD) >> M-BERT(SQuAD)** for Chinese. The M-BERT based model trained on English SQuAD fails to extract the exact language-specific answer span that result in significant performance deterioration in EM than in F1.

The RuBERT-based model outperforms the M-BERT based model by a small margin while trained on the SberQuAD training set **RuBert(SberQuAD) > M-BERT(SberQuAD)** in **validated mode**, these results are in accordance with the results presented by the authors of RuBERT [8].

Table 2: The models' performance (F1/EM) on SberQuAD test set (Russian)

Model Settings (training set)	Fixed mode		Validated mode	
	F1	EM	F1	EM
M-BERT(SQuAD)	63.5	34.7	73.49	42.94
M-BERT(SberQuAD)	78.78	58.49	83.21	64.56
RuBERT(SberQuAD)	81.14	61.78	84.19	65.83

Table 3: The models' performance (F1/EM) on DRCD test set (Chinese)

Model Settings (training set)	Fixed mode		Validated mode	
	F1	EM	F1	EM
M-BERT(SQuAD)	51.73	37.8	75.39	60.28
M-BERT(DRCD)	81.7	75.19	88.9	83.7
ChBert(DRCD)	84.46	78.01	88.3	82.67

The DCRD dataset is based on Chinese Wikipedia, it contains Latin symbols that were not properly handled by the character-based tokenization used for pre-training Chinese BERT. This leads to the almost comparable performance of the M-BERT based model with a language-specific Chinese BERT-based model.

Interestingly, the difference in performance between M-BERT and language-specific BERT is much more significant in **fixed mode**, that means careful training stopping criteria might mitigate the language-specific BERT superiority.

We build learning curves to measure how the language-specific monolingual dataset contributes to the performance of the M-BERT based model. All models were executed in the **validated mode**. The learning curves for Russian are depicted in **Figure 2**. First, we define two boundaries, the upper bound is the M-BERT based model performance trained solely on the entire SberQuAD dataset (red dashed line). The lower bound is the M-BERT based model performance trained on the English SQuAD dataset (green dashed line). The **-SQuAD** curve denotes the model trained on the part of the SberQuAD dataset and the **+SQuAD** curve denotes the model trained jointly on the same part of the SberQuAD dataset and on the entire English SQuAD

dataset. Adding an entire English SQuAD train to the training set significantly improves performance in comparison to the model trained only on the part of the SberQuAD set. Moreover, the model trained on the joint dataset with 5,000–10,000 language-specific monolingual training samples is only a few points behind the model trained on the entire language-specific training set. The similar findings hold for Chinese as depicted in [Figure 3](#).

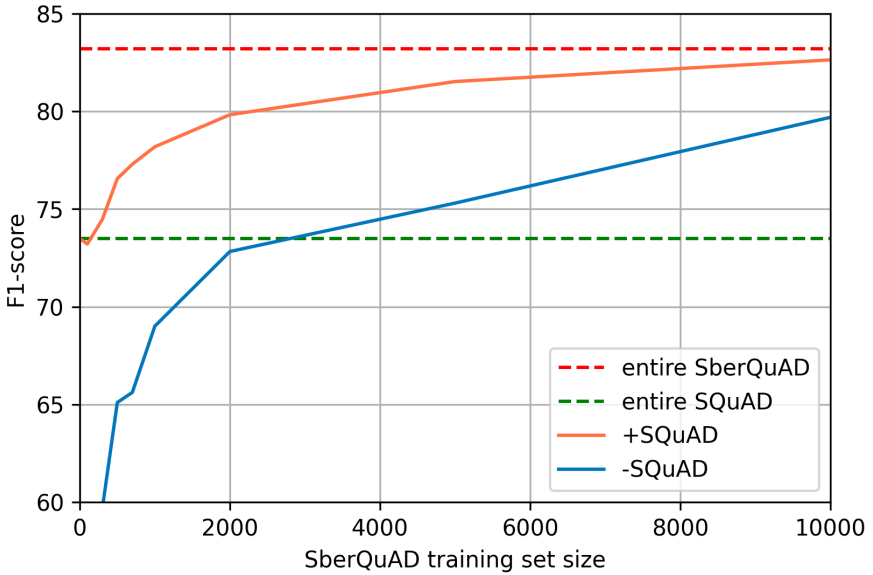


Figure 2: Learning curves on SberQuAD (Russian)

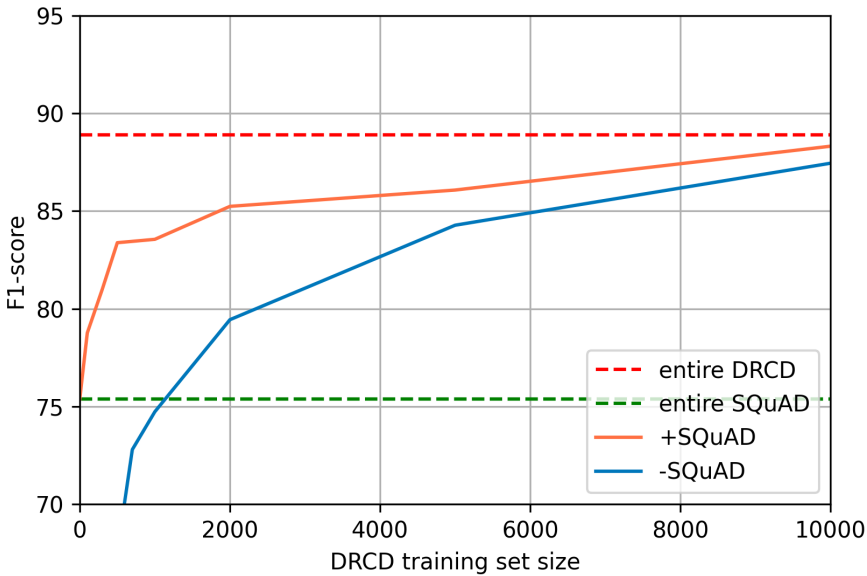


Figure 3: Learning curves on DRCD (Chinese)

6. Conclusion

We measured the multilingual BERT cross-language transferability on the Reading Comprehension task. Specifically, we demonstrated that M-BERT based models perform comparably with language-specific BERT while trained on the same training set. We also showed that M-BERT based models trained jointly on widely available English training data and a number of language-specific instances achieves comparable performance. Our results and analysis agree with the previous theories that M-BERT creates multilingual representations, that allow us to achieve promising performance in cross-lingual model transfer settings. To encourage researchers to further investigate the BERT’s cross-language transferability in different tasks we made our code publicly available¹.

7. Acknowledgements

The authors thank all the members of the Deep Learning and Neural Network lab at Moscow Institute of Science and Technology. This work was supported by National Technology Initiative and PAO Sberbank project ID 0000000007417F630002.

¹ <http://github.com/deepmipt/DeepPavlov/>

References

1. *Artetxe, M. et al.*: On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:1910.11856. (2019).
2. *Burtsev, M. et al.*: DeepPavlov: An open source library for conversational ai. ACL. (2018).
3. *Conneau, A. et al.*: Xnli: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:1809.05053. (2018).
4. *Devlin, J. et al.*: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).
5. *Efimov, P. et al.*: SberQuAD–russian reading comprehension dataset: Description and analysis. arXiv preprint arXiv:1912.09723. (2019).
6. *Google-Research*: TensorFlow code and pre-trained models for bert, (2018).
7. *Kaushik, D., Lipton, Z. C.*: How much reading does reading comprehension require? A critical investigation of popular benchmarks. arXiv preprint arXiv:1808.04926. (2018).
8. *Kurатов, Y., Arkhipov, M.*: Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213. (2019).
9. *Lewis, P. et al.*: Mlqa: Evaluating cross-lingual extractive question answering. arXiv preprint arXiv:1910.07475. (2019).
10. *Pires, T. et al.*: How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502. (2019).
11. *Rajpurkar, P. et al.*: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250. (2016).
12. *Schwenk, H., Li, X.*: A corpus for multilingual document classification in eight languages. arXiv preprint arXiv:1805.09821. (2018).
13. *Shao, C. C. et al.*: Drcd: A chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920. (2018).
14. *Wu, S., Dredze, M.*: Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. arXiv preprint arXiv:1904.09077. (2019).
15. *Yang, Y. et al.*: PAWS-x: A cross-lingual adversarial dataset for paraphrase identification. arXiv preprint arXiv:1908.11828. (2019).

DISFLUENCIES IN RUSSIAN SPOKEN MONOLOGUES: A DISTRIBUTIONAL ANALYSIS¹

Korotaev N. A. (n_korotaev@hotmail.com),

Podlesskaya V. I. (vi_podlesskaya@il-rgggu.ru)

Russian State University for the Humanities, Moscow, Russia

Smirnova K. V. (kategold1@gmail.com),

Fedorova O. V. (olga.fedorova@msu.ru)

Lomonosov Moscow State University, Moscow, Russia

The paper addresses the overall distribution of speech disfluencies in Russian spoken monologic discourse: basing on corpus data, we investigate qualitatively and quantitatively how disfluencies of different types group (or do not group) with each other and how isolated disfluencies and their sequences are sandwiched with periods of fluent speech in the course of speech production. Self-repairs, filled and silent pauses, and instances of hesitation lengthening were annotated in a subcorpus of the “Russian Pears Chats and Stories” (RUPEX). A distribution-oriented typology of disfluencies was proposed that distinguishes between isolated disfluencies, disfluency clusters, and quasi-clusters. We claim that disfluency tokens tend to cluster, as isolated occurrences are significantly less frequent in our data than it could have been expected basing on the relative frequency of tokens. This finding contradicts previous studies that treated disfluency clusters as a more marginal phenomenon, and emphasizes the importance of a distributional, rather than merely structural, approach to annotating disfluencies. Furthermore, individual types of disfluency tokens demonstrate significantly different distributional patterns. Compared to other types, self-repairs occur more often in isolation, while words with hesitation lengthening appear predominantly in clusters, and filled pauses most often group with silent pauses to form quasi-clusters.

Key words: speech disfluencies, spoken discourse, corpus, disfluency clusters

DOI: 10.28995/2075-7182-2020-19-454-466

¹ This study is supported by Russian Foundation for Basic Research (complex grant #18-00-01598, and subgrant #18-00-01485).

ДИСТРУБИТИВНЫЙ АНАЛИЗ РЕЧЕВЫХ СБОЕВ В УСТНОМ РУССКОМ МОНОЛОГИЧЕСКОМ ДИСКУРСЕ

Коротаев Н. А. (n_korotaev@hotmail.com),
Подлеская В. И. (vi_podlesskaya@il-rggu.ru)

РГГУ, Москва, Россия

Смирнова К. В. (kategold1@gmail.com),
Федорова О. В. (olga.fedorova@msu.ru)

МГУ им. М. В. Ломоносова, Москва, Россия

В статье рассматриваются дистрибутивные свойства речевых сбоев в устном русском монологическом дискурсе. На материале корпуса «Рассказы и разговоры о грушах» мы проанализировали, как различные типы речевых сбоев группируются или не группируются друг с другом в процессе реального речепорождения. Этим наша работа отличается от большинства исследований речевых сбоев, в которых это явление чаще изучается не с синтагматических, а с парадигматических позиций. Мы рассмотрели четыре типа речевых сбоев: самоисправления, заполненные паузы, абсолютные паузы при хезитации и хезитационные удлинения. Все эти типы могут встречаться как изолированно, так и в составе кластеров или квази-кластеров (эти понятия определены в тексте статьи); причем в целом кластеризация оказывается более частотным явлением, чем изолированное употребление. Кроме того, согласно предварительному анализу, типы сбоев варьируют по степени стремления к кластеризации: самоисправления значительно чаще встречаются в изоляции, хезитационные удлинения — в составе кластеров, заполненные паузы — в составе квази-кластеров.

Ключевые слова: речевые затруднения, устная речь, корпус, кластеризация речевых сбоев

1. Introduction

Speech disfluencies have been studied under the auspices of various linguistic disciplines, including discourse analysis, conversational analysis, experimental psycholinguistics, speech recognition, speech pathology, etc. The state of the art in this area is well represented, inter alia, in the proceedings of biennial Workshops on Disfluency in Spontaneous Speech (e.g., in the latest [DiSS 2019]), in multiple dissertations (e.g., [Bailey 2004], [Schnadt 2009], [Miller 2010], [Finlayson 2014]) and overviews ([Lickley 2015] is among the best), see also The Filled Pause Research Center—a very useful open online resource at <http://filledpause.com>.

Beginning with seminal works by [Willem J. M. Levelt 1983], [Elizabeth Shriberg 1994] and [Robert Eklund 2004], studies on disfluencies addressed primarily the typology of disfluencies as well as forms and functions of their individual types in individual

languages (cf. [Watanabe 2006], [2013] on Japanese, [Gósy 2019] on Hungarian, [Silber-Varod et al. 2019] on Hebrew, [Plug 2015] on Dutch, etc.) The most articulated research questions cover (a) the localization of disfluencies against hierarchical and linear structure of discourse, i.e. whether disfluencies appear at the boundaries of words, syntactic units, intonation units, turns etc., or inside the respective unit; (b) the influence of discourse genres (narratives, argumentative discourse, descriptions, etc.) and types of discourse (monologue vs. dialogue) on forms and frequencies of disfluencies; (c) segmental and suprasegmental characteristics of disfluent speech in comparison with fluent standards. In addition to these, an acute discussion in the literature on disfluencies, initiated primarily in a salient paper by [Herbert H. Clark and Jean E. Fox Tree 2002], concerns the functions of particular disfluencies (mainly, filled pauses) from the perspective of listeners: the central research question here is whether listeners comprehend disfluencies as intentional signals produced by speakers to inform about speech difficulties or as involuntary symptoms of those (cf. [Finlayson 2014] among recent contributions to this discussion).

However, much less is known and much less discussed is how disfluencies are distributed throughout big coherent portions of discourse; namely, how fluent and disfluent portions of speech alternate each other, which types of disfluencies cluster more easily, and which appear as isolated tokens surrounded by fluent speech? Literature on these distributional issues is much more scarce, some few exceptions to be mentioned include e.g. [Crible et al. 2017], [Betz et al. 2015], [Betz and Kosmala 2019], [Bóna 2019]. Our paper is in line with this distributional approach, which we advocate and check against Russian data. Since Russian is considerably underrepresented in disfluency studies, we aim at partially bridging also this gap.

Our main goal is to explore the overall distribution of speech disfluencies in Russian spoken monologic discourse. Basing on corpus data, we investigate qualitatively and quantitatively how disfluencies of different types group (or do not group) with each other and how isolated disfluencies and their sequences are sandwiched with periods of fluent speech in the course of speech production.

The following example may serve as a preliminary illustration of some (though not all!) phenomena we deal with in this study, including self-repairs as well as filled and silent pauses. These disfluencies can appear as singletons surrounded by fluent material or can group together.

(1) pears04: N-vE035—N-vE040²

N-vE035	/Podnimaetsja climbs.up	\nazəd. back
pN-016	(0.30)	
N-vE036	Prodolžæt continues	sobirat' \ŕgruši ^u , == to.pick pears

² See Section 2 for an overview of our data and transcripts details, and Section 3 for additional glossing conventions.

N-vE037	(*\D _a , yes						
N-vE038	v /kako _j -to in some	moment moment	proxo= passes.b=		(ə 0.25)	\bleet bleats	-/koza, goat
N-vE039	my we	slyšim hear	/↑ble _j anie, bleating				
pN-017	(0.23) SP						
N-vE040	(əu 0.58) FP		idēt walks	/čelove _k , man			

‘He climbs back up [the tree]. He resumes picking the pears... Oh yes, at some point a goat is pass-... bleating; we can hear the bleating and [we see] a man walking’

In (1), the first discourse unit, i.e. the first line in the transcript, is produced fluently, then comes a silent pause. At the end of the next line, N-vE036, the speaker abandons the current discourse unit and makes a switch in the narration line, as she suddenly decides to add a missing detail about a man with a goat who passes by (this type of repair, which affects the overall structure of a relatively big discourse fragment, is notated by a double equal sign). In line N-vE038, two disfluencies are grouped together: first, the word *proxo*= (initially intended as *proxodit* ‘passes by’) is truncated and further on substituted by the word *bleet* ‘bleats’ (this type of repair—caused by word searching difficulties inside the ongoing discourse unit—is notated by double vertical strokes); and then, immediately after that, comes a filled pause (ə). Later on, another filled pause, (əu), is found at the beginning of line N-vE040; notably, it is immediately preceded by a silent pause. We will refer to example (1) later, in Section 3.

The rest of the paper is organized as follows. In Section 2, we present our data. In Section 3, we discuss the principles for annotating disfluencies, propose a distributive typology thereof, and analyze some typical examples. Section 4 contains quantitative results and their statistical evaluation. In Section 5, we provide a conclusion.

2. Data

Our data come from the “Russian Pear Chats and Stories” corpus (RUPEX; see [Kibrik, Fedorova 2018]). RUPEX consists of annotated communication sessions that share the same general design. In each session, three participants with fixed roles (Narrator, Commentator, and Reteller) alternate in relating and discussing the contents of the “Pear Film” [Chafe (ed.) 1980]. RUPEX is a multichannel corpus; it registers not only speech, but also kinetic actions such as manual gesticulation, head movements, and gaze direction. The analysis in this study, however, is confined to vocal modality. We used the *vocal transcripts* available on RUPEX website (<https://multidiscourse.ru/corpus/?en=1>). In transcripts, each line corresponds to an elementary discourse unit (EDU; see [Kibrik, Podlesskaya (eds.) 2009]) or a pause between EDUs; pause durations are indicated in parentheses; accents placement and pitch

directions are iconically indicated with slashes and arrows; punctuation marks at the end of EDUs code illocutionary and phase values; see [Korotaev et al. 2020] for more detail. Additional conventions are discussed in this paper, when needed.

Three recorded sessions were analyzed for this study. In examples, we provide the session ID (04, 22, or 23) and IDs of the EDUs involved (in these IDs, N stands for Narrator, C for Commentator, and R for Reteller). Since we were primarily interested in how fluent and disfluent stretches are distributed across relatively long speech intervals, we concentrated on monologic fragments and didn't consider dialogical parts³. Overall, we annotated 32 mins of audio that contained 4,780 words.

3. Annotating disfluencies

When annotating disfluencies, we implemented a *distribution-oriented typology* that draws a primary distinction between isolated disfluencies and disfluency clusters. *Isolated* are single disfluency tokens surrounded by fluent stretches; *clusters* are sequences of disfluency tokens that come together. Before discussing this distinction in more detail (see Section 3.2), we will first present the types of tokens that were taken into account.

3.1. Basic disfluency tokens

Basic disfluency tokens are what most traditional classifications are about; they are elementary signals (or symptoms⁴) of speech difficulties. Since our analysis was primarily aimed at revealing the distributional patterns of disfluencies, we used a simplified typology of basic tokens that includes the following options.

3.1.1. Self-repairs

In self-repairs, speakers interrupt their speech, reject an already pronounced fragment, and then restart or correct it. Following the classical model introduced in [Schriberg 1994], we analyze self-repairs as having three elements: reparamandum, interruption point (often signaled by words truncation), and repair (or, reparans). In example (1) provided above, there are two instances of self-repair, one at a macro-level, and the other one—which is combined with a filled pause—inside an EDU. Another instance of a self-repair inside an EDU is found in (2). Here, a so-called false-start takes place. As the speaker starts uttering the word *nežno* 'gently', she decides that it should not precede but rather follow the verb *delat* 'does'; so she immediately interrupts her production [Levelt, 1993: 478], provides the verb, and then restarts the suspended verbalization. In examples provided in this paper, we gloss interruption points as BR (from "break").

³ In our prior work, we found that monologues and dialogues differ in what speech disfluencies occur more frequently. Compared to monologues, speakers in dialogues tend to produce significantly less filled pauses [Korotaev 2019] and more repairs [Podlesskaya et al. 2019]. This is another reason why we decided to stick to the monologic data in our current research.

⁴ See Section 1 for a brief overview of this delimitation in the literature. As for now, we are not ready to adhere to a particular decision as to whether speech disfluencies are rather signals of difficulties in planning, or mere (non-voluntary?) symptoms thereof. In this paper, we try to abstract ourselves from this distinction and use somewhat neutral terms.

(2) pears04: N-vE034

N-vE034 tak vsë očēn' ne= || delaet /nežno i \ljubovno^u.
 this.way everything very tend= BR does tenderly and lovingly

'And he is tend-... doing all this in a very tender and loving way'

Excerpts (1) and (2) illustrate different types of self-repairs. Numerous classifications have been brought to account on these and still other differences, including the scope and the localization of the repair, the type of the repaired feature (lexical, grammatical, constructional), etc.; see, inter alia, [Schegloff 2013]; [Podlesskaya 2015]. In this study, for the sake of simplicity, we do not distinguish between different types of self-repairs and treat them all uniformly. Self-repair is a complex phenomenon that involves several parts; in the quantitative analysis below, cf. Section 4, counted are frequencies of interruption points.

3.1.2. Filled pauses (FPs)

There are two instances of filled pauses in example (1) above, one combined with a self-repair, the other following a silent pause. An example of an FP surrounded by fluent speech is given in (3), where it occurs between the verb and the adjunct.

(3) pears22: R-vE163

R-vE163 i /gruši razletelis' (ə 0.49) \vokrug.
 and pears got.scattered FP around

'And the pears got scattered all around'

Filled pauses may have slightly different functions depending on their phonetic realizations [Clark, Fox Tree 2002], but their general role of hesitation markers seems undisputable. In all the examples above, an FP is used when the speaker hesitates on how to continue her production and seeks to gain some additional time⁵.

3.1.3. Lengthening

Functionally and formally close to FPs are some cases of lengthening, or phoneme prolongation. In (4), the speaker prolongates the final vowel of the complex preposition *v storonu* 'towards' before providing the noun *dereva* 'tree'. A plausible interpretation of this lengthening is that the speaker first seeks for a more concrete nomination (like, for instance, 'a pear tree') but then rejects this idea and uses a generic term. In transcripts, lengthening is conveyed with a hyphenated notation.

(4) pears22: R-vE207

R-vE207 poka značit əmal'čiki idut v storonu-u \ʔdereva,
 while so boys go towards tree

'So, while the boys are heading towards the tree...'

⁵ In RUPEX, several types of FPs are distinguished—(ə) and (e) for *uh*- and *ah*-like pauses, (u) for *um*-like pauses, (ʔ) for glottal stop fillers, as well as various mixed cases (see (ə*m*) in (1)). In our analysis, we neglect these distinctions.

As has been repeatedly pointed out in literature, in cases like (4) lengthening is similar to filled pauses as it also signals hesitation by means of extended vocalization (see, inter alia, [Eklund 2004: 241–252]). However, there are numerous uses of this formal device beyond hesitation. Lengthening may express emphasis [Braver et al. 2014], intensification and spatio-temporal relations [Kodzasov 2009: 112–125, 137–146], appear in phonetically motivated contexts [Wightman et al. 1992], take part in complex discourse strategies such as addressing someone or remembering something [Yanko 2008: 98–117], and so on. In this study, we annotated prolongations as disfluency tokens only when we were quite sure of their hesitation function in the given context. As for now, we are not aware of ways to make this procedure more formal.

3.1.4. Silent pauses (SPs)

Similar problems arise when annotating silent, or absolute, pauses. Sometimes, when an SP appears in an “unexpected” position and/or is “unexpectedly” long, it can also be interpreted as a hesitation phenomenon. For instance, in (5) the speaker suspends vocalization after providing the adverb *tuda* ‘there’, keeps silent for more than 600 ms, and then completes the clause by producing the verb *skladyvaet* ‘puts’. This can hardly be interpreted otherwise than an instance of hesitation, as the speaker is apparently trying to find a good way to express the idea of putting an already indicated referent into an indicated place.

(5) pears23: R-vE156

R-vE156	on	èti	/gruši	tuda	(0.61)	\skladyvaet.
	he	these	pears	there	SP	puts

‘He puts these pears there’

Still, just like lengthening, SPs don’t only occur as hesitation devices. Most importantly, silent pauses, as well as pauses filled with loud inhalations⁶, tend to coincide with boundaries between elementary discourse units and other discourse constituents (see [Selkirk 1984]; [Chafe 1994]; [Kibrik, Podlesskaya (eds.) 2009]; etc.). Perception-based criteria are usually used to tell these two functions apart (see, e.g., [Lickley 2015: 458]; [Trouvain et al. 2016]); and we followed the same principle. Here, it suffices to say that the SP in (5) was treated as a disfluency token, while a shorter pause placed between lines N-vE035 and N-vE036 in (1) was not.

3.1.5. Other

There are still other markers of disfluency: placeholders (*ètot* ‘that one’, *kak ego* ‘whatchacallit’), lexical search items (*nu* ‘well’), editing terms (*to est* ‘that is’, *net* ‘no’), etc. They were not annotated for this study.

3.2. Isolated vs. clustered disfluencies

As noted above, the distribution-oriented typology we propose in this study is based on the distinction between isolated disfluencies and disfluency clusters. Disfluencies

⁶ SPs per se and those filled with loud inhalations are annotated separately in the corpus but are not differentiated for counts in this study.

presented in examples (2)–(5) were annotated as isolated since they are surrounded by fragments of fluent speech. Contrary to these cases, clusters are sequences of disfluency tokens that come adjacent to one another. For example, in line N-vE038 of excerpt (1) above, the interruption point of a self-repair (truncated verb *proxo*= ‘passes b=’) is immediately followed by an FP (ə 0.25). A more complex cluster is present in (6). The line starts with an FP, then the final vowel of the pronoun *oni* ‘they’ is prolonged and gradually evolves into another FP, after that the speaker truncates her first attempt to provide the verb *razminajutsja* ‘pass by each other’ and the interruption point is once again followed by an FP. Overall, five adjacent disfluency tokens precede a stretch of fluent speech.

(6) pears04: N-vE097

N-vE097 (e 0.30) Oni-i^ə (e 0.36) r= || (ə 0.23) razminajutsja na-↑doroge^u,
 FP they FP p- BR FP pass.by.each.other on road

‘They pass by each other on the road’

We consider the opposition of isolated vs. clustered disfluencies significant, as the latter type presumably indicates more severe difficulties than the former (see, e.g., [Robb et al. 2009]; [Bóna 2019] for clinical accounts on clustering). However, there is no clear-cut between these two types. We find it useful to point out at least two intermediate cases. First, sometimes disfluency tokens don’t follow one another immediately, but are still apparently connected. In (7), the speaker enters the “disfluency mode” as she prolongates the final consonant of *potom* ‘then’ in line N-vE083. After that, she successfully provides the matrix verb *ponimaet* ‘understands’ but resumes hesitating on what to say next and uses a combination of lengthening and FP right in the beginning of the complement clause.

(7) pears04: N-vE083

N-vE082 snačala mal’čik xočet vzjat’ /odnu ↑grušu,
 at.first boy wants to.take one pear

N-vN023 (q 0.49)

N-vE083 /potom-m ponimaet čto-o (’ 0.45) ničto emu ne /grozīt,
 then understands that FP nothing to.him not threatens

‘At first, the boy wants to take only one pear, but then he understands that he’s in no danger...’

Cases like (7) were also annotated as clusters. To discriminate between clusters and isolated disfluencies we used a one-word threshold: tokens separated by no more than one fluent notional word were considered elements of clusters⁷.

Second, quite frequent (see Section 4) are sequences of single disfluency tokens immediately followed and/or preceded by silent pauses. Such sequences were annotated as *quasi-clusters*; see (8), where an FP is preceded by an SP. Occurring between the topical noun phrase and the rhematic verb phrase, the SP doesn’t indicate

⁷ A looser, two-word, threshold is used in [Gósy 2012].

disfluency in itself, but followed by an FP, it may be reinterpreted as a hesitation device. (A similar quasi-cluster is present in lines pN-017—N-vE040 of excerpt (1).)

(8) pears04: R-vE497

R-vE497 V èto /vremja /-fermer (0.55) (ə 0.31) spuskaetsja s /↓→lestnicy,,
 in this time farmer SP FP goes.down from ladder

‘At this point, the farmer gets down the ladder’

4. Results and discussion

440 disfluencies were annotated in our data. **Table 1** shows the distribution of these disfluencies across isolated, clustered, and quasi-clustered types.

Table 1. Isolated disfluencies, disfluency clusters, and quasi-clusters in annotated data

	Isolated	Quasi-clusters	Clusters	Overall
N (%)	126 (28.6%)	174 (39.6%)	140 (31.8%)	440 (100%)

Numbers in **Table 1** suggest that, contrary to some previous studies (see, e.g., [Betz, Cosmala 2019]), clustering of disfluencies is not a marginal phenomenon, but rather a standard tendency. Even if we don’t consider quasi-clusters (the most frequent, and somewhat intermediate, case), proper clusters still occur more frequently than isolated disfluencies in our data.

To provide a preliminary statistical evaluation of these findings, we analyzed the distribution of disfluency tokens across isolated and clustered disfluencies. Our assumption was that tokens don’t appear independently of one another, i.e. that clustering is not a matter of coincidence. For the sake of simplicity, we eliminated all silent pauses from the analysis, so tokens that appear in quasi-clusters were counted as isolated. The results are as follows. Out of 672 disfluency tokens (self-repairs, FPs, words with lengthening), 391 are used in clusters, and 281 are isolated. We compared the observed distribution with the distribution expected under the null hypothesis, as shown in **Table 2**. The expected distribution was calculated based on the relative frequency of disfluency tokens in the analyzed data (672/4,780 words; i.e. 0.14). If disfluency tokens appeared independently of one another, the probability for a word to have at least one disfluency token at a 2-word distance (see discussion in **Section 3.2**) would equal $0.45 (1 - 0.86^4)$, as the distance should be laid out in both directions). The difference between the observed and the expected distribution is significant (chi-square test; $p < .00001$); the null hypothesis was rejected.

Furthermore, we observed significant differences in how specific types of disfluency tokens comply with the overall tendency of clustering. **Fig. 1** presents the normalized distributions of filled pauses, words with lengthening, and self-repairs across isolated, quasi-clustered, and clustered disfluencies.

Table 2. Observed vs. expected distributions of disfluency tokens across isolated and clustered types

Tokens occurring:	Observed	Expected
as isolated disfluencies or in quasi-clusters	281	370
in clusters	391	302

**Figure 1.** Appearance of disfluency tokens in isolated, quasi-clustered, and clustered disfluencies

The differences shown in **Fig. 1** are statistically significant (chi-square test; $p < .00001$). While all three types are predominantly used in clusters, self-repairs seem to be the most independent type of disfluencies, as 29.9% of them (50 out of 167; remember that we count interruption points here) occur in isolation. Filled pauses, on the contrary, are isolated in just 8.3% (27 out of 325), but often (38.8%) appear in quasi-clusters, i.e. come exclusively grouped with silent pauses and no other disfluency tokens. As for lengthening, tokens of this type most frequently appear in clusters—71.8% (128 out of 180), compared to 52.9% and 54.5% of filled pauses and self-repairs, respectively.

5. Conclusions

We analyzed the speech disfluencies (self-repairs, filled pauses, hesitation lengthening and silent pauses) found in the monologic parts of three communicative sessions included into the “Russian Pear and Chats Stories” corpus. Our main concern was to find out whether disfluencies tend to group together, or they most frequently appear in isolation of one another. We annotated our data using a distribution-oriented typology of disfluencies that distinguishes between isolated cases, clusters, and quasi-clusters. Clusters turned out to be quite frequent, as they accounted for nearly 32% of all

disfluencies. This allowed us to assume that clustering, often considered as manifesting especially severe speech difficulties, is in fact a standard feature of natural spoken discourse. We provided a preliminary statistical evaluation of this claim by further investigating into how often basic disfluency tokens (filled pauses, lengthening, and self-repairs) group or do not group together. We found that over 58% of 672 tokens appeared in clusters, which is way more often than it could have been expected basing on the relative frequency of tokens in our data. Also, specific types of disfluency tokens behave somewhat differently, as self-repairs occur more often in isolation, while words with hesitation lengthening appear predominantly in clusters, and filled pauses most often group with silent pauses to form quasi-clusters.

We foresee several expansions of our study. First, more accurate results on clustering could be obtained when all instances of disfluency tokens are annotated separately. As we concentrated on a holistic analysis of disfluencies, this has not been properly done yet for our data. Second, a typology of clusters could be introduced basing on number, types, and respective order of disfluency tokens. For example, we found that clusters containing one or several filled pause(s) and one or several instance(s) of lengthening are the most frequent in our data (50 out of 140 clusters, or 35.7%); a finer-grained annotation could yield more specific results. Finally, we understand that our methodology could be applied for annotating other related phenomena like speech disfluencies in people who stutter or clustering in manual gesticulation. Also, it has already been used for selecting stimuli in an event-related fMRI experiment [Smirnova et al. 2020].

Acknowledgments

We express our gratitude to Svetlana Toldova for her help with a statistical assessment of our data. We also thank anonymous reviewers for their insightful observations. All remaining mistakes are, of course, our responsibility.

References

1. *Bailey K. G. D.* (2004), *Disfluent speech and the visual world: An application of the visual world paradigm to the study of spoken language comprehension*, PhD dissertation, Michigan State University, USA.
2. *Betz S., Wagner P., Schlangen D.* (2015), *Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis*, Proceedings of Interspeech 2015, Dresden, Germany, pp. 2222–2226.
3. *Betz S., Kosmala L.* (2019), *Fill the silence! Basics for modeling hesitation*, Proceedings of DiSS 2019, The 9th Workshop on Disfluency in Spontaneous Speech 2019, Eötvös Loránd University, Budapest, Hungary, pp. 11–14.
4. *Bóna J.* (2019), *Clustering of disfluencies in typical, fast and cluttered speech*, *Clinical Linguistics & Phonetics*, Vol. 33, Iss. 5, pp. 393–405.
5. *Braver A., Dresher N., Kawahara S.* (2014), *The Phonetics of Emphatic Vowel Lengthening in English*, Proceedings of the 2014 Annual Meeting on Phonology, available at <https://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/3754/3473>.

6. *Chafe W.* (Ed.) (1980), *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*, Norwood: Ablex.
7. *Chafe W.* (1994), *Discourse, consciousness, and time*, University of Chicago Press, Chicago.
8. *Clark, H. H., Fox Tree J. E.* (2002), Using uh and um in spontaneous speaking, *Cognition*, Vol. 84, pp. 73–111.
9. *Crible L., Degand L., Gilquin G.* (2017), The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency, *Languages in Contrast*, Vol. 17, Iss. 1, pp. 69–95.
10. *Corley M., Stewart O. W.* (2008), Hesitation disfluencies in spontaneous speech: The meaning of um, *Language and Linguistics Compass*, Vol. 2, Iss. 4, pp. 589–602.
11. *Eklund R.* (2004), *Disfluency in Swedish human-human and human-machine travel booking dialogues*, PhD dissertation, University of Linköping, Sweden.
12. *Finlayson I. R.* (2014), *Testing the roles of disfluency and rate of speech in the coordination of conversation*, PhD dissertation, Queen Margaret University, Edinburgh, UK.
13. *Gósy M.* (2012), BEA—A multifunctional Hungarian spoken language database, *The Phonetician*, Vol. 105, Iss. 10, pp. 50–61.
14. *Gósy M.* (2019), Halt command in word retrieval, *Proceedings of DiSS 2019, The 9th Workshop on Disfluency in Spontaneous Speech*, Eötvös Loránd University, Budapest, Hungary, pp. 3–6.
15. *Kibrik A. A., Fedorova O. V.* (2018), Language production and comprehension in face-to-face multichannel communication, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2018”* [Komp’juternaya Lingvistika i Intellektual’nye Texnologii: Trudy Mežunarodnoj Konferencii “Dialog 2018”], Moscow, RGGU, pp. 305–316.
16. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009), *Night Dream Stories: A corpus study of spoken Russian discourse* [Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa], LRC, Moscow.
17. *Kodzasov S. V.* (2009), *Studies in Russian Prosody* [Issledovanija v oblasti russkoj prosodii], LRC, Moscow.
18. *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 meždunarodnoj konferencii “Korpusnaja lingvistika-2019”]*, Saint Petersburg University, Saint Petersburg, pp. 48–54.
19. *Korotaev N. A., Kibrik A. A., Podlesskaya V. I.* (2020), Annotating the vocal modality, *O. V. Fedorova, A. A. Kibrik* (eds.), *The MCD handbook: A practical guide to annotating multichannel discourse*, in print.
20. *Levelt W. J. M.* (1983), Monitoring and self-repair in speech, *Cognition*, Vol. 14, Iss. 1, pp. 41–104, 1983.
21. *Levelt W. J. M.* (1993), *Speaking: From Intention to Articulation*, The MIT Press.
22. *Lickley R. J.* (2015), Fluency and Disfluency, *Melissa A. Redford* (ed.) *The Handbook of Speech Production*, John Wiley & Sons, pp. 445–474.
23. *Miller T. A.* (2010), *Generative models of disfluency*, PhD dissertation, University of Minnesota.
24. *Plug L.* (2015), Prosodic Marking and Predictability in Lexical Self-Repair, *Proceedings of the 18th International Congress of Phonetic Sciences 2015*, Glasgow, UK, available at <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0032.pdf>.

25. *Podlesskaya V. I.* (2015), A corpus-based study of self-repairs in Russian spoken monologues, *Russian Linguistics*, Vol. 39, Iss.1, pp. 63–79.
26. *Podlesskaya V. I., Korotaev N. A., Mazurina S. I.* (2019), A corpus study of self-repairs in Russian monologues and dialogues [Samoispravljenija govornjaščego v ruskom monologičeskom i dialogičeskom diskurse: opyt korpusnogo issledovanija], *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019” [Komp’juternaja Lingvistika i Intellektual’nye Texnologii: Trudy Meždunarodnoj Konferencii “Dialog 2019”]*, RGGU, Moscow, pp. 508–522.
27. *Robb M. P., Sargent A., O’Beirne G. A.* (2009), Characteristics of disfluency clusters in adults who stutter, *Logoped Phoniatri Vocol.*, Vol. 34, Iss. 1, pp. 36–42.
28. *Schegloff E. A.* (2013), Ten operations in self-initiated, same-turn repair, Makoto Hayashi, Geoffrey Raymond, Jack Sidnell (Eds.), *Conversational repair and human understanding*, Cambridge: Cambridge University Press, pp. 41–70.
29. *Schnadt M.* (2009), *Lexical influences on disfluency production*, PhD dissertation, University of Edinburgh, UK.
30. *Selkirk E. O.* (1984), *Phonology and Syntax: The Relation Between Sound and Structure*, Cambridge, MA: The MIT Press.
31. *Shriberg E.* (1994), *Preliminaries to a theory of speech disfluencies*, PhD dissertation, University of California, Berkeley, USA.
32. *Silber-Varod V., Gósy M., Eklund R.* (2019), Segment prolongation in Hebrew, *Proceedings of DiSS 2019, The 9th Workshop on Disfluency in Spontaneous Speech 2019*, Budapest, Hungary, pp. 47–50.
33. *Smirnova K. V., Korotaev N. A., Panikratova Y. R., Lebedeva I. S., Pechenkova E. V., Fedorova O. V.* (2020), Using the RUPLEX Multichannel Corpus in a Pilot fMRI Study on Speech Disfluencies, *Twelfth International Conference on Language Resources and Evaluation (LREC)*, pp. 195–203, available at <http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.25.pdf>.
34. *Trouvain J., Fauth C., Möbius B.* (2016), Breath and Non-breath Pauses in Fluent and Disfluent Phases of German and French L1 and L2 Read Speech, *Proceedings of Speech Prosody (SP8) 2016*, Boston, USA, pp. 31–35.
35. *Watanabe M., Hirose K., Den Y., Miwa S., Minematsu N.* (2006), Factors influencing ratios of filled pauses at clause boundaries in Japanese, *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics 2006*, Athens, Greece, pp. 253–256.
36. *Watanabe M.* (2013), Phrasal complexity and the occurrence of filled pauses in presentation speeches in Japanese, *Proceedings of DiSS 2013, the 6th Workshop on Disfluency in Spontaneous Speech and TMH-QPSR 2013*, KTH Royal Institute of Technology, Stockholm, Sweden, Vol. 54, Iss. 1, pp. 71–72.
37. *Wightman C. W., Shattuck-Hufnagel S., Ostendorf M., Price P. J.* (1992), Segmental durations in the vicinity of prosodic phrase boundaries, *The Journal of the Acoustical Society of America*, Vol. 91, Iss. 3, pp. 1707–1717.
38. *Yanko T. E.* (2008), Intonation strategies of Russian speech from a contrastive perspective [Intonacionnye strategii russkoj reči v sopostavitel’nom aspekte], *LRC*, Moscow.

R-BERT FOR RELATIONSHIP EXTRACTION ON RUSSIAN BUSINESS DOCUMENTS

Korzun V. A. (sdernal2@gmail.com)

МИПТ ABBYY Lab, Moscow, Russia

This paper provides results of participation in the Russian Relation Extraction for Business shared task (RuREBus) within DialogueEvaluation 2020. Our team took the first place among 5 other teams in Relation Extraction with Named Entities task. The experiments showed that the best model is based on R-BERT model. R-BERT achieved significant result in comparison with models based on Convolutional or Recurrent Neural Networks on the SemEval-2010 task 8 relational dataset. In order to adapt this model to RuREBus task we also added some modifications like negative sampling. In addition, we have tested other models for Relation Extraction and Named Entity Recognition tasks.

Key words: BERT, relationship extraction, named entity recognition, russian business, deep learning

DOI: 10.28995/2075-7182-2020-19-467-473

R-BERT ДЛЯ ИЗВЛЕЧЕНИЯ ОТНОШЕНИЙ НА ДОКУМЕНТАХ РУССКОГО БИЗНЕСА

Корзун В. А. (sdernal2@gmail.com)

МФТИ ABBYY Lab, Москва, Россия

1. Introduction

At the moment, many natural language processing (NLP) tasks are solved with large pre-trained language models like BERT [1]. Relationship extraction (RE) is one of them. One of state-of-the-art results [2] is based on pre-trained BERT with some modifications. There are also models based on Convolution or Recurrent Neural Networks, like [3] and [4].

Relationship extraction task can be considered as a sentence classification task. Given a sentence and a pair of nominals, the objective is to identify the relation between nominals. In this case task could be called relationship classification.

Named entity recognition (NER) could be a sub-task for relationship extraction, when pairs of nominals are not given and first must be found. NER task can be considered as a sequence labeling task. Given a sequence of tokens in a sentence, the objective is to classify each token. Classes should correspond to entity types. Quite often for each type of entity classes are added, corresponding to the beginning, middle or end of the entity.

In a broader sense in relationship extraction task there are also possible cases, when sentence contains more than two entities. One way to solve this problem is to duplicate sentence for each possible pair of entities. Then, the resulting samples are classified separately. Thus the relationship extraction task could be reduced to the relationship classification task.

Our contribution is to adapt state-of-the-art method for relationship classification to case when more than two entities could be found in the sentence by using special sampling approach.

2. Shared task overview

RuREBus is a public task for named entity recognition and relationship extraction for Russian on business documents. The dataset provided is new and was developed for this particular task. Authors describe their corpus as a real set of documents from business and task could be more complicated than for known datasets for relationship classification. For more information refer to [5].

2.1. Problem description

The whole competition divided into 3 tasks:

- **Named Entity Recognition.** The objective is to find entities spans and classify them. The object metric is exact micro F-measure.
- **Relationship Extraction with Named Entities.** Given an entities for each document find a relationship between them and provide relation type. The objective metric is micro F-measure.
- **End-to-end Relationship Extraction.** Similar to the previous, but without entities markup. First, the objective is to find entities, than identify relations between them.

The competition took place in 2 phases. First for 1st and 3rd task and second for the 2nd task.

2.2. Data description

Shared task organizers provided manually annotated documents and the large corpus of unlabelled documents for training models. The train set consists of 188 documents. Each document includes a raw text file and an annotation file in BRAT format. Annotation file contains a list of entities and relations in the following format:

`'T{idx}\t{type} {span start} {span end}\t{value}'` for entities
`'R{idx}\t{type} Arg1:T{e1 idx} Arg2:T{e2 idx}'` for relations

Spans are given as offset in symbols from the beginning of a document.

The test set provided for the 1st phase contains 544 raw text file for NER and end-to-end RE tasks. For the 2nd phase there are also annotations files with entities for RE with NE's task.

2.3. Data problems

We found some difficulties working with dataset:

- **Inconsistent punctuation.** Some documents contain punctuation marks separated from other tokens by space. It could help in tokenization, but other documents have punctuation marks attached to tokens.
- **New line separation.** New line separators can be found between sentences as well as inside sentences. This brings a lot of problems with sentence segmentation.
- **Long sentences.** There are also a lot of long sentences in the dataset. Some of them represent enumerations divided by semicolon. It also brings problems with training models such BERT which GPU memory consumption depends on sequence length.

To train BERT on GPU with 8GB RAM, sequence length should be limited. Therefore, we decided to split sentences as follows:

1. split input document by sentences using nltk [6] sentence tokenizer
2. split the remaining long sentences by the following substrings consistently:

`','`, `'\n\n'` and `'\n'`

3. item split the remaining sentences by maximum length (120)

However, to train models based on recurrent networks we decided to use only nltk sentence tokenizer.

3. Solution for NER task

For the NER task we have tried different models based on LSTM with Self-attention and BERT.

3.1. Self-attention LSTM

For the NER task we mostly used Bidirectional LSTM followed by Self-attention. First, we have tried to use only Russian fasttext word embeddings [10] as token features and got an adequate score. Then, we have added casing features, part-of-speech tags from pymorphy [7] and character embeddings. Results and model dimensions are given below.

Table 1: Results

Features	F-score on test
only fasttext	0.4357
+ casing + char	0.4559
+ POS-tags	0.4638

Table 2: Model dimensions

LSTM hidden	400
word embeddings	300
char embeddings	20
chars hidden	50
POS-tag embeddings	30
casing embeddings	10
dropout	0.4

3.2. BERT for NER

Further, we have tried to use output from pre-trained BERT as token features instead of using fasttext embeddings and other features. We have tried the simple Multilayer perception and the same encoder from previous experiments (BiLSTM + Self-Attention) as encoders. We have also tried to freeze and unfreeze BERT parameters. And as pre-trained BERT models we used RuBERT from DeepPavlov [8] and multilingual-cased from Transformers library [9]. The results are listed in **Table 3**.

Table 3: BERT results on NER

Model	F-score
RuBERT(frozen) + MLP	0.1916
RuBERT(frozen) + SALSTM	0.2941
*RuBERT(unfrozen) + SALSTM	0.43
*Multilingual cased BERT(unfrozen) + MLP	0.5144
*RuBERT(unfrozen) + MLP	0.5469
lst-place competitor	0.561

* the results were obtained after phase end

We found that using BERT output as embeddings for subsequent encoder degrades the overall quality. To get the maximum benefit from BERT, we must unfreeze its parameters and use the simplest classifier. Furthermore, the latest release of RuBERT for PyTorch outperforms standard multilingual BERT. However, using just BERT is not enough to take state-of-the-art result for this task. To improve our results we could use ensemble of BERT models or other state-of-the-art approaches for NER.

4. Solution for RE task

For relationship extraction task we used two models based on BERT. For both models we used pre-trained multilingual-cased BERT. To solve this task, first, we reduced it to relationship classification task like SemEval-2010 Task 8 by sampling sentence for each pair of entities.

4.1. Negative sampling

In SemEval-2010 Task 8 only two entities are given for each sentence. In RuREBus task we have various number of entities for each sentence. Therefore, we create a sample containing a sentence and a pair of entities, for each two entities in the sentence. A sample is called negative if its pair of entities is not in the markup, the rest of samples are positives. Now the task is reduced to SemEval task, where samples are taken instead of sentences.

The number of samples created such way was enormous and the training epoch would last a very long time. Therefore we suggested to reduce the number of negative samples. The way of reduction is taking all positive pairs and some number of negatives. For negative samples we took random pairs from sentence in number similar to positive samples. There were also taken some negative samples from sentences without positive ones. And for each train epoch negative sample are taken randomly. This allowed us to reduce number of training batches by about 3 times. Nevertheless, all possible pairs of nominals are taken for evaluation.

4.2. NER embeddings over BERT

The first idea is to use BERT with NER embeddings. We concatenated BERT output with NER embeddings for each token. Then it is passed through attention to create sample feature vector. The vector obtained is used by MLP classifier to predict classes distribution.

4.3. R-BERT

The next model is inspired by [2]. Their model got one of the best results in SemEval-2010 Task 8 [11].

The approach is as follows. First, special tokens are inserted before and after entities and modified sentence is passed through the pre-trained BERT. Then, the BERT outputs corresponding to the [CLS] token and each entity are taken. Entities outputs are averaged and all 3 received vectors pushed through FeedForward layers. Then they are concatenated and passed into a Linear Layer followed by Softmax. All hyperparameters are taken from the original article except the max sentence length which was 120. The table with results of used models and the closed competitor are listed below:

Table 4: Results on RE

Model	F-score
BERT + NER embeddings	0.2066
R-BERT	0.44
Closest competitor	0.394

Hence, R-BERT outperformed our first model and the closest competitor and got state-of-the-art result on this task. It seems to us that the main feature of R-BERT is special token insertion. This could help BERT to attend on entities and generate

output more suitable for the task. R-BERT authors also claim that output corresponding the [CLS] token represents the semantics of the sentence and the output corresponding the entities represents their semantics. As a result, this model achieved the best score in the competition. However, some samples could be lost due to tokenization. And the large unlabeled corpus was not used. It can be used to pre-train BERT for the domain and probably get better results.

5. Conclusion

In this paper we present a winning solution of Russian Relation Extraction for Business task within Dialogue Evaluation 2020. We took one of the best models for Relation Classification for English and combined it with special sampling procedure. We have also compared BERT and RNN based approaches for Named Entity Recognition task. This result with such score is not sufficient for fully automatic relationship extraction, given also the low score on named entity recognition task. Nevertheless, this approach can be used as an auxiliary for manual markup.

References

1. *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.* (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
2. *Wu, S., & He, Y.* (2019, November). Enriching pre-trained language model with entity information for relation classification. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2361–2364).
3. *Wang, L., Cao, Z., De Melo, G., & Liu, Z.* (2016, August). Relation classification via multi-level attention cnns. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1298–1307).
4. *Zhang, S., Zheng, D., Hu, X., & Yang, M.* (2015, October). Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia conference on language, information and computation (pp. 73–78).
5. *Ivanin V., Artemova E., Batura T., Ivanov V., Sarkisyan V., Tutubalina E., Smurov I.* (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
6. *Bird, Steven, Edward Loper and Ewan Klein* (2009), Natural Language Processing with Python. O’Reilly Media Inc.
7. *Korobov M.* (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts (pp 320–332).
8. *Kuratov, Y., Arkhipov, M.* (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213.
9. *Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Brew, J.* (2019). Huggingface’s transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.

10. *Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.*
11. *Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Séaghdha, D. O., Padó, S., ... & Szpakowicz, S. (2010, July). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 33–38). Association for Computational Linguistics.*

TAXONOMY ENRICHMENT FOR RUSSIAN: SYNSET CLASSIFICATION OUTPERFORMS LINEAR HYPONYM-HYPERNYM PROJECTIONS

Kunilovskaya M. (maria.kunilovskaya@wlv.ac.uk),
Kutuzov A. (andreku@ifi.uio.no), **Plum A.** (a.j.plum@wlv.ac.uk)
University of Wolverhampton, UK
University of Oslo, Norway
University of Wolverhampton, UK

We present the description of our system that was ranked third in the noun sub-track of the Taxonomy Enrichment for the Russian Language shared task offered by Dialogue Evaluation 2020. Our best-performing system appears against the backdrop of other methods and their combinations attempted, and its results argue in favour of Occam's razor for this task. A simple supervised classifier was trained on static distributional embeddings of hyponym words as features and their numeric hypernym synset identifiers from the taxonomy as class labels. It outperformed more complicated approaches based on learning linear projections from hyponym embeddings to hypernym embeddings and returning synset identifiers for the nearest neighbours of the predicted vectors. Training specially tailored word embeddings for ruWordNet multi-word expressions proved to be one of the key factors for both approaches.

Key words: taxonomy enrichment, hypernymy relations, distributional semantics, word embeddings, projection learning, supervised machine learning

DOI: 10.28995/2075-7182-2020-19-474-484

ПОПОЛНЕНИЕ ТАКСОНОМИИ ДЛЯ РУССКОГО ЯЗЫКА: ЛИНЕЙНЫЕ ГИПО-ГИПЕРОНИМИЧЕСКИЕ ПРОЕКЦИИ ИЛИ КЛАССИФИКАТОР СИНСЕТОВ

В настоящей статье описывается способ расширения таксономии, который занял третье место в соревновании, объявленном в рамках Dialogue Evaluation 2020 (задача определения гиперонимических синсетов для существительных). Мы сравниваем наш наиболее эффективный подход с другими методами, которые были применены к решению поставленной задачи. Наши опыт и результаты свидетельствуют

в пользу выбора более простого подхода, который изначально не представлялся многообещающим. Таким методом оказался классификатор, обученный на векторах гипонимов и идентификационных номерах соответствующих гиперонимических синсетов. Его результат значительно выше чем для метода, основанного на выучивании линейной трансформации вектора гипонима в вектор гиперонима с последующим поиском слов (и идентификаторов их синсетов), семантически похожих на предсказанные гиперонимы. Для обоих подходов важную роль играет наличие качественных дистрибутивных векторных репрезентаций для многословных единиц тезауруса ruWordNet.

Ключевые слова: пополнение таксономии, гипо-гиперонимические отношения, векторные репрезентации, линейная трансформация векторов, машинное обучение с учителем

1. Introduction

This paper reports our approaches to the shared task offered by the Taxonomy Enrichment for the Russian Language competition within Dialogue Evaluation 2020¹ (see the overview [12]). Taxonomy enrichment is a natural language processing (NLP) task where a system is required to add new entries to an existing lexical database (ontology). Dialogue Evaluation 2020 provided a list of noun and verb unigrams to be attached to the appropriate hypernym synsets in ruWordNet [10]. At first glance, this task is derived from the lexical-level hypernymy detection and extraction challenges, when hypernymy, or a ‘is-a’ relationship, is established between two lemmas or two senses in a WordNet-format taxonomy. However, taxonomy enrichment can be re-cast as a classification task where a learner is presented with the target hyponym words and the associated hypernym synset IDs as the class labels. This approach avoids learning the semantic relations between words *per se*, but directly links the semantics of hyponyms to their generic concepts represented by the synsets.

1.1. Task description

The participants were asked to return up to 10 hypernym synsets from the existing taxonomy (ruWordNet) for every target hyponym from the test set. The systems were primarily evaluated with Mean Average Precision (MAP), which measures the accuracy of target synset detection. This measure was tuned to take into account accuracy per connected component, i.e. it returned higher results if the predicted synsets include at least one correct answer for each specified set of hypernyms associated with each sense of a polysemous hyponym. For example, the Russian word ‘yasli’ [ясли] has two distinct senses and two respective sets of hypernyms: (1) a day nursery and (2) a trough. A successful system is expected to predict synsets from both sets. Further more, the organisers limited the set of correct answers by the first and second order hypernyms, which are treated indiscriminately for evaluation purposes.

¹ <https://competitions.codalab.org/competitions/22168>

The paper is structured as follows: **Section 2** provides an account of the research that motivated some of our solutions. In **Section 3**, we introduce our methods and their settings, while in **Section 4** we describe the linguistic resources used to solve the task. Our experimental setup is discussed in **Section 5**, followed by the presentation of the results in **Section 6**.

2. Related work

Hypernymy, as a semantic relation between a specific and a generic term, underlies several interrelated NLP tasks: (1) hypernymy detection, when a provided pair of words is being classified as hypernymic or not, (2) hypernym discovery/extraction, when an input word is given a set of possible super-ordinates, (3) taxonomy induction/construction, when the required output is a hierarchy of concepts in a general or special domain and (4) taxonomy enrichment, when new items need to be integrated into an existing taxonomy. In this variety of hypernymy related tasks, the input lexical items are usually represented either by formal, morpho-syntactic and contextual properties of the queries (co-occurrence and pattern-based statistical approaches) or by word vectors within the distributional approach to capturing word semantics. Methodologically, in the former case a query's hypernym is directly extracted from a pattern match, while the latter approach follows one of two scenarios: the task can be cast either as a classification of word pairs (possibly constructed from the entire corpus vocabulary) that outputs the probability of hypernymy relation existing between the two words or as learning a transformation matrix from a hyponym vector to a vector in the subspace of possible hypernyms. There have been attempts to benefit from a combination of patterns-based information and vector representations.

The obvious limitations of the pattern-based approaches include lack of coverage, when many terms receive no candidate hypernyms, and noisy results, when patterns return figurative or broad generalisations. To give a curious example from our experiments, 'Самая серьезная **проблема** — **это человек**' ['The most serious **problem is (a) man**'], where the 'is-a' pattern returns 'man' as a hypernym for 'problem'. Often, such methodologies use lexico-syntactic patterns similar to, or adapted from those introduced by Marti Hearst in 1992 and named after her [5]. For example, Michael Oakes applied them to automatically extract hyponymy relations from a corpus of an English pharmaceutical texts [13]. Panchenko et al. successfully used sub-string matching and pattern searches to extract candidate hypernyms from raw text in four languages (but not for Russian) in the SemEval 2016 task of automatic taxonomy generation (Task 13) [14]. The key to their success was the introduction of a classifier-based filter for the noise returned by the patterns. Hearst patterns for Russian were adapted by Kristina Sabirova and Artem Lukanin in 2014 in an attempt to facilitate the creation of the Russian thesauri and to improve query expansion algorithms in information retrieval systems [16].

Another approach to represent words in the hypernymy-related tasks is to use word embeddings. However, based on the results of the SemEval-2016: Task 14 challenge, word embeddings are not very effective, if the hypernymy detection task is formulated as a classification task. In the SemEval-2016: Task 14 setting the participants

were provided with an out-of-taxonomy word sense and its definition and were expected to demonstrate high integration accuracy as to (i) the identification of the exact sense of the existing taxonomy entry that is hypernymic to the query, and (ii) the choice of the correct integration operation (add a new lemma to the existing synset or introduce a new concept). The best performing system cast the problem as a pairwise classification task to detect the correct hypernym for each hyponym [17]. Their classifier demonstrated only very moderate improvement after the feature set constrained to the formal and grammatical properties extracted from WordNet was extended to include the distributional semantics features. Overall, it is the only system (out of 13) that is slightly above the naive First Word, First Sense baseline, which picked the first word that matched the query part of speech (PoS) in the provided definition and attached the query to the first sense of it [6].

Instead of classification, Ustalov et al. implemented an approach that learns “a hypernym embedding on the basis of a hyponym embedding” [19]. It is known as projection learning and was initially applied in bilingual tasks like word translation. Ustalov et al. combined the projection learning system applied to hypernymy extraction from [4] and the negative sampling regularisation idea from [20]. They presented a distributional hypernym extraction system equipped with clustering and two options for negative sampling: their algorithm is either penalised for predicting vectors similar to the vector of the *input hyponym* (asymmetric regularisation that uses inverted hypernymy relations) or for predicting vectors similar to the vectors of the *input hyponym synonyms* (neighbour regularisation).

In essence, the task at hand is more in line with the hypernymy modelling as offered at SemEval-2018: Task 9 [3], which asked participants to retrieve suitable hypernyms from a target corpus (in our case, from ruWordNet) given an input term. The pairwise classification setup would require to evaluate all possible pairs of the ruWordNet senses in the respective PoS category for every test hyponym. This was the reason we did not use the pairwise classification approach, but went for projection learning and another kind of classification setup described below.

Note that a major practical drawback of using pre-trained word embeddings in any NLP task, including taxonomy enrichment, is the issue of out-of-vocabulary words (OOV). As we show in the next sections, this is particularly important in the context of the current shared task: more than half of noun lemmas in ruWordNet are actually multi-word expressions (MWE), which often lack representations in the pre-trained embeddings. **Subsection 3.2** describes how we handled this problem.

Now we move on to explain the approaches we devised based on this analysis in more detail, including most of the ideas discussed above.

3. Methodology

3.1. Baseline approach

All our approaches are based on distributional Continuous Skipgram or Continuous Bag-of-Words embeddings [11]. For the baseline system, we represent all pairs

of hyponym-hypernym words from the provided training data with the available pre-trained vectors, discarding all items not found in the embeddings vocabulary and ignoring connected components. Then we learn the optimal linear transformation of hyponym vectors into hypernym vectors and predict a hypernym vector for each new test word. We return the synset ids of k ruWordNet words most similar (based on vector cosine similarity) to the predicted vector. Duplicate ids in the list were removed. For the test words not found in the embedding model vocabulary, we return ids of the synsets that are most frequent as hypernyms (but not domains) in ruWordNet. As an alternative strategy for OOV words handling, we tried using **fastText** vectors [2], but they yielded lower results. We comment on the performance of a number of available pre-trained word embedding models using this setup in **Section 4**.

3.2. Improvements

Some improvement was gained when we enhanced our baseline projection learner with the negative sampling regularisation as suggested in [19]. Similar to their findings, we noticed better performance for the synonym based approach, when the model is penalised for predicting hypernym vectors which are close to ‘the semantic neighbourhood of the hyponym’.

An additional increase in performance was seen after we introduced the refinement of the projection learning results using hypernyms extracted with Hearst patterns. The list of the top k words that were found to be most similar to the predicted hypernym vector was reordered to give priority to the hypernyms matched by seven Hearst patterns, inspired by Sabirova and Lukanin [16]. We implemented this filter only for nouns, but the MAP results with the filter were consistently higher for all our settings and test sets.

The obvious limitation of most of the available pre-trained word embeddings is that many multi-word entities (MWEs) are missing from their vocabularies (if any MWEs are present there at all). MWEs account for about 57.4% of the ruWordNet training data word list for nouns and for 46% for verbs. Moreover, the inability to represent MWE rendered almost 19% of noun synsets entirely inaccessible for us at the prediction stage, because they do not have any one-word lemmas (‘senses’) attached to them. For example, it was impossible for us to predict ‘butterfly’ [ба́ттерфляй] as ‘a swimming stroke’ [стиль пла́вания]. We were sceptical of representing MWEs as averaged vectors of their components and decided in favour of tokenising them in a large corpus and training distributional embeddings from scratch on this tailored corpus. As expected, this yielded good results and increased the performance of the baseline system.

3.3. Classifier as an Alternative

The real boost in performance was achieved with the approach that did not initially look promising because of its simplicity. The error analysis for the previous methods demonstrated that though we were returning reasonably good results in terms of hypernym words, it was tricky to output the correct synset ids. For example, even if we predicted ‘bird’ [пти́ца] for ‘lapwing’ [чи́бис] and ‘art’ [иску́ство] for

‘folk-rhyme’ [частушка], our system did not score because the gold answers for these test words include only (5681-N, 7396-N), verbalised as ‘wild bird’ and ‘bird of passage’ and (108434-N, 7372-N), verbalised as ‘vocal (art) piece’ and ‘song, songlet’ respectively. To account for this, our simple classifier approach (see more on it in [Section 5](#)) **avoids learning relations between words** in the hyponym-hypernym pairs entirely. Instead, it directly uses hypernym synsets (or, rather, their ids) as classes of hyponym words (represented with their embeddings). At test time, the trained classifier was given a vector representation of a query word, and predicted a probability distribution over all classes (hypernym synsets from ruWordNet) for this vector. Our best submission includes 10 synset ids with the highest probabilities per test hyponym.

3.4. Failed Attempts

We attempted a number of other strategies to improve the results. This subsection contains a brief account of what did not work for this task.

As an approximate for Hearst patterns statistics, we used the corpus co-occurrence counts to refine our raw output for nouns and move the frequently co-occurring hypernyms up the list. To that end, we concatenated the four Russian corpora that were at our disposal (see Section 4 for details) and produced a frequency dictionary for sentence-based co-occurrences of the test words with each ruWordNet noun. If a lemma was seen in the top n predicted hypernyms and also in the top k co-occurring items for the test word, we moved it up the prediction list. The gains from this laborious effort did not live up to our expectations: The improvements were only marginal.

Our initial attempt to avoid the intermediary step of extracting hypernym words consisted of producing synset vectors as average vectors of all lemmas attached to the synset (‘de-worded approach’). The idea was that it would allow us to directly link hyponym words to the synset ids. However, the results for this setting were even lower than the baseline.

Finally, we unsuccessfully tried to reduce the number of same-level concepts predicted as hypernyms by weeding out the direct hyponyms of the higher level concepts in the output, hoping to move the true hypernyms up the list of predictions. The attempts to disambiguate hypernyms in the output and return only relevant ids instead of all of them (based on measuring the similarity between the predicted vector and the averaged synset id vector) did not work either: not the least because this improvement, even if successful, affected only 30% of the output.

For brevity, we report the settings that actually yielded some improvements and omit negative results from [Table 1](#) below.

4. Linguistic Resources

4.1. ruWordNet as a dataset

RuWordNet is a lexical database in the WordNet format that was semi-automatically converted from RuThes-lite, a publicly available part of the linguistic ontology of the Russian language originally created for NLP applications [10]. According to the authors of RuThes-lite, it includes a subsection of the full database and comprises words that are particularly frequent in the news [9]. RuWordNet has inherited some of the properties of RuThes that are important for the task discussed here. It contains a considerable number of multi-word items, with many synsets having no single-word lemmas (unigrams) attached to them. The total number of all hyponym-hypernym pairs for nouns in the training data (provided by the organisers of the shared task) is 431,937; filtering out MWEs leaves us with 94,115 pairs. Note that polysemous words in ruWordNet are assigned to several synsets, and have different sets of hypernyms. However, our estimate, based on the training data, indicates that 76.9% of the listed hypernyms are monosemous.

4.2. Comparative Performance of Embeddings

We have tested nine available pre-trained embedding models, produced within different frameworks on raw and PoS-tagged Russian corpora of various size and make-up. Most models were obtained from the RusVectores repository [8]. Our observations indicated that (1) **fastText** models trained on the same corpora performed worse than **word2vec** (for example, MAP on the public test set (nouns) for **word2vec** trained on UD-tagged Russicum Maximum corpus and **fastText** vectors learnt on the lemmatised version of the same corpus compare as 0.2540 vs. 0.1406, even though **fastText** returns no OOV), and (2) lemmatised and tagged embeddings outperformed those trained on raw text (including extremely large vectors from Russian Distributional Thesaurus project [15]). The best performing vector model was a **word2vec** skipgram model trained on the Araneum Russicum Maximum corpus [1] (about 10 billion words). It significantly outperformed similar vectors (size 300, window 2) trained on concatenated Wikipedia and Russian National Corpus, the news corpus provided by the shared task organisers and on the Taiga corpus².

4.3. Extended MWE Support: Training our own Vectors

Co-occurrence statistics and Hearst patterns results were produced from a large corpus obtained by the concatenation of the four Russian corpora: Araneum Russicum Maximum, Russian Wikipedia, Russian National Corpus (both the main and newspaper parts), and the news corpus. All texts were lemmatised and PoS-tagged with UDPipe [18]. The overall size of the pre-processed corpus amounted to 900 million sentences, 8.7 billion tokens (after stop words removal). The same corpus was used to train customised embeddings with extended MWE support. To this end,

² https://tatianashavrina.github.io/taiga_site/

we tokenised 39,000 noun phrases (NP) and 12,000 verb phrases (VP) from ruWordNet training data in this corpus, obtaining multi-word tokens like *'wild_ADJ::bird_NOUN'*. A standard CBOW embedding model was trained on the resulting corpus for three epochs, with a symmetric context window of size 2. Note that we aimed at covering as many ruWordNet lemmas as possible. For that, we set an unusually high vocabulary size of approximately 1 million top frequent word types (this essentially means each word with an absolute frequency of more than 60 received its vector). This measure rendered more than 85% of NP and 62% of VP from ruWordNet accessible and boosted the size of the effective training data for our classifier.

5. Experimental Setup

5.1. Projection Learning Approach

The baseline hypernym predictions for test words were obtained by finding (via exact normal equation solving) a linear projection matrix which minimises the sum of the squares of the differences between the target (hypernym) vectors and those returned by the multiplication of the projection matrix and the source (hyponym) vector; this largely follows the setup described in [7]. Experiments returned higher scores for normal equation solving without the regularisation term λ .

To learn projections with the loss function regularised by negative sampling using reversed hypernymy relations or synonyms, we adapted the code published by Ustalov et al. (2017) [19]. For the asymmetric regularisation we used the hyponym-hypernym word pairs from the provided training data; for synonyms as negative examples we used senses from the same synset as they appear in the training data.

5.2. Classification Approach

To obtain a model that predicts the hypernym synsets directly, we implemented a standard neural classifier. It uses the training pairs of hyponym word vectors and the associated hypernym synset ids. We discarded ids that appeared less than five times in the training data. The size of the training data, when using our own vectors with extended MWE support (see Section 4.3), is 74,235 train instances (hyponym words) and 4,068 classes (synsets) for nouns. The classifier has one hidden layer of size 386, a dropout of 0.1, a ReLU activation function and a softmax output layer. It was trained with batch size 32, with the maximum number of epochs set to 25 (in practice, the training almost always stopped at around epoch 15, since the accuracy stopped increasing). At test time, the trained classifier was given a vector representation of a new hyponym and predicted the probability distribution over all possible classes (hypernym synset ids) for this vector. We return 10 synset ids with the highest probability values.

Note that the classifier was implemented as a shallow feed-forward neural network, but in fact, comparable results could be achieved with many other non-neural classifiers (logistic regression, etc). We chose a neural architecture because of the comparative ease of implementation.

6. Results

In Table 1, we report the mean average precision (MAP) scores for our methods described above on the private and public test sets (and the train-test split provided by the organisers) for nouns. Our best system (using tokenised MWE embeddings and the classifier approach) was eventually ranked 3rd (of 17 participants total) in the shared task leader board on the test set, with the MAP of 0.4976. The systems ranked 1st and 2nd showed MAP of 0.5522 and 0.5054 respectively. For verbs the performance of our approach is less competitive, with MAP of 0.2470 against the best result achieved in the competition of 0.4483.

Table 1: MAP scores for the methods we used and their modifications on noun test sets

Approach	Araneum embeddings			embeddings with MWE		
	private	public	provided	private	public	provided
baseline projection	0.2615	0.2470	0.2251	0.2846	0.2925	0.2693
+ Hearst patterns	0.2653	0.2520	0.2298	0.2866	0.2942	0.2736
neg-sampling synonymy [19]	0.2604	0.2484	0.2384	0.2974	0.3153	0.2925
+ Hearst patterns	0.2674	0.2506	0.2397	0.2996	0.3178	0.2959
classifier	0.4590	0.4219	0.3236	0.4976	0.4706	0.5105

The figures in the right-hand part of the table are consistently higher than in the left-hand part. The difference indicates the gains in performance attributable to the customised embeddings trained on the version of the corpus with the tokenised MWEs.

The results for several test sets indicate how stable the improvements are and link the performance to the specificity of the test set. In our scenario, this specificity is defined by the number of OOV items, as our solutions are contingent on the embeddings vocabulary size. With the standard vectors trained on Araneum Maximum, the number of OOV items was 3%, 4% and 22% for private, public and provided test sets respectively, which obviously affected the performance of projection-learning approaches. Our customised vectors have a higher coverage and close the gap between the number of OOV among test words, returning only 0.26%, 0.66% and 5.66% for the same test sets.

Table 1 highlights that refining the order of predicted hypernyms within the range of top 15 and 50 hypernym-ids tuples with hypernyms extracted using Hearst patterns (for Araneum and customised vectors respectively) consistently improves the results, albeit marginally: most changes are in the third decimal digit. Enforcing projection learning with negative sampling based on synonyms secures additional small gains in most settings. We saw a significant precision growth in this task only when we fell back to a simpler and more straightforward classification approach, which required no fine-tuned thresholds or filtering.

7. Conclusion

This paper demonstrates the limited effectiveness of some traditional approaches to taxonomy enrichment based on Hearst patterns or projection learning. In our case, the most successful method to solve the given task of hypernym finding for Russian nouns was a comparatively simple classifier trained on static word embeddings of hyponyms as features and ruWordNet noun synset ids as classes, returning the 10 most probable synsets at test time for each query. The code and the word embedding model we used to solve the task are available at <https://github.com/kunilovskaya/hypohyper/>.

It is important that substantial gains in performance were obtained from using word embeddings trained on a large specifically tailored Russian corpus with MWEs from ruWordNet merged into one token. This ensured better coverage of hyponym single-word and multi-word entities, and, as a result, significantly more training data for the classifier.

Other techniques which proved to be useful for us (but, eventually, not the best) were Hearst patterns and negative sampling with synonyms when learning the projection matrix.

References

1. *Benko, V.*: Aranea: Yet Another Family of (Comparable) Web Corpora Aranea: A Family of Comparable Gigaword Web Corpora View project Dictionary of Contemporary Slovak Language View project Aranea: Yet Another Family of (Comparable) Web Corpora. In: TSD 2014. pp. 257–264 (2014).
2. *Bojanowski, P. et al.*: Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*. 5, 135–146 (2017).
3. *Camacho-Collados, J. et al.*: SemEval-2018 Task 9: Hypernym Discovery. In: Proceedings of the 12th international workshop on semantic evaluation. pp. 712–724 (2018).
4. *Fu, R. et al.*: Learning Semantic Hierarchies via Word Embeddings. In: Proceedings of the 52nd annual meeting of the association for computational linguistics. pp. 1199–1209 Association for Computational Linguistics (2014).
5. *Hearst, M. A.*: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on computational linguistics-volume 2. pp. 539–545 Association for Computational Linguistics (1992).
6. *Jurgens, D., Pilehvar, M. T.*: SemEval-2016 Task 14: Semantic Taxonomy Enrichment. In: Proceedings of semeval-2016. pp. 1092–1102 (2016).
7. *Kutuzov, A. et al.*: Clustering comparable corpora of Russian and Ukrainian academic texts: Word embeddings and semantic fingerprints. In: Ninth workshop on building and using comparable corpora. pp. 3–10 (2016).
8. *Kutuzov, A., Kuzmenko, E.*: WebVectors: A toolkit for building web interfaces for vector semantic models. In: Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers. pp. 155–161 Springer International Publishing, Cham (2017).

9. *Loukachevitch, N. et al.*: RuThes-lite, a publicly available version of Thesaurus of Russian language RuThes. In: Computational linguistics and intellectual technologies: Papers from the annual conference “dialogue”. pp. 340–349 (2014).
10. *Loukachevitch, N. et al.*: Creating Russian Wordnet by Conversion. In: Computational linguistics and intellectual technologies: Papers from the annual conference “Dialogue”. pp. 22–32 (2016).
11. *Mikolov, T. et al.*: Distributed representations of words and phrases and their compositionality. In: Burges, C. J. C. et al. (eds.) Advances in neural information processing systems 26. pp. 3111–3119 Curran Associates, Inc. (2013).
12. *Nikishina, I. et al.*: RUSSE’2020: Findings of the First Taxonomy Enrichment Task for the Russian Language. In: Computational linguistics and intellectual technologies: Papers from the annual conference “Dialogue”. (2020).
13. *Oakes, M. P.*: Using Hearst’s Rules for the Automatic Acquisition of Hyponyms for Mining a Pharmaceutical Corpus. In: RANLP text mining workshop. pp. 63–67 (2005).
14. *Panchenko, A. et al.*: TAXI at SemEval-2016 Task 13: a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling. In: Proceedings of semeval-2016, pp. 1320–1327 (2016).
15. *Panchenko, A. et al.*: Human and Machine Judgements for Russian Semantic Relatedness. Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7–9. 221–235 (2017).
16. *Sabirova, K., Lukanin, A.*: Automatic Extraction of Hypernyms and Hyponyms from Russian Texts. In: AIST (supplement). pp. 35–40 (2014).
17. *Sejr Schlichtkrull, M., Martínez Alonso, H.*: MSejrKu at SemEval-2016 Task 14: Taxonomy Enrichment by Evidence Ranking. In: Proceedings of semeval-2016. pp. 1337–1341 (2016).
18. *Straka, M., Straková, J.*: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. pp. 88–99 (2017).
19. *Ustalov, D. et al.*: Negative sampling improves hypernymy extraction based on projection learning. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers. pp. 543–550 Association for Computational Linguistics, Valencia, Spain (2017).
20. *Yamane, J. et al.*: Distributional Hypernym Generation by Jointly Learning Clusters and Projections. In: Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers. pp. 1871–1879 (2016).

СЕМАНТИЧЕСКИЕ ЭФФЕКТЫ ВРЕМЕНИ ВО ВВОДНЫХ КОНСТРУКЦИЯХ С МЕНТАЛЬНЫМИ ГЛАГОЛАМИ¹

Кустова Г. И. (galinak03@gmail.com)

Институт русского языка им. В. В. Виноградова РАН,
Москва, Россия

Вводные конструкции с глаголом мнения (*как я думаю*) в статье рассматриваются как результат редукции главной клаузы: *Я думаю, что приглашение прислал профессор Уилер* → *Приглашение, как я думаю, прислал профессор Уилер*. Показано, что значение времени ментального глагола влияет на интерпретацию предложения. В настоящем времени *как я думаю* вводит предположение с нейтральным статусом: *Это произойдет, как я думаю, в самом ближайшем будущем* [Ю. Семенов] = 'неизвестно, Р или не-Р'; в прошедшем времени *как я думал* вводит неправильное предположение: *Дядя, который, как я думал, давно забыл о подаренных часах, воспринял эту новость болезненно* [Ф. Искандер] 'я думал, что забыл, а на самом деле — не забыл'.

Ключевые слова: парентетические (вводные) конструкции, ментальные глаголы, категория времени, категория лица

DOI: 10.28995/2075-7182-2020-19-485-499

SEMANTIC EFFECTS OF VERB TENSE IN PARENTHETICAL CONSTRUCTIONS WITH MENTAL VERBS

Kustova G. I. (galinak03@gmail.com)

Vinogradov Russian Language Institute of the Russian Academy
of Sciences, Moscow, Russia

Parenthetical constructions with verbs of opinion (*as I think*) are considered as the result of the reduction of the main clause: *Ja dumaju, chto priglasenie prislal professor Wiler* → *Priglasenie, kak ja dumaju, prislal professor Wiler*. The meaning of the mental verb tense affects the interpretation of the sentence.

¹ Работа выполнена при поддержке РФФИ, проект № 17-29-09154 офи_м «Динамика языковой системы: корпусное исследование синхронной вариативности и диахронических изменений в текстах разных типов». Примеры извлечены из Национального корпуса русского языка, www.ruscorpora.ru.

In the present tense, construction *as I think* introduces an assumption with a neutral status: *Eto proizojdet, kak ja dumaju, v samom blizhajshem budushchem* [Ju. Semenov]—‘no one knows, P or non-P’. In the past tense, construction *as I thought* introduces a wrong assumption: *Djadja, kotoryj, kak ja dumal, davno zabylo o podarenykh chasakh, vosprinjal etu novost' boleznenno*.

Keywords: parenthetical constructions, mental verbs, tense, person

1. Вводные предложения как редуцированные конструкции

Вводные (парентетические) конструкции изучаются с разных точек зрения и в разных аспектах, ср. [Appelt 1995], [Banik 2009], [Kehler 2002], [Urmson 1963], [Остроумова, Фрамполь 2009]. В данной работе они будут рассматриваться как результат редукции главной клаузы.

В русском языке существует несколько механизмов редукции главной клаузы с предикатом пропозициональной установки, т.е. механизмов понижения статуса матричного предиката и его вытеснения из главной клаузы. Один из них — перевод предиката с пропозициональной валентностью из главной клаузы в подчиненную (в относительное придаточное): *Петя еще вчера предупредил, что не успеет к началу праздника* — *Петя не успеет к началу праздника, о чем предупредил еще вчера*; *Многие опасались, что отпуск сократят* — *Отпуск сократили, чего многие опасались* (преобразования могут сопровождаться изменением модальных статусов пропозиций). Другой механизм — так называемые вводные (парентетические) конструкции (ВК). Считается, что вводные конструкции синтаксически не связаны с предложением, хотя семантически материал предложения заполняет пропозициональную валентность глагола вводной конструкции.

Данная работа посвящена свойствам вводных конструкций с *как* и глаголами мнения:

- (1) *Приглашение, как я думаю, было по инициативе профессора Уилера, который, вероятно, помнил меня по Тбилиси*
[А. Д. Сахаров. Воспоминания (1983–1989)].

Принято различать два вида мнения (см. [Дмитровская 1988], [Зализняк 1991]) — мнение-предположение (*Я думаю, что фильм хороший* — ‘не видел, но предполагаю, что понравится’) и мнение-оценку (*Я считаю, что фильм хороший* — ‘видел и оценил положительно’). В статье речь пойдет только о предположениях.

Основным материалом послужат предложения с глаголом *думать*, эпизодически будут привлекаться глаголы *бояться* и *надеяться*, которые, кроме предположения, дополнительно выражают еще оценку (отношение к ситуации Р): *Как он надеется, Р* — ‘Р ожидается и желательно’; *Как он боится, Р* — ‘Р ожидается и нежелательно’.

Вводные конструкции с глаголами мнения естественно считать результатом редукции главной клаузы сложноподчиненного предложения с изъяснительным придаточным: *Я думаю, что он придет* — *Он, я думаю, придет*. Можно

различать два вида ВК — вводные слова и вводные предложения. Вводными словами (сокращенно — ВСл) будем называть конструкции вида (я) *думаю* (*Он, я думаю, уехал*); они допускают личное местоимение, но его всегда можно опустить: *Он, думаю, уехал*. Вводными предложениями (сокращенно — ВПр) будем называть конструкцию с *как*: *Он, как я думаю, уехал*.

С синтаксической точки зрения удобно иметь редуцированную конструкцию с глаголом пропозициональной установки, чтобы «разгрузить» текст. Так, есть случаи, когда вводная конструкция позволяет разрешить синтаксический конфликт — например, если она вставляется в придаточное: *Вейнрейху отправил 5 экземпляров, которые, как я надеюсь, он уже получил*. Конфликт состоит в том, что на одно придаточное (*он получил 5 экземпляров*) претендуют две главные клаузы (*отправил 5 экземпляров* [*которые он получил*] и *надеюсь* [*что он получил*]), ср. разговорную островную конструкцию: *Вейнрейху отправил 5 экземпляров, которые я надеюсь, что он уже получил*. К одной и той же пропозиции могут относиться два оператора, ср.: *Если он, как мы надеемся, приехал утром, он уже отключил сигнализацию* ('если он приехал' & 'надеемся, что он приехал'). С другой стороны, у вводных конструкций есть определенные ограничения — не только по сравнению с главной клаузой, но даже по сравнению с другими редуцированными конструкциями, — например, они не допускают отрицания, ср.: **Граф, как никто не ожидал, не оставил завещания* — в отличие от относительных придаточных: *Граф не оставил завещания, чего никто не ожидал* (формально отрицание может находиться внутри ВК, но обычно оно стоит при импlicative глаголе, и семантически это «положительное» сообщение, ср.:

- (2) *Он был неграмотный, как не преминул бы заметить Бабель, до глубины души* [Лев Овруцкий. Мера закона и безмерность беззакония // Библиотека «Огонек», 1989] —

не преминул бы заметить → 'заметил бы').

Конструкция с *как* используется в том случае, если говорящий передает что-то известное ему из какого-то источника (источником может быть другое лицо или он сам). Особенно наглядно это показывают конструкции с глаголами речи: *как говорят*; *как сообщили*; *как сказал X*; *как я уже говорил*. *Как*, в общем случае, коррелирует с *так*: — *Как сказал Гераклит, все течет, все изменяется*. — *Он действительно так сказал? Как соотносится не только с пропозициональным так* [Пекелис 2019], но и с пропозициональным *это*: *Я так не говорил — Я этого не говорил*. Мы не можем подробно рассматривать эту проблему, отметим только, что если *это* — просто «механическое» вместилище antecedenta, *так*, будучи обстоятельством образа действия, включает идею выбора, ср. [Кустова 2019] и подчеркивает вариативность способов передачи информации (которая, тем самым, может быть передана неточно или даже, в предельном случае, — неправильно). Поэтому, выбирая ВПр с *как*, говорящий как бы хочет подчеркнуть, что он не несет ответственности за истинность пропозиции. Это относится не только к передаче речи (*как говорят*), но и к передаче мысли (*как он думает*).

Вводные предложения имеют больше возможностей, чем вводные слова, которые на шкале редукции дальше продвинулись в сторону модальных показателей и имеют существенные ограничения в наборе форм и грамматических значений.

ВСл допускают первое лицо: *Петя, думаю, уехал*, — но не третье: **Петя, думает, уехал*, т.к. в третьем лице невозможно вычислить субъекта. Однако и эксплицитно обозначить его тоже нельзя. Конструкции, обозначающие субъекта (**Петя, Маша думает, уехал*), неупотребительны, и в Корпусе не встретились (теоретически они могут встретиться, но должны быть квалифицированы как нарушение литературной нормы). Подчеркнем, что предложения типа *Петя, думает Маша, уехал* (с обратным порядком компонентов) — не вводные конструкции, а цитирование, т.е. прямая речь — в данном случае, внутренняя речь, см. [Сердобольская 2017] (*думает Маша* — так наз. слова автора, где сказуемое обязательно предшествует подлежащему). То же самое — если субъект обозначен в предыдущих предложениях: *Маша обыскала всю квартиру. Наверное, думает, я потеряла эту шкатулку* — это тоже прямая речь (= *Наверное, думает Маша, ...*).

Кроме того, ВСл имеют ограничение на время — не употребляются в прошедшем времени: *Маша, я думал, уехала* (при этом прошедшее время *я думал* возможно в главной клаузе при бессоюзном присоединении придаточного: *Я думал, вы уехали, но не Вы, я думал, уехали*).

ВПр не имеют ограничений на лицо и время, т.е. допускают, наряду с 1-м, 3-е лицо:

(3) *За эти годы еще два раза хватала за блесну «та самая», как он думает, щука* [А. А. Ливеровский. Журавлиная родина. Рассказы охотника (1966)],

и, наряду с настоящим, прошедшее время:

(4) *Маленький пароходик так валяло со стороны на сторону, что семья из Сан-Франциско пластом лежала на диванах в жалкой кают-компании этого пароходика, закрыв от дурноты глаза. Миссис страдала, как она думала, больше всех* [И. А. Бунин. Господин из Сан-Франциско (1915)].

Сразу оговоримся, что мы не рассматриваем предложения с предикатами 2-го лица, обозначающими адресата (*как ты думаешь*), т.к. у ментальных и подобных глаголов (как, отчасти, и у физических глаголов) форма 2-го лица ведет себя особым образом и представляет собой отдельный класс употреблений, см. [Кустова 2018].

Далее мы рассмотрим ВПр с ментальными глаголами настоящего и прошедшего времени (при необходимости будет учитываться также параметр «лицо»: 1 л. vs. 3 л.).

2. Настоящее время

2.1. 1-е лицо: КАК Я ДУМАЮ

Путативные глаголы (мнение) противопоставлены фактивным (знанию), — хотя возможны семантические сдвиги в обе стороны. С одной стороны, как отмечал Дж. Урмсон, предложения с парентетическими глаголами мнения претендуют на истинность: говоря *Он, я думаю, уже дома*, человек в действительности выражает не предположение, а уверенность, своего рода знание [Urmson 1963]. С другой стороны, как отмечал Ю. Д. Апресян, фактивный глагол *знать* может, кроме знания, выражать, например, мнение (*Просто не знаю, что делать* — ‘не имею мнения’ [Апресян 1995]).

Предложения с конструкциями настоящего времени *как я думаю / как мы думаем* (в отличие от прошедшего времени, дающего эффект квазифактивности, см. ниже) выражают именно предположение, гипотезу (*надеяться* или *бояться* добавляют соответствующую оценку ситуации Р). Настоящее время *думаю* имплицитно, что в момент речи говорящий еще не располагает знанием (т.е. истинность Р в момент речи не известна): *Как я думаю, Р* = ‘я не знаю, Р или не-Р’ — иначе *думаю* было бы неуместно (невозможно). Пропозиция Р (вмещающее предложение) может относиться к любому временному плану:

- Р = БУДУЩЕЕ время:
 - (5) *Судно это /.../ собственность Геза, и он, как я думаю, не откажет мне в просьбе уступить вам каюту* [А. С. Грин. Бегущая по волнам (1926)]; *Это поможет юношам и девушкам в выборе профессий, связанных с автомобилизмом, и, как мы надеемся, вызовет приток квалифицированных молодых кадров* [«Техника — молодежи», 1977]
 - (6) *Я все-таки хотѣлъ бы и очень бы васъ просилъ, чтобы настоящая роль Загряцкаго осталась неразоблаченной. /.../ Но если, какъ я боюсь, это окажется практически невозможнымъ, — сказалъ онъ съ подчеркнутой ироніей, — то въдомству вашему тревожиться нечего* [М. А. Алданов. Ключ (1929)]
- Р = НАСТОЯЩЕЕ время:
 - (7) *Кажется, что граф Тотлебен поступает не по долгу своей присяги и, как я думаю, находится в переписке с неприятелем* [В. Я. Шишков. Емельян Пугачев (1934–1939)]
- Р = ПРОШЕДШЕЕ время:
 - (8) *С виду был симпатичный старичок, но полезных нам сведений дать не мог; он, как я думаю, сам ничего не знал и в торговых делах разбираться не мог* [Николай Варенцов. Слышанное. Виденное. Передуманное. Пережитое (1930–1935)].

В приведенных примерах *как я думаю* относится к предикату (ко всей пропозиции), т.е. сферой действия пропозициональной установки является вся

ситуация целиком. В этом случае конструкция 1-го лица настоящего времени *как я думаю* по сравнению с *я думаю* кажется, во-первых, избыточной: она эквивалентна вводу слову, и ее всегда можно заменить на более экономное [*я думаю*]; во-вторых, прагматически сомнительной: естественно с помощью такой «отстраняющей» конструкции передавать чужие мысли, а не свои собственные. Тем не менее такие примеры, пусть и немногочисленные, в Корпусе встречаются.

Однако в сфере действия пропозициональной установки *думаю* (о понятии сферы действия см. [Богуславский 1996]) может попадать не вся ситуация, а какой-то ее аспект, признак, обстоятельство:

- (9) *Сближение наше шло очень медленно. У меня это происходило скорее от неуверенности в себе, у него же, как я думаю, от осторожности* [И. Н. Потапенко. Несколько лет с А. П. Чеховым. К 10-летию со дня его кончины (1904–1914)] —

в сфере действия *думаю* — обстоятельство причины;

- (10) *Я помню только очень печальное происшествие, за которое Андрей избежал розог только благодаря горячему заступничеству дяди Коли, бабушки и, как я думаю, колебанию самой матери* [Н. А. Лухманова. Девочки (1894)] —

в сфере действия *думаю* — один из членов однородного ряда. В подобных случаях конструкция *как я думаю* выглядит оправданной, т.к. увеличивает коммуникативный вес выделяемого фрагмента, который должен быть противопоставлен остальной части высказывания.

2.2. 3-е лицо: КАК ОН ДУМАЕТ

Третье лицо, в общем случае, похоже на первое: говорящий может передавать и свое предположение (*как я думаю*), и предположение другого лица (*как он думает*), если оно ему откуда-то известно. Однако в случае третьего лица говорящий может отождествлять себя с этим лицом, а может сохранять дистанцию. Отсюда — две возможности.

В первом случае (тривиальном) говорящий просто передает мысль (предположение) другого лица, снимая с себя ответственность за истинность пропозиции Р: ‘субъект предположения не знает, Р или не-Р; говорящий тоже не знает, Р или не-Р’. Такая позиция наиболее естественна, если Р относится к будущему:

- (11) *Это будет большая комната, метров на тридцать, как он думает* [П. Ф. Нилин. Варя Лугина и ее первый муж (1936)]; *Кое-что о кредитах, которые Украина получает или, как там надеются, вот-вот получит: их сумма может подкатить аж к полусотне миллиардов долларов* [«Эксперт», 2014] —

или если пропозиция Р является общим суждением:

- (12) *И логика рассуждений такого человека предельно ясна: над ним, как он думает, господствуют неведомые силы* [В. В. Мезенцев, К. С. Абилянханов. Чудеса: Популярная энциклопедия (1991)].

Во втором случае (дистанция) говорящий знает, что не-Р, и конструкция *как он думает* используется как модальный показатель недоверности Р: Он понятия не имеет, что с ним происходит, носит на шее индейский, *как он думает*, амулет в виде летучей мыши [Полина Волошина, Евгений Кульков. Маруся (2009)] — если бы это был факт (амулет действительно индейский), *ВК он думает* не понадобилась бы.

Итак, в настоящем времени у ВК 1-го лица (*как я думаю*) нет импликации неправильного предположения, у ВК 3-го лица (*как он думает*) такая импликация может быть, но это лишь возможность.

3. Прошедшее время: КАК Я / ОН ДУМАЛ

3.1. Импликации вводных предложений в прошедшем времени

ВК с путативными глаголами прошедшего времени устроены иначе (чем ВК настоящего времени), как и конструкции с прошедшим временем в главной клаузе.

Начнем с рассмотрения главной клаузы. Формально прошедшее время установки означает ее отнесенность к прошлому и больше ничего. Однако это верно лишь для синхронной точки отсчета (о понятии точки отсчета см. [Падучева 1996]), когда повествователь находится в прошлом и не знает, как развивались дальнейшие события:

Я / он думал, что последний автобус уже ушел, и пошел пешком — ушел или не ушел последний автобус, остается неизвестным. Пропозиция может быть вообще неверифицируемой, ср.

- (13) *В нём было, как я думал, какое-то своё, крепкое знание жизни* [Максим Горький. В людях (1915–1916)] —

но наблюдатель-повествователь все равно находится в прошлом.

При ретроспективной точке отсчета (когда повествователь находится в настоящем и знает, как развивались события вплоть до момента речи) прошедшее время *я думал, что Р* имплицирует: ‘сейчас уже не думаю → сейчас знаю, что не-Р’: *Я думал, что Петя улетел, а у него рейс отменили*. Если знание о Р не получено, должно сохраняться настоящее время ментального предиката: *Я думаю, что Петя улетел*. *Думал* имплицирует, что предположение Р было неправильным (т.е. приобретает свойства фактивного (квазифактивного) глагола, см. [Кустова 1998], ср. также [Шатуновский 1988]). Если же предположение было правильным, нужно сказать: *Я так и думал / так и знал, что Петя не улетит*.

Таким образом, у глаголов мнения в главной клаузе время показывает не только время установки, но и статус пропозиции. Эту особенность наследуют и вводные конструкции: *Как я думал, Р* означает ‘сейчас уже известно,

что не-Р', т.е. как я думал имплицитно предполагает неправильное предположение: *Я брел из последних сил, оступался, спотыкался. Но впереди у меня было, как я думал, тепло избушки, был чай и крепкий сон в тепле. Я пришел к тоне, когда начало темнеть. Избушка была пуста — рыбаки куда-то уехали* [Юрий Казаков. Нестор и Кир (1961)] — предположение 'меня ждет чай и крепкий сон в тепле' было неверным. Верное предположение должно обозначаться иначе — с помощью конструкции с отождествительным *и*: *Петя, как я и думал, уже вернулся.*

При этом правила вычисления сферы действия ментального предиката для вводной конструкции намного сложнее, чем для главной клаузы, и семантическая интерпретация предложений с вводными ментальными глаголами в прошедшем времени представляет специальную проблему с точки зрения автоматической обработки текста. Некоторые из этих правил мы рассмотрим ниже.

Сразу оговоримся, что в определенных контекстах пропозиция, связанная с ВПр прошедшего времени, может быть неверифицируемой, т.е. иметь нейтральный статус ('неизвестно, Р или не-Р'):

(14) *Преступник, как думали следователи, сам отправлял письма. Впрочем, доказать этого не удалось; Никита Ильич впервые поехал туда в 1962 году, чтобы, как он думал, просто изучать лексику — он хотел создать словарь Полесья* [С. М. Толстая, Л. Ермолинская. Этнолингвист Светлана Толстая: Нас обвиняли в том, что мы проповедуем религию (2015.10.27)];

Они поднялись на холм, откуда, как думал Петя, будет лучше видно — это мотивировка: 'потому что думал, что оттуда лучше видно', не важно, как на самом деле — Р или не-Р (то же верно и для главной клаузы: *Не знаю, как сейчас, но вчера Петя думал, что Вера уехала* — неизвестно, Р или не-Р).

Но мы будем рассматривать случай, когда прошедшее время путативного глагола имплицитно предполагает неверное предположение.

Прошедшее время глагола в главной клаузе имплицитно предполагает неверное предположение как в диалогическом, так и в нарративном режиме (о диалогическом и нарративном режиме см. [Падучева 1996]). Что касается вводной конструкции с *думал*, то в диалогическом режиме ее нельзя использовать для выражения неправильного предположения, ср.:

(15) — *Почему ты не поехал к Пете?* — *Я думал, что Петя в командировке* — здесь я думал имплицитно предполагает, что предположение ошибочное, и Петя не был в командировке, но:

(16) — *Почему ты не поехал к Пете?* — *Петя, как я думал, в командировке*

Дальше мы будем рассматривать только нарративный режим интерпретации высказываний с вводными предложениями.

ПРИМЕЧАНИЕ

Сразу заметим, чтобы больше не возвращаться к этому вопросу, что время вводящего предложения Р в рассматриваемых ниже примерах может быть разным (как и в примерах из [раздела 2](#)), но оно не влияет на интересующие нас

параметры интерпретации ситуации Р, — в частности, потому, что оно, с одной стороны, несамостоятельно (обычно это относительное время, которое вычисляется по определенным правилам), а с другой, финитный глагол может относиться к нереальной ситуации (которая не имела места):

- (а) *Сергей озвучил ее последние слова, которые, как он думал, она произнесла перед тем, как бросить трубку* [Фазиль Искандер. Морской скорпион (1977)];
- (б) *И эти места, как она думала, займут два других стажера, а ее поставят на подмену* [«Военный вестник Юга России» (Ростов-на-Дону), 2003.03.03] —

ситуация (а) [произнесение слов] должна была иметь место раньше некоторой точки отсчета (задаваемой ВК), ситуация (б) [назначение стажеров] — после некоторой точки отсчета, однако ни та, ни другая не имели места (см. следующий раздел); это точечные ситуации, обозначаемые глаголами СВ. Длительная или постоянная ситуация может обозначаться как прошедшим временем (как предшествующая времени рассказчика), так и настоящим временем (как синхронная времени персонажа):

- (17) *У шефа в шкафу, где, как все думали, лежали важные документы, на самом деле были спрятаны портреты в рамках* [Елена и Валерий Гордеевы. Не все мы умрем (2002)], ср.:
- (18) *... в шкафу, где, как все думали, лежат важные документы; Галечников и песков, которые, как я думал, слагают островные горы, видно не было* [«Знание — сила», 1999] —

постоянная ситуация; наконец, ВПр может вообще относиться не к глаголу, а к другому фрагменту семантической структуры:

- (19) *И я съездил и купил, только не на поезд, как она думала, а на самолет* [Виктор Слипечук. Зинзивер (2001)].

Есть две основные конструкции, в которых участвует ВПр как *X думал*, — без отрицания и с контрастивным отрицанием.

3.2. Предложения без отрицания

В предложениях без отрицания ВПр как *X думал* стоит перед ремой, которая и является неправильным предположением:

- (20) *Лидия Петровна пожала большими плечами, и они пошли, как Катерина думала, в соседнюю школу. Но неожиданно для нее школьную изгородь они миновали, продвинулись еще с полквартила вглубь, и остановились возле такого же решетчатого металлического забора, сквозь который хорошо было видно громоздкое трехэтажное здание* [Марина Зосимкина. Ты проснешься (2015)] —

то, что предположение неправильное, выясняется из последующего текста. Вот еще пример, где обыгрывается сходство звучания soap ‘мыло’ и soup ‘суп’:

(21) *Он [официант] несколько брезгливо пожал плечами и удалился, как я думал, на кухню. Вернулся он минут через двадцать. На подносе стояли пять бумажных стаканов с густой розовой жидкостью, отдающей мылом. Познакомившись с напитком поближе, я убедился, что это и было жидкое мыло, которое наш официант терпеливо слил из контейнеров в туалетных умывальниках [Александр Генис. Довлатов и окрестности (1998)]*

(русские посетители заказали суп, а официант принес мыло). Отрицание в таких предложениях теоретически возможно (ср.: *Петя, как я думал, не писал письмо в прокуратуру. Но потом выяснилось, что все-таки написал*), — но подобные примеры в Корпусе не встретились. Для нас существенно, что такое отрицание не контрастивное.

3.3. Предложения с контрастивным отрицанием

Если вводное предложение стоит после ремы-предположения, то в предложении, парадоксальным образом, выражается: (1) неправильное предположение P, (2) его отрицание не-P, кроме того, часто — (3) реальная ситуация Q:

(22) *Оказывается, он не отвалил, как я думал, а пошел к своей машине, которая немного в стороне стояла [Фазиль Искандер. Сандро из Чегема (1989)] —*

‘я думал, что P [отвалил], а на самом деле не P, а Q [пошел к своей машине]’;

(23) *Он охотно согласился, разумеется, за двойные прогоны, считая прогоны не за две версты, как я думал, а за двадцать [Т. Г. Шевченко. Прогулка с удовольствием и не без морали (1855–1858)];*

(24) *Дня три тому назад меня из 8-й роты перевели, и не в 4-ю, как я думал, а первую [А. Гнедин. Письма (1939–1941)]*

В отрицательных предложениях вводная конструкция должна стоять после отрицания:

Петя приехал не вечером, как я думал, а утром — ‘думал, что приехал вечером’; при другом порядке слов смысл предложения изменится на противоположный: *Петя, как я думал, приехал не вечером, а утром* — ‘думал, что приехал утром’.

Если положительное и отрицательное придаточное при матричном ментальном глаголе обозначают две разные (противоположные) ситуации, ср.: *Я думал, что Петя приехал* и *Я думал, что Петя не приехал*, то положительное и отрицательное предложение с вводной конструкцией (в разных позициях) обозначают одну и ту же ситуацию:

(25) *Меж тем Тоббоган согласился сыграть — из вежливости, как я думал, — но когда оба мы выложили на стол по несколько золотых, его глаза выдали игрока* [А. С. Грин. Бегущая по волнам (1926)],

ср.: *Тоббоган согласился сыграть из вежливости, как я думал* [но потом выяснилось, что он игрок] VS. *Тоббоган согласился сыграть не из вежливости, как я думал, в потому что он игрок.*

При этом предложения без отрицания (*Тоббоган согласился сыграть из вежливости, как я думал*) содержат только неправильное предположение. Опровержение появляется в дальнейшем повествовании (или реконструируется). Предложения с отрицанием состоят как бы из двух слоев и содержат, парадоксальным образом, как неправильное предположение (P), так и его опровержение (отрицание — не-P), ср.:

(26) *Она призналась мне, что днем не сидела в номере, как я думал, а ходила к дому Добрынина...* [Вера Белоусова. По субботам не стреляю (2000)]:

— ‘я думал, что] она сидела в номере (P)’;

— ‘она не сидела в номере (не-P) (а Q: ходила)’

Т.е. в одном и том же предложении содержится и пропозициональная установка (ложное предположение), и сообщение об истинном положении дел.

При этом возникает конфликт: *как я думал* должно относиться к неправильному предположению P, а формально в предложении сочетается с не-P (что соответствует реальному положению дел). Конфликт разрешается за счет того, что отрицание контрастивное и не входит в сферу действия ВПр *как я думал*.

Еще один вариант неправильного предположения — предложение с антонимами, ср.:

(27) *Однако эта трудновыполнимая, как я думал, задача оказалась на удивление простой* [Евгений Рубин. Пан или пропал. Жизнеописание (1999–2000)]

Эта трудная, как я думал, задача оказалась простой — здесь компонентам не-P и Q соответствуют антонимы, ср.: *Задача оказалась не трудной, как я думал, а простой.*

Главная клауза тоже может входить в контрастивную конструкцию, но эта конструкция будет устроена противоположным образом: *Я думал, что Петя уехал, а он не уехал / остался* (ср.: *Петя не уехал, как я думал, а остался*), т.е. пропорция следующая:

(а) *Я думал, Петя уехал*

(б) *Петя не уехал, как я думал...*

(в) *Петя, как я думал, уехал* [однако потом выяснилось: не уехал]

Вводный глагол *думал* в прошедшем времени предполагает, что предположение ошибочное, при этом, в зависимости от позиции, он может формально присоединяться как к неправильному утверждению (*уехал*), так и к правильному утверждению (*не уехал*). Если ВПр стоит после ремы с контрастивным отрицанием, ср. (б), это отрицание не входит в сферу действия ВПр (‘не [уехал, как я думал = думал, что уехал]’); если ВПр стоит перед ремой, ср. (в), то рема входит целиком

в сферу действия ВПр, а наличие в ней отрицания — обычного или контрастивного — роли не играет: (i) *Петя, как я думал, уехал*; (ii) *Петя, как я думал, не уехал* (= остался); (iii) *Петя, как я думал, не уехал, а остался* — во всех трех случаях предположение неверное, причем (ii) ('не уехал') — вариант (iii) ('не уехал, а остался').

ПРИМЕЧАНИЕ

Интересно, что только у глаголов мнения прошедшее время имплицитно ложность пропозиции. У глаголов речи из прошедшего времени не следует ложность утверждения, поскольку это просто цитация, передача чужих слов:

(28) *Иванов, как он утверждал на допросе, вернулся в 9 часов.*

Говорящий снимает с себя ответственность, передавая речь другого лица, — но не обязательно сомневается в факте. Именно поэтому прошедшее время можно заменить на настоящее: *Иванов, как он утверждает, вернулся в 9 часов* — что невозможно для глаголов мнения (*как я думал* ≠ *как я думаю*).

Еще одна сложность возникает при интерпретации формально одинаковых предложений с конструкциями *как я думал*, которые вводят неправильное предположение, и конструкциями *как я и думал* (конструкциями подтверждения с отождествительным *и*), которые вводят правильное предположение:

Петя, как я думал, пришел — 'думал — пришел, реально — не пришел';

Петя, как я и думал, пришел — 'думал — пришел, реально — пришел';

Вероятность смешения двух видов конструкций возникает в предложениях с отрицанием.

Рассмотрим исходное предложение:

(29) *Она призналась мне, что днем не сидела в номере, как я думал, а ходила к дому Добрынина...* [Вера Белоусова. По субботам не стреляю (2000)].

Альтернативная реальная ситуация («она выходила») не обязательно обозначается в предложении — она может домысливаться, тогда предложение существует в «коротком» варианте: *Она не сидела в номере, как я думал* (домысливается 'она куда-то выходила').

В результате, при определенном порядке слов, конструкции неправильного предположения и конструкции подтверждения (т.е. правильного предположения) формально могут совпадать, что вызывает трудности при интерпретации, в том числе — при автоматической обработке текста:

(а) *Она не сидела в номере, как я думал*

(б) *Она не сидела в номере, как я и думал*

ВПр *как я думал* и *как я и думал*, теоретически, должны относиться к противоположным ситуациям: *как я думал* вводит неверное предположение 'думал, что сидела', *как я и думал* — верное предположение 'думал, что не сидела'. Однако формально они относятся к одному и тому же отрезку текста «не сидела». Более того, этот отрезок текста описывает одну и ту же ситуацию «она не сидела в номере», и в обоих предложениях эта ситуация имела место.

На семантическом уровне эти предложения устроены, конечно, по-разному: в (а), как уже говорилось выше, отрицание не входит в сферу действия вводного компонента, в (б), напротив, входит.

Почему конструкция предположения имеет обе формы времени (*как я думал / как я думаю*), а конструкция подтверждения — только прошедшее (*как я и думал*), но не настоящее (**как я и думаю*)? Потому что в первом случае (*как я думаю*) в момент речи знания еще нет, и установка *думаю* сохраняется; во втором — конструкции подтверждения — в момент речи знание уже есть, и *знаю*, как более «сильный» предикат, в соответствии с постулатом Грайса [Grice 1975], исключает более слабое *думаю*.

Вообще, в связи с рассмотренным материалом естественным образом встает вопрос о правилах вычисления конвенциональных и неконвенциональных импликатур (ср. об этом [Blakemore 1989]; [2000]; [2001]; [Potts 2002a]; [2002b]; [2003]; [2005]). Не имея возможности подробно разбирать этот вопрос, отметим лишь, что отменяемость импликации, в частности, зависит от наличия опровержения неверного предположения в последующем тексте (*Р, как я думал. Но потом... / Но оказалось...*), что с трудом поддается формализации.

Итак, мы показали, что в предложениях с ментальными предикатами статус пропозиции зависит не только от их семантического класса, но и от их грамматических параметров. Кроме того, редуцированные предикации вводных конструкций взаимодействуют с пропозицией более сложным образом, чем главные клаузы, и это должно отражаться в правилах интерпретации таких конструкций. Наконец, рассмотренный материал демонстрирует разные фазы процесса редукции исходной «полноценной» синтаксической единицы к грамматическому показателю, и вводные предложения на этой шкале занимают промежуточное положение между главными клаузами и вводными словами.

References

1. Appelt, D. E. (1985), *Planning English sentences*, Cambridge University Press, Cambridge.
2. Apresjan, Ju. D. (1995), The problem of factivity: *znat'* and its synonyms [Problema faktivnosti: *znat'* i ego sinonimy], Apresjan, Ju. D. Selected Works [Izbrannye trudy]. V. 2, *Jazyki russkoj kul'tury*, Moscow, pp. 405–433.
3. Banik, E. (2009), Parenthetical Constructions — an Argument against Modularity, Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks, ACL-IJCNLP 2009, Suntec, Singapore, 6 August 2009. 2009 ACL and AFNLP, pp. 46–53.
4. Blakemore, D. (1989), Denial and Contrast: A Relevance Theoretic Analysis of 'but', *Linguistics and Philosophy* 12, pp. 15–37.

5. *Blakemore, D.* (2000), Procedures and Indicators: ‘Nevertheless’ and ‘But’, *Journal of Linguistics* 36, pp. 463–486.
6. *Blakemore, D.* (2001), *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*, Cambridge Studies in Linguistics 99, Cambridge: Cambridge University Press.
7. *Boguslavskij, I. M.* (1996), The Scope of lexical units [Sfera dejstvija leksicheskikh edinic], *Jazyki slavyanskoy kul’tury*, Moscow.
8. *Dmitrovskaya, M. A.* (1988), Knowledge and opinion: the image of the world, the image of man [Znanie i mnenie: obraz mira, obraz cheloveka], *Logical analysis of language. Knowledge and opinion [Logicheskij analiz yazyka. Znanie i mnenie]*, Nauka, Moscow, pp. 6–18.
9. *Grice, H. Paul* (1975), *Logic and Conversation*, Peter Cole and Jerry Morgan (eds.), *Syntax and Semantics*, New York: Academic Press, volume 3: *Speech Acts*, pp. 43–58.
10. *Kehler, A.* (2002), *Coherence, Reference and the Theory of Grammar*. CSLI.
11. *Kustova, G. I.* (2018), Mental predicates in metatext [Mental’nye predikaty v metatekstovykh konstrukcijah 2-go lica], *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii]*, RGGU, Moscow, pp. 380–390.
12. *Kustova, G. I.* (2019), Conceptualization of non-fully controlled situations: verbs and pronouns [Konceptualizacija ne polnost’ju kontroliruemykh situacij: glagoly i mestoimenija], *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii]*, RGGU, Moscow, pp. 340–350.
13. *Kustova, G. I.* (1998), Some problems of describing mental predicates [Nekotorye problemy opisaniya mental’nykh predikatov], *Scientific and technical information. Series 2 [Nauchno-tehnicheskaja informacija. Serija 2. Informacionnye processy i sistemy]*, VINITI, Moscow, pp. 22–28.
14. *Ostroumova, O. A., Frampol’, O. D.* (2009), *Dictionary of Parenthetical Words, Combinations, and Sentences [Slovar’ vvodnykh slov, sochetanij i predlozhenij]*, Moscow.
15. *Paducheva, E. V.* (1996), *Semantic investigations [Semanticheskie issledovaniya]*, *Jazyki russkoi kul’tury*, Moscow.
16. *Pekelis, O. E.* (2018), Complement and adverbial at the same time: on a pronoun that replaces the sentential argument [Komplement i adverbial odnovremenno: o mestoimenii tak, zameshchayushchem sentencional’nyj aktant], *Typology of morphosyntactic parameters [Tipologiya morfosintaksicheskikh parametrov]*. Moscow.
17. *Potts, C.* (2002a), The Lexical Semantics of Parenthetical-As and Appositive Which, *Syntax* 5, pp. 55–88.
18. *Potts, C.* (2002b), The Syntax and Semantics of As-Parentheticals, *Natural Language and Linguistic Theory* 20, pp. 623–689.

19. Potts, C. (2003), Conventional Implicatures, a Distinguished Class of Meanings', Gillian Ramchand and Charles Reiss (eds.), *The Oxford Handbook of Linguistic Interfaces*, Studies in Theoretical Linguistics, Oxford: Oxford University Press.
20. Potts, C. (2005), The logic of conventional implicatures. Oxford, UK: Oxford University Press.
21. Serdobol'skaja, N. V. (2017), Conjunctionless actant sentences with the verb *dumat'* in Russian [Bessojuznye aktantnye predlozhenija s glagolom *dumat'* v russkom jazyke], *Topics in the study of language [Voprosy jazykoznanija]* 5, pp. 7–35. (In Russ.)
22. Urmson, J. O. (1963), Parenthetical verbs, Ch. E. Caton (ed.) *Philosophy and ordinary language*. Urbana, Chicago, London: University of Illinois Press, pp. 220–240.
23. Zaliznyak, Anna A. (1991), *Schitat'* and *dumat'*: two kinds of opinions [*Schitat'* i *dumat'*: dva vida mneniya], *Logical analysis of language. Cultural concepts [Logicheskij analiz jazyka. Kul'turnye koncepty]*, Nauka, Moscow, pp. 187–194.

SHIFTRY: WEB SERVICE FOR DIACHRONIC ANALYSIS OF RUSSIAN NEWS

Kutuzov A. (andreku@ifi.uio.no)

University of Oslo, Oslo, Norway

Fomin V. (wadimiusz@gmail.com)

National Research University Higher School of Economics,
Moscow, Russia

Mikhailov V. (Mikhaylov.V.Nikola@sberbank.ru)

National Research University Higher School of Economics;
Sberbank, Moscow, Russia

Rodina J. (julia.rodina97@gmail.com)

National Research University Higher School of Economics,
Moscow, Russia

We present the ShiftRy web service. It helps to analyze temporal changes in the usage of words in news texts from Russian mass media. For that, we employ diachronic word embedding models trained on large Russian news corpora from 2010 up to 2019. The users can explore the usage history of any given query word, or browse the lists of words ranked by the degree of their semantic drift in any couple of years. Visualizations of the words' trajectories through time are provided. Importantly, users can obtain corpus examples with the query word before and after the semantic shift (if any). The aim of ShiftRy is to ease the task of studying word history on short-term time spans, and the influence of social and political events on word usage. The service will be updated with new data yearly.

Key words: diachronic word embeddings, semantic shifts, web service, Russian, news

DOI: 10.28995/2075-7182-2020-19-500-516

SHIFTRY: ВЕБ-СЕРВИС ДЛЯ АНАЛИЗА ДИАХРОНИЧЕСКИХ ИЗМЕНЕНИЙ СЛОВ В НОВОСТНЫХ ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

Кутузов А. (andreku@ifi.uio.no)

Университет Осло, Осло, Норвегия

Фомин В. (wadimiusz@gmail.com)

Высшая Школа Экономики, Москва, Россия

Михайлов В. (Mikhaylov.V.Nikola@sberbank.ru)

Высшая Школа Экономики; Сбербанк, Москва, Россия

Родина Ю. (julia.rodina97@gmail.com)

Высшая Школа Экономики, Москва, Россия

В статье представлен веб-сервис ShiftRy, который помогает анализировать диахронические изменения в употреблении слов в российских новостных текстах. Мы используем диахронические дистрибутивно-семантические векторные модели, обученные на большом русскоязычном новостном корпусе, который включает в себя тексты за временной период с 2010 по 2019 годы. Пользователь веб-сервиса имеет возможность не только исследовать семантическую историю любого заданного слова, но и проанализировать списки изменившихся слов, упорядоченных по степени семантических изменений между двумя рассматриваемыми годами. Кроме того, ShiftRy предоставляет визуализации семантических «траекторий» движения слов во времени и генерирует корпусные примеры употребления заданного слова до и после семантического сдвига, если он имел место. Исследователи истории языка могут использовать ShiftRy для облегчения анализа динамики словоупотребления на коротких промежутках времени, а также для изучения влияния социально-политических событий на семантику русской лексики. Мы планируем обновлять ShiftRy ежегодно.

Ключевые слова: диахронические эмбединги, семантический сдвиг, веб-сервис, русский язык, корпус новостных текстов

1. Introduction

In this paper, we describe *ShiftRy*: the web service aimed to help the analysis of temporal shifts in word usage in Russian news texts¹. It leverages several existing methods of semantic shift detection using diachronic word embedding models.

Words change their meaning over time. These processes can be triggered either by linguistic or by social and cultural causes [6]. In the latter case, it is often not a discrete lexicographic shift (acquiring a new sense, losing an old one or changing an existing one), but rather a change in contextual usage, an attitude, or associations bound to an object in public opinion. A typical example is a country name which can acquire new associations after a military campaign starts in this country. Such changes are often short-term and can be traced on the time periods spanning across years or even months (unlike ‘linguistic’ semantic shifts which usually occur on the scale of decades or centuries). This is the most common type of semantic change found in our material: Russian news texts.

News texts are deeply interconnected with social, cultural and technological dynamics of human society. Additionally, time-annotated news corpora are relatively easy to obtain. Thus, applying diachronic computational semantic methods to this type of texts is interesting both for evaluating the corresponding algorithms and for getting insights about changes in society reflected through the lens of media.

ShiftRy performs this kind of analysis on Russian news texts published in the time span between 2010 and 2019 (in the future, it will be updated yearly). Russian material is attractive for our purpose, since it includes both state-sponsored media always supporting the stance of the authorities, and relatively free media presenting alternative points of view (see more on that in **Section 3**). Additionally, it provides numerous examples of word usage drift, caused by political and social events in the last decades, both outside and inside Russia.

The intended type of *ShiftRy* usage is asking it about a particular query word. The service produces the word’s semantic trajectory across the whole 2010–2019 time span, or across any sub-span determined by the user. It also tries to guess whether any substantial semantic shifts had happened in the course of these years (see more about the classifier performing this job in **subsection 5.1**). For deeper analysis, the user can inspect the lists of the nearest neighbors (semantically similar words) for the query word in the distributional representations trained on different time periods. Additionally, we provide samples of sentences from the corresponding corpora containing the query word, so that its real contexts can be examined. These examples are mostly intended to be used with the words which had undergone a semantic shift. For this reason, they are sampled in a way which maximizes the semantic difference between examples from different time bins (see **subsection 5.3**).

¹ <https://shiftry.rusvectors.org>

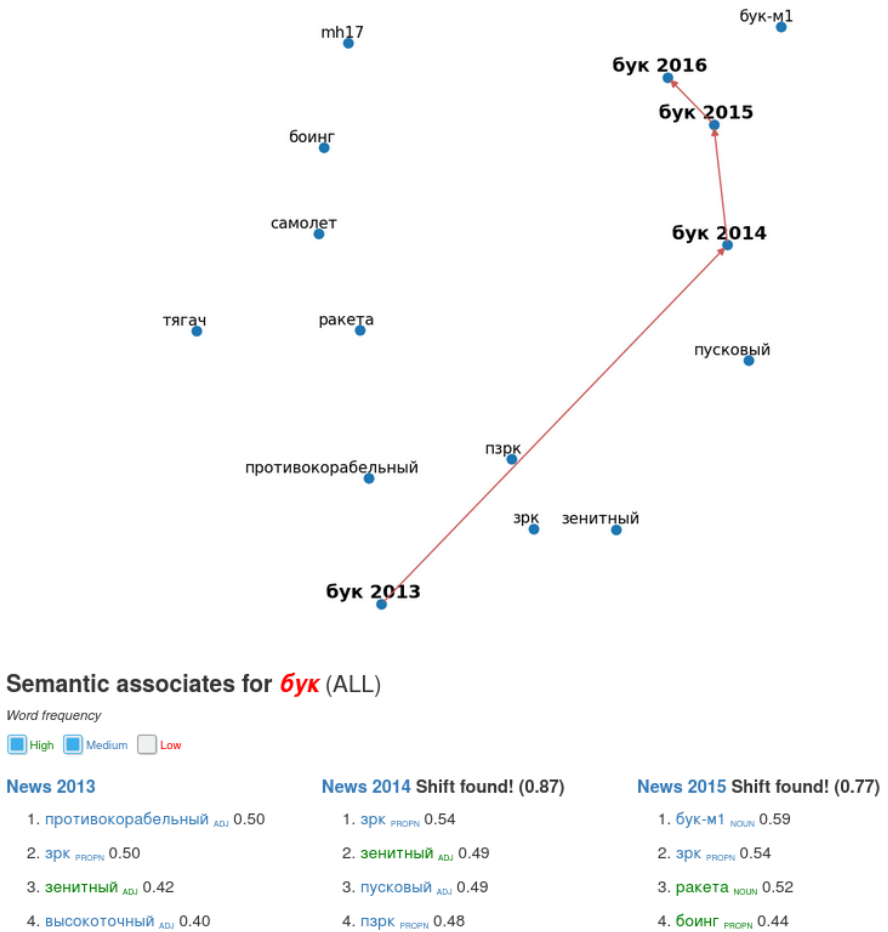


Figure 1: *ShiftRy* interface with the semantic trajectory and time-specific nearest neighbours of the Russian proper name ‘Бук’ ‘Buk’ (missile system) from 2013 to 2016

Figure 1 shows a screenshot of *ShiftRy* interface, featuring the analysis of the Russian proper name ‘Бук’ denoting the ‘Buk’ missile system (or a type of trees, but this sense is rare in news texts). Note how its typical neighbours drift from other missile systems to the words ‘боинг’ (‘Boeing’) and ‘МН-17’. This reflects the downing of the Malaysia Airlines Boeing Flight МН-17 in 2014, allegedly by Russia-sponsored Eastern Ukraine insurgents using this particular missile system. Note that this is a case of cultural context variation: the word did not acquire a new lexicographic sense. However, it had significantly changed its typical contexts and the meaning behind it had undoubtedly shifted for Russian speakers. Thus, we believe this to be an example of diachronic semantic change.

Our contributions in this paper are:

1. We provide a publicly available web service which allows to visually explore temporal changes in word usage, based on yearly Russian news texts (from 2010 to 2019).
2. A modification of Procrustes alignment for word embedding models is described, allowing to preserve full models' vocabularies, not only their intersection.
3. Diachronic word embedding models trained on the Russian news corpora and aligned using the aforementioned technique are released.
4. We describe a simple novel method of sampling sentences from corpora to better illustrate lexical semantic change.

The rest of the paper is organized as follows. In **Section 2**, we put our project in the context of the previous related work. **Section 3** describes the news corpora we employed for training word embedding models, which are in turn presented in **Section 4**, along with the explanation of our alignment method. **Section 5** provides a detailed account of *ShiftRy* features, including the word trajectory visualizations, the classifier and the corpus examples sampling. Finally, in **Section 6**, we conclude and account for the future work.

2. Previous Work

Recent years saw a growing amount of research aimed at automatically tracing diachronic semantic shifts and detecting their nature. This increase in popularity arguably started with the paper [6], following the advent of dense word embeddings [16], although some research was published even before that. Because of space limit, we do not list here all the relevant publications. Instead, we refer the reader to the surveys in [12] and [21], and to the proceedings of the first Workshop on Computational Approaches to Historical Language Change [20].

Among recent research on semantic shifts in Russian, one can mention [2], who extensively described 20 hand-picked examples of words having changed their meaning across two centuries, and [4], who evaluated several embedding-based semantic shift detection approaches on both long-term and short-term time spans. *ShiftRy* mostly uses the Procrustes alignment approach [5], which outperformed its competitors both for Russian in [4] and for English in [18].

Note that, in principle, semantic change modeling problem can also be cast in the context of lexical databases, like WordNet [22] for English or RuWordNet [13] for Russian. In this case, a semantic shift means a new synset (or one of the existing synsets) is attached to a word or, vice versa, detached from it: it is then a matter of a word acquiring or losing senses. This approach arguably can return more interpretable results, but, unfortunately, it unavoidably requires expensive human annotation. Thus, any manually curated lexical database is always inherently limited. In comparison, the distributional approach taken by *ShiftRy* is completely unsupervised, and the only text annotation it requires is the time of creation. Thus, it can in theory handle shifts occurring to any words of interest.

Russian news texts were studied in [3], who discovered that in state-sponsored Russian media, the United States are regularly mentioned more frequently the month directly following an economic downturn in Russia. This represents an interesting analysis of propagandist agenda-setting and the frames employed in these texts. However, unlike our work, [3] did not focus on meaning or usage changes, and limited themselves to only one media source.

The web service closest to *ShiftRy* is arguably *JeSeMe*² [7]. It allows to trace diachronic semantic trajectory of query words across several English and German corpora, mostly with time spans of 2 or 3 centuries. Instead, we deal with Russian language, and time spans of years, making it possible to study small cultural changes, including those influenced by political events.

The *ShiftRy* code is based on the *WebVectors* framework [11], which we substantially extended and adapted to the task of semantic change analysis. More specifically, we added new visualizations of semantic trajectories, and integrated semantic shift classifier and corpus examples sampling.

3. Collection of data

To collect training corpora, we crawled texts published from 2010 to 2019 from the Russian news web sites listed in **Table 1**.

Table 1: Training corpora sources and stances

Nr.	Title	URL	Stance
1	Fontanka.ru	https://www.fontanka.ru/	Opposition
2	Gazeta.ru	https://www.gazeta.ru/	Loyal
3	Interfax	https://www.interfax.ru/	Neutral
4	Izvestia	https://iz.ru/	Loyal
5	KP	https://www.kp.ru/	Loyal
6	Lenta.ru	https://lenta.ru/	Mixed
7	Novaya Gazeta	https://novayagazeta.ru/	Opposition
8	N + 1	https://nplus1.ru/	Scientific
9	RBC	https://www.rbc.ru/	Neutral
10	The Village	https://www.the-village.ru/	Opposition

The selection of the source web sites does not claim to cover the whole media space of Russian news (that would be impossible within the natural constraints of an academic paper). However, it is balanced with regards to the sources' political stance. Sources 2, 4 and 5 as a rule are loyal to the point of view of the Russian authorities, while sources 1, 7 and 10 regularly publish critical opinions (source 7 is generally considered to be one of the most influential opposition newspapers). Source 6 belonged more to the 'opposition' category until March 2014, when its top staff was forcefully changed to include managers loyal to the authorities. Since that time,

² <http://jeseme.org>

source 6 falls into the ‘loyal’ category. Sources 3 and 9 position themselves as a neutral news-wire service and a business analytic media correspondingly. Finally, source 8 publishes scientific news. The stances are listed in the right column of [Table 1](#).

The collected corpora were tokenized, lemmatized and PoS-tagged using UD-Pipe 2.3 [19]. After removing functional words, the full corpus for 10 years contains about 156 million tokens, with yearly sub-corpora sizes varying from 9 million (2014) to 20 million (2015) tokens.

4. Word Embedding Models

Continuous Bag-of-Words (CBOW) word embedding models [16] were trained on each of the yearly news corpora, with vector size 300, symmetric context window size 5, and no down-sampling, following the hyper-parameters from [4].

One important change compared to the previous work was that each yearly model was initialized with word vectors trained on the full Russian National Corpus (RNC). The RNC size is about 250 million word tokens, and it is very well balanced with respect to text genres. The reason for this pre-training was that our yearly news corpora are comparatively small, and thus the models trained solely on them might end up in a sub-optimal state. We hypothesized that initializing them with quality vectors trained on a representative Russian corpus would provide a stabler foundation for further learning year-specific embeddings.

The RNC model was trained for 10 epochs with the vocabulary size of 50,000 (most frequent) words. Since news texts do not constitute the majority of the RNC, this vocabulary certainly lacks many proper names and toponyms frequent in news pieces. This is why when updating the RNC model with the co-occurrence data from the yearly news corpora, we expanded the vocabulary, increasing its maximum size from 50,000 to 100,000 words. It means that if a new (not present in the RNC model) word occurred in a yearly corpus with the frequency 10 or more, it was added to the yearly model vocabulary. As a result, the sizes of the yearly models’ vocabularies vary from about 60 thousand to about 95 thousand words. This operation can be thought of as a sort of ‘fine-tuning’ the original RNC model, mixing its general world knowledge with the knowledge about recent political and social events from the yearly news corpora. To make the news texts influence the resulting models more, we performed this ‘fine-tuning’ in 20 epochs, instead of 10.

We found out that the RNC pre-training greatly benefited the diachronic models. [Table 2](#) shows the intrinsic performance scores of our embedding models trained on yearly news corpora (for comparison, we also report the performance of the original RNC model). For evaluation, we used the SimLex965 semantic similarity test set [10] and the Russian translation of the Google Analogies test set [15]. Without the pre-training, the results on Google analogies were pretty much the same as in the table. However, the accuracies on SimLex965 (that is, ranking word pairs by their semantic similarity) dropped almost twice: the average score was about 0.16. The reason is that SimLex965 (unlike Google Analogies) does not contain proper names and toponyms. At the same time, proper names are heavily over-represented in news texts. This why the models trained solely on them perform good enough in the Google

Analogies evaluation, but fail with SimLex965. RNC pre-training allowed us to avoid this and to make the resulting models more ‘aware’ of general world knowledge.

Table 2: Intrinsic evaluation scores of the yearly models in comparison to a single model trained on the RNC; the right column shows the number of words not shared by **all** models

Model	SimLex965	Analogies	Unique words
RNC	0.35	0.14	
2010	0.33	0.12	18,001
2011	0.33	0.12	17,953
2012	0.33	0.12	17,258
2013	0.33	0.12	17,240
2014	0.33	0.11	11,334
2015	0.31	0.11	19,849
2016	0.31	0.11	18,996
2017	0.32	0.12	18,666
2018	0.32	0.11	14,005
2019	0.31	0.11	13,928

4.1. Model alignment

Our yearly models are trained independently (except the shared RNC initialisation), and thus must be aligned to make it possible to directly measure cosine similarity between word vectors produced by them [12]. For that, we used orthogonal Procrustes analysis [5], which became a tool of choice for diachronic semantic change modeling with word embeddings since [6]. We transformed all the yearly models’ vector spaces to match the vector space of the 2017 model (its vocabulary was the largest of all, so we chose it as our ‘baseline’ model). All the methods described below in Section 5 deal with these aligned models.

The novel part of our alignment procedure was that we did not remove any words from the models’ vocabularies, even if they were not shared across the two models being aligned. Since singular value decomposition (SVD) in Procrustes analysis requires two matrices of the same shape, applying this method typically results in excluding all the words not present in the intersection of the models’ vocabularies. However, with our 10 time bins, intersecting vocabularies of all models would result in excluding too many words (arguably, most time-specific and thus most interesting). Hence, we conduct alignment in three stages:

1. For the baseline 2017 model M_{2017} and the model to align M_i , the matrices S_{2017} and S_i are generated, containing vectors for the words occurring in both models’ vocabularies. Their dimensionalities are identical.
2. The standard Procrustes alignment with SVD is applied to S_{2017} and S_i , resulting in the optimal orthogonal projection matrix O .

3. The original embedding matrix E_i from the model M_i is multiplied by O , thus projecting all M_i vectors into the vector space of M_{2017} . Note that the dimensionality of E_i can be (and most often is) different from the dimensionality of the original 2017 matrix S_{2017} , because of the differences in their vocabularies. However, all the words from M_i are projected into the M_{2017} vector space, including those missing from the M_{2017} vocabulary.

This modification allowed us to keep the unique vocabularies for yearly models. The right column of **Table 2** lists the number of unique words for each models: that is, the words from the model vocabulary missing from the intersection of all vocabularies. This gives the intuition of how many (potentially crucial) words would be lost if we stuck to the classical Procrustes alignment workflow.

5. Features of *ShiftRy* as a web service

Primary way of user interaction with *ShiftRy* is by entering word queries. The system will generate a visualization of the semantic drift trajectory for the query word across all the featured years (currently from 2010 to 2019) or across the user-specified time span. It will also apply a pre-trained classifier (described in **subsection 5.1**) to make educated guesses about whether the query word experienced sharp semantic shifts between any two consequent years. Users can use this classifier for a particular pair of years in the ‘Shift classifier’ tab.

To support deeper analysis of word usage change, *ShiftRy* also provides lists of the nearest neighbors of the query word for the query years. The ‘Synchronic visualizations’ tab allows to generate synchronic 2-dimensional PCA or t-SNE projections of any set of word vectors for any particular year.

Finally, the ‘Shift lists’ tab takes a pair of years as an input, and produces a ranked list of words, the usage of which changed most when comparing these two years. Here, the usage change is defined as a simple cosine distance between word vectors in each of the two (already aligned) models. In this way, a user can get insights about words most likely to have undergone a usage shift in a given period of time.

5.1. Shift classifier

The shift classifier predicts whether or not a query word has experienced a semantic shift within a given pair of years. It is a logistic regression trained on the ‘Micro’ dataset described in [4].

The ‘Micro’ dataset contains 280 Russian adjectives, together with year pairs, for instance: [‘экстремистский’ ‘extremist (adjective)’, 2007 → 2008]. The dataset covers the years from 2000 to 2014. Three independent informants annotated each word with one of the three labels: the meaning of the word did not change at all (0), somewhat changed (1), or experienced a significant change (2). The gold truth consists of their averaged scores.

Since the ternary granularity is not needed for *ShiftRy*, we binarized the target labels by mapping 0 → 0, 1 → 0 and 2 → 1. This was done to prioritize precision and

minimize false positives: we want the classifier to capture only significant shifts. The classifier uses four features, each of which is a similarity score between embeddings of the query word in two yearly models:

1. Cosine similarity between the query word vectors.
2. Similarity score produced by the Global Anchors algorithm [23]. It first creates a shared vocabulary V by intersecting the vocabularies of 2 models under analysis. Then, for each model, a vector of cosine similarities between the embedding of the query word and the embeddings of all words in V is produced. The resulting semantic change score is the cosine similarity between these 2 vectors (also known as second-order similarity).
3. Jaccard similarity [8] between the sets of 50 nearest neighbors of the query word in in each embedding model.
4. Kendall's τ [9] on the intersection of the ranked lists of 50 nearest neighbors; it tries to determine whether there are any significant changes in the neighbours ranking.

We evaluated the classifier on the ‘Micro’ training set using cross-validation with 5 folds. Table 3 shows the results comparing classifiers employing all four features and cosine similarity only. Note that these are scores on the ‘shifted’ class, which is the most important for us. As one can see, the 4-feature classifier consistently outperforms the single-feature one. Thus, we use the 4-feature classifier. Note that the ‘Micro’ dataset contains only adjectives, while *ShiftRy* works with all parts of speech: for this reason, the real reliability of the *ShiftRy* predictions is arguably somewhat lower. However, until another manually annotated test set of short-term Russian semantic shifts appears, we believe the results in Table 3 to be a good approximation of the *ShiftRy* performance.

Table 3: Semantic shift classifier scores on the ‘shifted’ class from the ‘Micro’ dataset

Features	F ¹	Precision	Recall
All features	27.8%	18.0%	80.0%
Cosine similarity only	20.1%	13.3%	50.2%

5.2. Word trajectory visualization

Projection of high-dimensional embedding data into a low-dimensional space using dimensionality reduction techniques is a well-established method of visualizing word embeddings. It can be used for diachronic change analysis as well.

ShiftRy employs T-distributed Stochastic Neighbor Embedding (t-SNE) [14], a nonlinear dimensionality reduction technique, to plot semantic trajectory of a specific word via several embedding models trained on different time bins. These visualizations make it easier to explore the usage drift of a query word and to trace how it changed its typical neighbours through time. Our visualization procedure is similar to the one used in [6] and [17].

caused by the media coverage of large protest rallies at the Bolotnaya (literally ‘Swampy’) Square in Moscow in 2012. These rallies resulted in mass arrests of its participants and organizers, with the ‘Bolotnaya case’ being in court trials throughout 2013 and 2014. This is why the 2014 point on the plot is close to the cluster of the names of the arrested and to the words ‘митинг’ ‘rally’ and ‘сфабриковать’ ‘to fabricate the case’. Note also a separate cluster with the names of other famous squares where protest events took place both in Russia and throughout the world (e.g. ‘манежный’ ‘Manezhnyy’ and ‘трафальгарский’ ‘Trafalgar’), together with the word ‘независимость’ ‘independence’. In 2018, the meaning of the target word is slowly shifting back. Notably, our shift classifier does not find semantic shifts for this word after 2013 (for any consequent pair of years). A possible reason for this is that while the shift was rapid and strong in 2012 (due to political events), the ‘return’ of the meaning back to the original state after 2014 was slow and gradual.

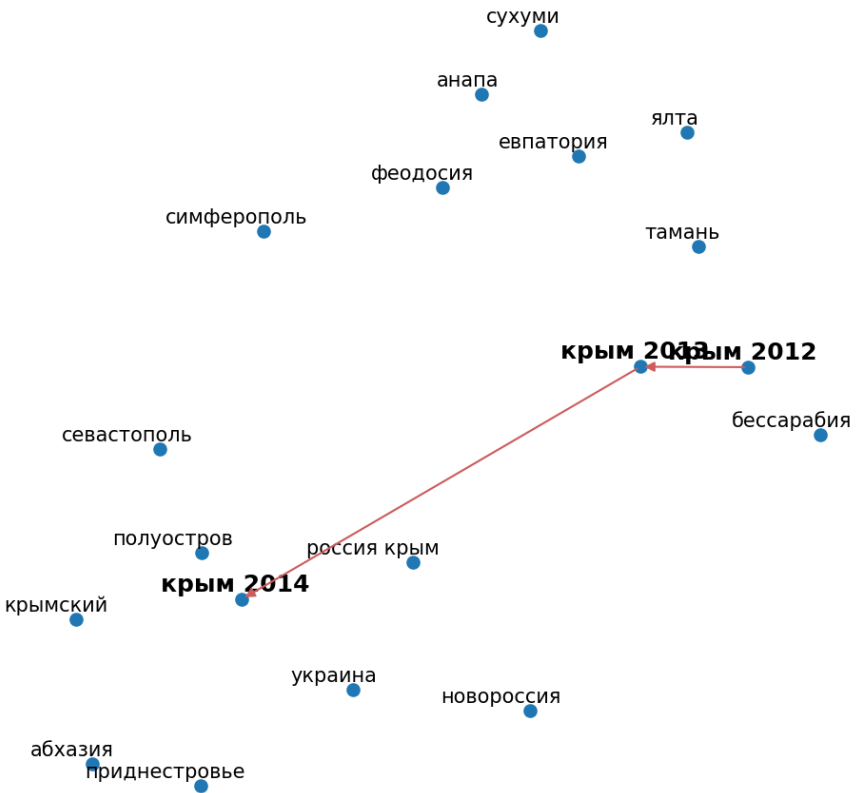


Figure 3: Alterations in the meaning of the word ‘Крым’ ‘the Crimea’: from ‘Ялта’ ‘Yalta’ and ‘Бессарабия’ ‘Bessarabia’ in 2012 to the ‘Россия Крым’ ‘Russia Crimea’ bigram in 2014

Another example is shown on **Figure 3**. Here, the semantic trajectory of the word ‘Крым’ ‘the Crimea’ moves from the cluster of Black Sea resort town names in 2012 and 2013 towards the words ‘Россия’ ‘Russia’, ‘Украина’ ‘Ukraine’ and ‘Новороссия’ ‘Novorossiia’ in 2014. This is, of course, caused by the Russian annexation of the Crimea, and the resulting territory conflict with the Ukraine.

Currently, the plots are static and non-interactive, but we plan to remedy that by making terms on them clickable in order to get lexical information and/or to explore the diachronic trajectories of neighbours. Another possible issue is that t-SNE is inherently stochastic, and the 2-dimensional projections can linearly change (rotate or scale) with the selected years, which in some cases can make it harder to explore the word’s history. This is, however, an engineering problem, which we plan to solve in further release versions.

5.3. Sampling examples from corpora

Given a query word w and a pair of years x and y , *ShiftRy* can produce lists of sentences containing w in both corresponding corpora, while making sure to sample the sentences so that x sentences are as far away semantically from y sentences as possible. This allows to manually check whether a usage shift is actually taking place or not.

To provide the user with the most representative examples of the word’s usage we first sample all the sentences that contain the query word from the text corpora of two chosen years. Then, we calculate the vectors of these sentences by averaging the embeddings of their words (recall that all the models are already aligned). Two sentences from different years with the highest cosine distance between their sentence vector are chosen as the first context pair. For each of them, we find 4 most similar sentences from the same corpora. This allows to get examples of the most specific contexts for each year, especially in the cases when the query word really shifted its meaning.

Below, we give an example with the word ‘пожар’ ‘fire’ and the years 2018 and 2019:

Old contexts (2018):

- В ночь перед пожаром мужчина со своей знакомой гостил у 70-летней погибшей. (The night before the fire, a man and his friend were visiting a 70-year-old friend.)
- Родители и четырехлетняя сестра погибли в пожаре. (Parents and a four-year-old sister died in a fire.)
- Сотрудница центра «Медицина катастроф» из Улан-Удэ Светлана Пежемская спасла из пожара 12 человек — родственников и соседей. (Svetlana Pezhemskaya, an employee of the Center for Medicine of Catastrophes from Ulan-Ude, rescued 12 people—relatives and neighbors—from the fire.)
- Незадолго до пожара мужчина грозился спалить весь дом и бросался в прохожих разными предметами. (Shortly before the fire, the man threatened to burn down the whole house and threw various objects at passers-by.)

- У кемеровчанина Игоря Вострикова в пожаре погибли трое детей, жена и сестра. (Kemerovo resident Igor Vostrikov lost three children in a fire, his wife and sister.)

New contexts (2019):

- В ведомстве полагают, что переход на новую модель организации позволит улучшить систему управления, оптимизировать излишние управленческие и обеспечивающие структуры подразделений, повысить эффективность реагирования на чрезвычайные ситуации и пожары. (The agency believes that the transition to a new model of organization will improve the management system, optimize redundant management and support structures of units, improve the efficiency of emergency response and fires.)
- Военная авиация привлечена для ликвидации крупного пожара близ города Новотроицк Оренбургской области, угрожающего населенным пунктам. (Military aviation has been involved in the elimination of a major fire near the town of Novotroitsk in the Orenburg region, which threatens populated areas.)
- Основными причинами пожаров стали нехватка финансирования, отсутствие техники для тушения, некомплектованность подразделений наземной и авиационной охраны в регионах и низкое качество противопожарных мероприятий, которые проводят арендаторы. (The main causes of the fires were lack of funding, lack of equipment for firefighting, understaffing of ground and aviation security units in the regions and poor quality of firefighting activities carried out by tenants.)
- Секретарь генсовета «Единой России» Андрей Турчак предлагает пересмотреть подходы к борьбе с лесными пожарами, заявил он РБК: «Существующая система борьбы с лесными пожарами меняется, исходя из необходимости оптимизировать расходы на лесоохрану». (Andrei Turchak, secretary of the United Russia general council, suggests revising approaches to forest firefighting, he said: “The existing system of forest firefighting has been changing, based on the need to optimize the cost of forest protection”.)
- Медведев также поручил подготовить предложения и представить в правительство, чтобы усилить группировку по борьбе с пожарами и помочь тем регионам, которые находятся в наиболее сложной ситуации: «Субъекты не справляются с тушением лесных пожаров, значит, необходимо усилить роль федерального центра в системе мониторинга и тушения пожаров», — пояснил Турчак. (Medvedev also instructed to prepare proposals and submit to the government to strengthen the group of firefighters and help those regions that are in the most difficult situation: “The subjects are not able to cope with extinguishing forest fires, so it is necessary to strengthen the role of the federal center in the system of monitoring and extinguishing fires,”—said Turchak.)

Examples from 2018 describe fires in buildings and generally within cities. At the same time, examples from 2019 are mostly related to the unprecedented large-scale Siberian forest fires which were widely covered in Russian media in 2019. Thus, they reflect a significant shift in this word usage.

6. Conclusion

In this paper, we described *ShiftRy*, which is a web service to explore the dynamics of word usage in Russian news texts from 2010 to 2019. It employs several existing methods of semantic change modeling with diachronic word embeddings to generate visualizations of temporal semantic trajectories for any user-entered word, and to provide guesses on whether this word experienced a sharp semantic shift within any pair of years. For deeper exploration of semantic shifts, *ShiftRy* also features the possibility to intellectually sample maximally different corpus examples for the query word from two different time periods. The aim of *ShiftRy* is to ease the task of studying word history on short-term time spans, and the influence of social and political events on word usage. Thus, it can be used for studies in digital humanities, as well as in historical linguistics.

Under the hood, the system runs 10 word embedding models trained on the corresponding yearly sub-corpora and aligned using a novel modification of the Procrustes analysis, which preserves time-specific vocabularies. The underlying diachronic word embeddings and the source code of the web service are available online (<https://shiftry.rusvectors.org>), making it easy to apply the same approach to other sources, for example, social media. For this, one would have only to crawl the relevant textual data and train the corresponding diachronic word embeddings.

We plan to maintain *ShiftRy* and update it with new data yearly, possibly also extending the covered time span backwards in the XX century. Another point for future work is to improve the semantic trajectory visualizations by making them more interactive and more robust with regards to t-SNE initialization randomness, and by trying other dimensionality reduction techniques such as UMAP [1]. Finally, we plan to apply contextualized embedding architectures both for semantic shift detection and for sampling of corpus examples with sentences containing one and the same word in different senses.

References

1. *Becht, E. et al.*: Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*. 37, 1, 38 (2019).
2. *Daniel, M., Dobrushina, N.*: Two centuries in twenty words (in Russian). NRU HSE (2016).
3. *Field, A. et al.*: Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. In: *Proceedings of the 2018 conference on empirical methods in natural language processing*. pp. 3570–3580 Association for Computational Linguistics, Brussels, Belgium (2018).

4. *Fomin, V. et al.*: Tracing cultural diachronic semantic shifts in Russian using word embeddings: Test sets and baselines. *Komp'yuternaya Lingvistika i Intellekturnye Tekhnologii: Dialog conference*. 203–218 (2019).
5. *Gower, J. C. et al.*: Procrustes problems. Oxford University Press on Demand (2004).
6. *Hamilton, W. et al.*: Diachronic word embeddings reveal statistical laws of semantic change. In: *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*. pp. 1489–1501 Association for Computational Linguistics, Berlin, Germany (2016).
7. *Hellrich, J. et al.*: JeSemE: Interleaving semantics and emotions in a web service for the exploration of language change phenomena. In: *Proceedings of the 27th international conference on computational linguistics: System demonstrations*. pp. 10–14 Association for Computational Linguistics, Santa Fe, New Mexico (2018).
8. *Jaccard, P.*: Distribution de la flore alpine: Dans le bassin des dranses et dans quelques régions voisines. *Rouge* (1901).
9. *Kendall, M. G.*: Rank correlation methods. Griffin (1948).
10. *Kutuzov, A., Kunilovskaya, M.*: Size vs. Structure in training corpora for word embedding models: Araneum Russicum Maximum and Russian National Corpus. In: *Aalst, W. M. van der et al. (eds.) Analysis of images, social networks and texts*. pp. 47–58 Springer International Publishing, Cham (2017).
11. *Kutuzov, A., Kuzmenko, E.*: WebVectors: A toolkit for building web interfaces for vector semantic models. In: *Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers*. pp. 155–161 Springer International Publishing, Cham (2017).
12. *Kutuzov, A. et al.*: Diachronic word embeddings and semantic shifts: A survey. In: *Proceedings of the 27th international conference on computational linguistics*. pp. 1384–1397 Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018).
13. *Loukachevitch, N., Lashevich, G.*: Comparing two thesaurus representations for Russian. In: *Proceedings of global wordnet conference gwc*. pp. 35–44 (2018).
14. *Maaten, L. van der, Hinton, G.*: Visualizing data using t-SNE. *Journal of Machine Learning Research*. 9, 2579–2605 (2008).
15. *Mikolov, T. et al.*: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. (2013).
16. *Mikolov, T. et al.*: Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. 26, 3111–3119 (2013).
17. *Rodina, J. et al.*: Measuring diachronic evolution of evaluative adjectives with word embeddings: The case for English, Norwegian, and Russian. In: *Proceedings of the 1st international workshop on computational approaches to historical language change*. pp. 202–209 Association for Computational Linguistics, Florence, Italy (2019).

18. *Shoemark, P. et al.*: Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). pp. 66–76 Association for Computational Linguistics, Hong Kong, China (2019).
19. *Straka, M., Straková, J.*: Tokenizing, pos tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Proceedings of the conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. pp. 88–99 Association for Computational Linguistics, Vancouver, Canada (2017).
20. *Tahmasebi, N. et al.*: Proceedings of the 1st international workshop on computational approaches to historical language change. Association for Computational Linguistics, Florence, Italy (2019).
21. *Tang, X.*: A state-of-the-art of semantic change computation. *Natural Language Engineering*. 24, 5, 649–676 (2018).
22. *University, P.*: About WordNet, (2010).
23. *Yin, Z. et al.*: The global anchor method for quantifying linguistic shifts and domain adaptation. In: Advances in neural information processing systems. pp. 9433–9444 (2018).

SENTENCE COMPRESSION FOR RUSSIAN: DATASET AND BASELINES

Kuvshinova T. (tatiana.kuvsh@yandex.ru)

Higher School of Economics, National Research University,
Moscow, Russia

Sentence compression is the task of removing redundant information from a sentence while preserving its original meaning. In this paper, we approach deletion-based sentence compression for the Russian language. We use the data from the plagiarism detection corpus (ParaPlag) to create a corpus for sentence compression in Russian of almost 3,000 pairs of sentences. We align source sentences and their compressions using the Needleman-Wunsch algorithm and perform human-evaluation of the corpus by readability and informativeness.

Then we use bidirectional LSTM to solve sentence-compression task for Russian, which is a typical baseline for the problem. We also experiment with RuBert and Bert-multilingual. For the latter, we use transfer-learning, firstly pretraining the model on English data, which improves performance. We conduct human evaluation by readability and informativeness and do error analysis for the models. We are able to achieve f-measure of 74.8%, readability of 3.88 and informativeness of 3.47 (out of 5) on test data. We also implement post-hoc syntax-based evaluator, which can detect some of the wrong compressions, increasing overall quality of the system.

We provide the data and baseline results for future studies.

Keywords: sentence compression, summarisation, bert, rubert, lstm, corpus

DOI: 10.28995/2075-7182-2020-19-517-528

СЖАТИЕ ПРЕДЛОЖЕНИЙ РУССКОГО ЯЗЫКА: ДАТАСЕТ И БЕЙЗЛАЙНЫ

Кувшинова Т. (tatiana.kuvsh@yandex.ru)

Национальный исследовательский университет
«Высшая школа экономики», Москва

Сжатие предложений — задача по удалению избыточной информации из предложения при сохранении его первоначального смысла. В этой статье мы обращаемся к сжатию предложений на основе удаления для русского языка. Мы используем данные из корпуса выявления плагиата (ParaPlag) для создания корпуса сжатых предложений русского языка, содержащего более чем 3000 пар предложений. Мы выравниваем

исходные предложения и их сжатия, используя алгоритм Нидлмана-Вунша, и проводим ручную оценку корпуса по читаемости и информативности.

Затем мы используем двунаправленную LSTM для решения задачи сжатия предложений русского языка, что является типичным способом решения этой задачи. Мы также экспериментируем с RuBert и многоязычным Bert. В последнем случае мы используем трансферное обучение, сначала обучая модель на английских данных, что улучшает качество работы системы. Мы проводим ручную оценку по читаемости и информативности и анализ ошибок для моделей. Мы достигли f-меры 74,8%, читаемости 3,88 и информативности 3,47 (из 5) на тестовых данных. Кроме того, мы разработали синтаксический оценщик, который может распознать некоторые из неправильных сжатий предложений, позволяя увеличить общее качество системы компрессии.

Мы предоставляем данные и результаты бейзлайнов для будущих исследований.

Ключевые слова: сжатие предложений, сокращения, суммаризация, корпус

1. Introduction

Sentence compression is a text-to-text rewriting task of shortening a sentence by omitting redundant information while preserving the main points. Usually, a “perfect” compression is a difficult paraphrase task, involving removing, reordering and inserting operations. However, the problem is often narrowed down to deletion-based sentence compression.

Deletion-based sentence compression is performed by removing some words from the original sentence. The remaining words form compression in the exact same order as they appeared in the original sentence. This problem can be solved as a sequence labeling binary classification task—we classify each token in the original sentence with “stay” or “delete” label and then remove all the tokens with the “delete” label.

Sentence compression for English has attracted many researchers. An important milestone in the field was the release of a large parallel compression corpus by Google [4] and as well as applying deep learning techniques, especially LSTMs to the problem [3]. The most recent results in the field are achieved by complex models, combining LSTM with some additional techniques, for example, language model evaluator [11], [18].

However, to the best of our knowledge sentence compression has never been solved for Russian, which is easily explained by the lack of parallel data and complex structure of the language. In this paper, we approach deletion-based sentence compression for Russian. We present a parallel sentence-compression corpus for Russian with almost 3 thousand sentences based on The ParaPlag [15] data. We also conduct experiments with several data-driven models for sentence compression, which are bi-directional LSTM, RuBert-compression, and MultiBert-compression. We build syntax-tree based evaluator over compression systems to improve compression quality and perform an automatic and human evaluation of the results.

We hope that our work will serve as a pivot point for future researches.

2. Data

Data-driven solutions to natural language processing problems are usually data-hungry. Deletion-based sentence compression requires a large parallel corpus with sentence-compression pairs. Such data has no evident natural source (as, for example, parallel translation data) and no specific sentence compression corpora have yet been released for Russian.

In this paper, we use data from the Russian dataset for paraphrased plagiarism detection [15] to construct a deletion-based sentence-compression corpus for Russian. The corpus contains about 20 thousand manually paraphrased source-plagiarism pairs. Sometimes, when a person plagiarises a sentence, he or she summarizes or expands it. Fortunately, such examples are annotated in the corpus. We use them as a source for our dataset.

There are 7,298 such pairs of sentences in the dataset. Some of them are also paraphrased, in such cases we resolve the paraphrase before further processing. For example, for the source pair of sentences:

Sentence: Книги Толкина послужили основой для создания множества настольных, компьютерных и видеоигр как для PC, так Mobile.

Compression: Книги Толкина стали основой для создания настольных, компьютерных и видеоигр.

As a result of paraphrase resolution the compression will become:

Книги Толкина послужили основой для создания настольных, компьютерных и видеоигр.

We exclude those sentences where a paraphrase cannot be resolved automatically (for example when it is adjacent to deletion and its boundaries cannot be clearly established).

Then we tokenize a source sentence and its compression and align them using the Needleman-Wunsch algorithm for sequence alignment [13]: a dynamic programming approach firstly developed for DNA sequence alignment. The result of the alignment for the example sentence given above will be:

Finally, we exclude all the pairs with a compression rate higher than 0.7 (traditionally, small compression rate corresponds to shorter compressions, i.e. a compression rate is a proportion of words to stay). We make that decision because the sentences with high compression rate form poor training material: they are only partly compressed, some words that *could be* removed are not removed from them. Thus, a compression system would be confused by the training data inconsistency.

The remaining data contains 2,955 pairs of sentences. We split it into the train and test set as follows:

Table 1: The parallel sentence-compression corpus statistics and train-test split

set	no. sentences	no. tokens	compression rate
train	2,659	66,238	0.55
test	296	7,450	0.56

Common metrics for compression quality evaluation are readability and informativeness [12]. Readability addresses grammatical correctness and naturality of a compressed sentence. Informativeness shows a degree to which original meaning was preserved in the compression. We asked three native Russian speakers to annotate the first 150 sentence-compression pairs from the test set by readability and informativeness using a 5-grade scale. The first annotator has annotated all 150 pairs of sentences, the second—the first 100 pairs and the third—the pairs from 101 to 150, thus giving two annotations for every pair of sentences. As a result, we got a mean readability of 4.68 and mean informativeness of 4.03.

We make the sentence-compression corpus available online for future researches.¹

3. Experiments

3.1. LSTM

Bidirectional long-short term memory networks form strong baselines and are used as baselines in the majority of recent papers about sentence compression in English [3, 7, 11, 18]. That is why we decided to use biLSTM as our baseline as well.

We build a 3-layered biLSTM network with input and hidden layers of 300 dimensions and a *softmax* output layer. We use binary cross-entropy loss function and Adam optimizer. We initialise our model with *tayga_none_fasttextcbow_300_10_2019* vectors by RusVectōrēs project [10]. We also allow the model to finetune embeddings during training. We lemmatize the corpus before training making it compatible with embeddings using the Russian UdPipe Syntagrus model [16].

We implement our model using PyTorch and train it on sentence-compression data for Russian with the following parameters: learning rate of $1e - 3$, weight decay of $1e - 3$, dropout of 0.35 and mini-batch size of 20 sentences. We do not finetune any parameters and train the model until test loss starts increasing, which is 20 epochs on our data.

We then evaluate the model and present the results in the **Evaluation** section.

¹ download at <https://goo-gl.su/EqQvj8>

3.2. RuBert

BERT model has outperformed state of the art results in many natural language processing tasks [1], so it was only natural to approach sentence compression task with it.

We add a linear classification layer on top of the BERT hidden-states output and use Cross-Entropy for the loss function. We use Adam optimizer.

We implement the model using PyTorch transformers and use the pretrained RuBert model by DeepPavlov [8] to initialize the model. We convert DeepPavlov RuBert checkpoint to PyTorch-transformers Bert checkpoint to be able to use the same architecture and training procedure for RuBert and original Multilingual Bert model (described in the next subsection). Then we train it on Russian compression data with a learning rate of $3e - 5$ and minibatch of 8 sentences until the loss starts increasing on the test set, which takes 6 epochs for our data. We do not finetune any parameters.

The evaluation results are presented in the **Evaluation** section.

3.3. M-Bert

Multilingual Bert was proven to be good at cross-lingual model transfer [14], i.e. finetuning M-Bert to solve a task on a one language data allows solving the same task for other languages with little drop in quality. Authors hypothesize that the reason behind that behavior may be that similar words in different languages are surrounded by the same universal symbols (such as punctuations, numbers, etc.) and therefore learn similar vector representation. Therefore it makes sense to finetune M-Bert on English sentence-compression data before solving the task for Russian.

We use the same model architecture as for RuBert model. Firstly, we finetune the model on 8,000 English sentence-compression pairs from the first release of the Google sentence-compression dataset [4]. We train the model until the loss of 1,000 English test set starts increasing, which takes 3 epochs. Then we further train it on Russian sentence-compression pairs, again, until an increase in test loss, which takes 3 epochs.

For comparison we also train M-Bert model on Russian data only with the same parameters (i.e. skipping the first part of training procedure). This enables us to distinguish between the impact of fine-tuning and initial model quality.

Finally, we evaluate the models and present the results in the **Evaluation** section.

4. Post-hoc syntax evaluator

4.1. Foundation

Compression models performances can be inconsistent: they can compress correctly one sentence and fail with another. This sort of behaviour is highly undesirable for a service and neural network algorithms are hardly controlled and debugged. That is why we considered building post-hoc evaluator to separate correctly compressed sentences from damaged ones.

Inspired by tree cutting approaches to sentence compression [5], [6], [9], we decided to build a syntax-tree based evaluator. The tree cutting approach is based on the idea that parts of a sentence close to dependency graph root are preserved during compression, while leaves are dropped out. At the same time it is ungrammatical to remove a head preserving its child, i.e. the resulting graph ought to be a subgraph of the source graph with the same root.

Consider the following example: *Сокращая предложения, будьте очень внимательны.* and its dependency graph below, as well as dependency graphs of its deletion-based compressions.



Figure 1: Source sentence dependency graph



Figure 2: Correct compression graph

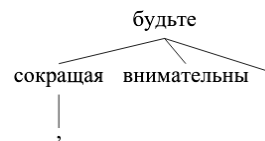


Figure 3: Correct compression graph

We see that potentially correct compression *Сокращая, будьте внимательны.* is a subgraph of the source sentence graph. On the opposite, potentially wrong compression has a “cut off” node *очень*, which has lost its head. Such transformation is usually ungrammatical and should not be allowed in the compression system.

4.2. Implementation

We use Universal Dependencies syntax parser [17] and Russian Udpipes Syntagrus model [2] for syntax parsing. Firstly, we build syntax trees for source and compressed sentences. Then we count tree distance—number of nodes that have changed their parents (which presumably happened because their original head was removed by compression system).

Mostly, any deviation from a source sentence subgraph indicates an ungrammatical compression. However, due to possible syntactic parser mistakes and several corner cases such as conjunction treatment, we would like to allow at least some fluency to an algorithm. To determine a threshold after which we consider a compression wrong, we calculate tree-distances between source sentences and gold compressions from the train dataset. We normalize distances over compression length to get a proportion of tokens that changed their heads and get the following normalized tree distance distribution:

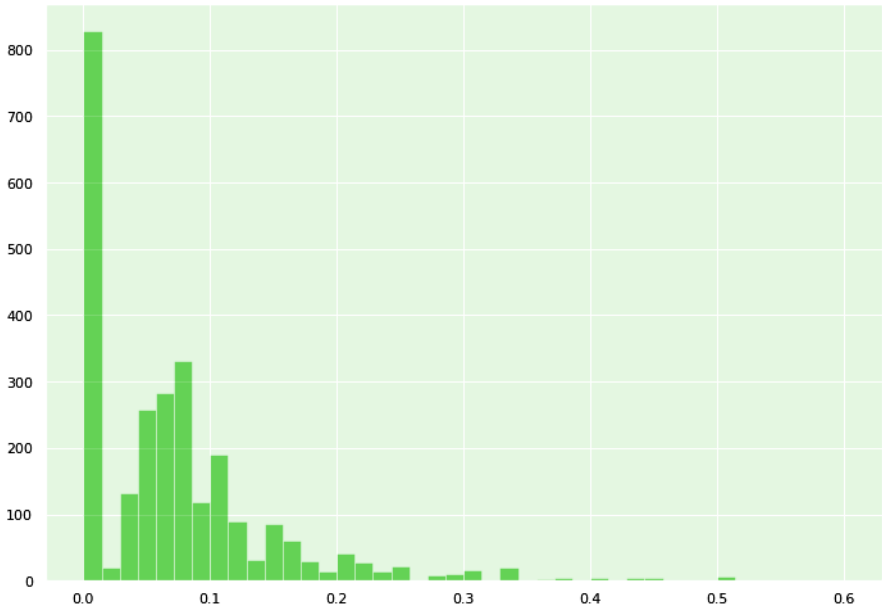


Figure 4: Normalized tree distance distribution over the train set

Based on the distribution and examples analysis we chose a threshold of 0.1—all compressions with higher proportion of misplaced nodes are considered wrong.

Our syntax evaluator doesn't directly improve the models performances, it only shows probably ungrammatical compressions, cases where the model have failed. Leaving those sentences uncompressed allows to trade off recall for precision.

5. Evaluation

While automatic evaluation using test data is a good way to access model performance, it is not sufficient for sentence compression task. For example, removing a subject from a sentence can give us accuracy of $> 90\%$, and yet the sentence can become completely ungrammatical and uninformative. It is possible that in an experiment a model with the maximum F-score and a model producing the most readable compressions are different ones [11]. That is why we perform both automatic and human evaluation of the results.

5.1. Automatic evaluation

We evaluated the models' performance on our 296 sentences test-set. For every model we evaluated its performance as a standalone system and combined with syntax evaluator. When using syntax evaluator, we excluded those compressions that were marked as probably wrong, i.e. evaluated only sentences approved by syntax evaluator. You can see evaluation results in the table below:

Table 2: F1-score on test data for LSTM, RuBert and M-Bert models; “f1-delete” is F1 measure for delete labels, “f1-stay” is F1 measure for stay labels and “f1” is weighted-averaged F1 measure between labels; SE stands for Syntax evaluator

model	f1, %	f1-delete, %	f1-stay, %	training data
LSTM	74.8	71.3	77.5	2,659 Russian pairs
LSTM + SE	77.1	70.1	81.3	2,659 Russian pairs
RuBert	67.9	63.0	71.6	2,659 Russian pairs
RuBert + SE	70.2	68.5	71.7	2,659 Russian pairs
M-Bert (ru)	67.1	63.8	72.1	2,659 Russian pairs
M-Bert (ru) + SE	68.2	64.9	73.5	2,659 Russian pairs
M-Bert	69.3	64.4	73.0	8,000 English + 2,659 Russian pairs
M-Bert + SE	70.8	65.5	74.5	8,000 English + 2,659 Russian pairs

As we can see, baseline LSTM surprisingly performs better than Bert-base models. M-Bert trained with both English and Russian data exceeds both Ru-Bert and M-Bert trained with Russian data only, it shows the promise of using transfer-models for sentence compression task.

Systems using syntax evaluator show higher performances than row models, however they leave some sentences uncompressed.

5.2. Human evaluation

We asked three native Russian speakers to evaluate systems’ compressions as well as gold compressions of the same sentences by readability and informativeness using a 5-grade scale. The first annotator has evaluated compressions for the first 150 sentences of the test-set (600 sentences), the second—the first 100 sentences of the test-set (400 sentences) and the third—compressions of sentences from 101 to 150 of test-set (200 sentences). Compressions were presented to annotators in random order so that they don’t know which system (or gold compression) they are evaluating. Agreement rate (Pearson correlation) between annotators is presented in the table below:

Table 3: Agreement rate (Pearson correlation) between annotators; the p-value is less than $1e - 9$ for every test

annotator pair	readability	informativeness	sentences in common
1 & 2	0.75	0.66	1–100
1 & 3	0.86	0.78	101–150

We consider the agreement rate satisfactory.

We also evaluate test data after processing it with syntax evaluator. We exclude sentences and compressions marked by syntax evaluator as inappropriate before calculating metrics. Note, that this inevitably increases compression rate, as some sentences remain uncompressed. We evaluate gold compressions by the same procedure.

Table 4: Mean readability, informativeness and compression rate for LSTM, RuBert, M-Bert models and gold data for the first 150 sentences of the test set; SE stands for syntax evaluator; higher compression rate corresponds to longer compressions and compression rate includes uncompressed sentences for models with SE

model	readability	informativeness	sentences excluded by SE	compression rate
Gold	4.68	4.03	0	0.56
Gold + SE	4.73	4.13	35	0.73
LSTM	3.84	3.29	0	0.54
LSTM + SE	4.10	3.61	49	0.78
RuBert	3.46	3.26	0	0.58
RuBert + SE	3.59	3.42	81	0.85
M-Bert	3.88	3.47	0	0.62
M-Bert + SE	4.07	3.71	55	0.84

We provide evaluation results for gold compressions from the test set for better comparison. We encourage any future researches of sentence compression in Russian to evaluate gold compressions as well. The reason is different annotators may have different ideas about subjective metrics as readability and informativeness, so human-evaluation results are hardly comparable between studies. However, evaluating the same set of gold compressions will allow leveling systems' results *in comparison* with gold compressions.

We see that M-Bert is better than other models in both readability and informativeness, even though it had had much lower f-measure than the LSTM model. We see that apparently training on a large English corpus allows the model to better generalize in sentence-compression task.

Applying syntax evaluator after compression system (excluding syntactically broken compressions) increases compressions readability and informativeness at the expense of increasing compression rate. The difference between a raw model and a model with post hoc evaluator is higher for initially worse models (where more compressions had to be excluded). However, RuBert model shows relatively low readability and informativeness even after syntax evaluator excluded more than a half of its compressions from analysis. After applying syntax evaluator the systems remove only about 15–20% of words on average, but this may be desirable results in some cases.

5.3. Examples and error analysis

While we consider evaluation results more than satisfactory for the first experience in sentence compression for Russian, there are still many cases when the models fail to produce grammatical and informative compression. We will discuss them and possible approaches to improving systems' performance in this subsection. We mark those compressions which were considered inappropriate by syntax evaluator with asterisk (*).

One of the frequent mistakes made by the models is the meaning travesty. Words deletion can sometimes radically change meaning as in the following examples:

- (1) **Source sentence:** *К тому времени Фрида уже не могла встать с постели.*
LSTM: *Фрида не могла встать.*
RuBert: *Фрида могла встать с постели.*
M-Bert: *Фрида уже не могла встать с постели.*
- (2) **Source sentence:** *В октябре 2011 года робот CubeStormer II, специально собранный из 4 наборов конструктора Lego Mindstorms, побил рекорд человека и собрал кубик за 5,53 секунды.*
LSTM: *В октябре 2011 года робот CubeStormer II, специально собранный из 4 наборов конструктора Lego Mindstorms, побил рекорд человека.*
RuBert: *В октябре 2011 года робот CubeStormer II, специально собранный из 4 наборов Lego Mindstorms побил.*
*** M-Bert:** *В октябре 2011 года робот CubeStormer II, специально собранный из 4 наборов конструктора Lego Mindstorms, побил человека.*

In the first example, RuBert system removes negation giving the sentence an opposite meaning. The syntax evaluator does not penalise those compressions, because in Universal Dependencies grammar negation is dependent on its verb. This case can be, however, addressed by merging negation to the following verb at preprocessing stage (which, of course, will require embeddings trained in such way).

In the second example M-Bert system removes word *рекорд*, the sentence stays completely grammatical, but the meaning changes from *The robot broke human record* to *The robot beat a man...* The head in noun phrase *рекорд человека* is *рекорд*. The complement *человека* cannot stay when its head was removed, but the system is apparently confusing the noun phrase for a typical *Adjective+Noun* phrase and is removing the first word. Providing the models with some syntactic information could improve performance. At the same time, in this case syntax evaluator is able to detect wrong transformation and marks the compression as inappropriate.

Another typical situation when the systems produce wrong compressions is the presence of a point inside a sentence (in an abbreviation or as a part of a number). It would appear that the system mistake this point to end-of-sentence point and get rather confused seeing tokens passed after it. The resulting compressions are barely readable:

- (3) **Source sentence:** *В финале соревнования по игре Painkiller в 2005 г. его выигрыш составил, ни много ни мало, 150 000 долларов.*
*** LSTM:** *В финале соревнования по игре.*
*** RuBert:** *В финале соревнования по игре Painkiller. 150 000 долларов.*
*** M-Bert:** *В финале соревнования по игре Painkiller в 2005 г. его выигрыш составил ни.*

To avoid such mistakes it is possible to remove inline points or change them to different symbols while preprocessing. Syntax evaluator detects severe changes in syntax structure in all systems' compressions.

6. Results

In this paper, we have addressed the problem of deletion-based sentence compression in Russian. We have prepared and preprocessed a parallel corpus of 2,955 sentence-compression pairs based on the data of ParaPlag corpus [15]. We asked three native Russian speakers to evaluate the corpus quality and got mean scores of 4.68 for readability and 4.03 of informativeness (out of 5).

We then build three sentence compression systems for Russian: bidirectional LSTM, RuBert based model and multilingual Bert based model. We train LSTM and RuBert with our corpus and multilingual Bert on both English sentence-compression data from [4] and our corpus.

We also implement post-hoc syntax evaluator based on source and compression syntax trees comparison. The evaluator does not improve compressions, but is able to detect some of wrong compressions, which could prove useful in applying systems to real data.

We conduct both automatic and human evaluation of the results. LSTM model achieves the highest f-measure on the test set, which is 74.8%. The multilingual Bert based model is the best judging by human evaluation: it achieves readability of 3.88 and informativeness of 3.47 (out of 5).

Applying syntax evaluator (evaluating only compressions approved by syntax evaluator) allows to achieve f-measure of 81.3% (for LSTM model), readability of 4.10 and informativeness of 3.71 (for multilingual Bert). However, the system with syntax evaluator simply doesn't compress lots of sentences, increasing average compression rate as the result.

We give several examples of models' performance and do error analysis.

Our work is the first to the best of our knowledge dedicated to sentence compression in Russian. We consider the achieved results as a good start and certainly hope that future researches will consider the topic and will be able to surpass our results.

References

1. *Devlin, J. et al.*: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).
2. *Droganova, K. et al.*: Data conversion and consistency of monolingual corpora: Russian ud treebanks. In: Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018), december 13–14, 2018, oslo university, norway. pp. 52–65 Linköping University Electronic Press (2018).
3. *Filippova, K. et al.*: Sentence compression by deletion with lstms. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 360–368 (2015).

4. *Filippova, K., Altun, Y.*: Overcoming the lack of parallel data in sentence compression. In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 1481–1491 (2013).
5. *Filippova, K., Strube, M.*: Dependency tree based sentence compression. In: Proceedings of the fifth international natural language generation conference. pp. 25–32 Association for Computational Linguistics (2008).
6. *Filippova, K., Strube, M.*: Sentence fusion via dependency graph compression. In: Proceedings of the conference on empirical methods in natural language processing. pp. 177–185 Association for Computational Linguistics (2008).
7. *Klerke, S. et al.*: Improving sentence compression by learning to predict gaze. arXiv preprint arXiv:1604.03357. (2016).
8. *Kuratov, Y., Arkhipov, M.*: Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213. (2019).
9. *Kurisinkel, L. J. et al.*: Domain adaptive neural sentence compression by tree cutting. In: European conference on information retrieval. pp. 475–488 Springer (2019).
10. *Kutuzov, A., Kuzmenko, E.*: WebVectors: A toolkit for building web interfaces for vector semantic models. In: Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers. pp. 155–161 Springer International Publishing, Cham (2017).
11. *Kuvshinova, T., Khritankov, A.*: Improving a language model evaluator for sentence compression without reinforcement learning. In: Proceedings of the tenth international symposium on information and communication technology. pp. 92–97 (2019).
12. *Mani, I.*: Summarization evaluation: An overview. (2001).
13. *Needleman, S. B., Wunsch, C. D.*: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*. 48, 3, 443–453 (1970).
14. *Pires, T. et al.*: How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502. (2019).
15. *Sochenkov, I. et al.*: The paraplac: Russian dataset for paraphrased plagiarism detection. In: Computational linguistics and intellectual technologies: Papers from the annual international conference “dialogue. pp. 284–297 (2017).
16. *Straka, M., Straková, J.*: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipeline. In: Proceedings of the conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. pp. 88–99 (2017).
17. *Straka, M., Straková, J.*: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipeline. In: Proceedings of the conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. pp. 88–99 Association for Computational Linguistics, Vancouver, Canada (2017).
18. *Zhao, Y. et al.*: A language model based evaluator for sentence compression. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers). pp. 170–175 (2018).

«UNDERSTATEMENT» И САРКАЗМ: ЛЕКСИКАЛИЗАЦИЯ РИТОРИЧЕСКОГО ПРИЕМА¹

Левонтина И. Б. (irina.levontina@mail.ru)

Институт русского языка им. В. В. Виноградова РАН

DOI: 10.28995/2075-7182-2020-19-529-540

“UNDERSTATEMENT” AND SARCASM: LEXICALIZATION OF A RHETORICAL DEVICE

Levontina I. B. (irina.levontina@mail.ru)

Russian Language (Vinogradov) Institute RAS

Understatement is a rhetorical device, based on making a statement weaker than it could be made in a given situation (i. e. underrating, less confident, presented as unimportant). In modern Russian, especially in colloquial speech, an extremely popular rhetorical figure is a combination of understatement and sarcasm; recently, several new ways of forming this figure have appeared: *na minutochku, esli chto, nichego chto..?* [Eto na minutochku moyа professiya; Eto, esli chto, moyа professiya; A nichego, chto eto moyа professiya?] ([literally *This is my profession, for a minute; This is my profession, just in case; Doesn't it mean anything that this is my profession?*]). For some language units, the corresponding meaning is partially or completely lexicalized. So, *na minutochku* and *na sekundochku* do not initially possess a “degrading” sense (if it is not really about time, meaning that you need a tiny bit of time for something); they are always used sarcastically. That said, as opposed to *na minutochku* and *na sekundochku*, other word forms (*na minutu, na minutku, na sekundu, na mig, na mgnovenie*) are not used this way. Thus, here we have a completely lexicalized figure of speech. In general, sarcasm is extremely difficult to formalize. Therefore, detection of linguistic manifestations of sarcasm appears to be extremely valuable.

Keywords: understatement, rhetorical device, sarcasm, lexicalization, discursive words, Russian

1. Вводные замечания. Understatement

Understatement — риторический прием, состоящий в том, что делается утверждение более слабое, чем могло бы быть сделано в данной ситуации (приуменьшенное, менее уверенное, представленное как неважное). Можно сказать,

¹ Работа выполнена при поддержке РФФИ, грант 19-012-00291.

что в этом случае происходит отступление от постулата количества Г. П. Грайса. Приведем фрагмент о филологе Романе Тименчике из книги А. К. Жолковского «Звезды и немного нервно»: «Тименчик подобен тому английскому джентльмену, на примере которого иллюстрируется понятие *understatement*. Когда среди его гостей возникает спор о том, что такое Занзибар, и кто-то говорит, что это такая птица, кто-то — что это рыба, и т. д., — он долго отмалчивается, пока наконец не позволяет себе осторожно предположить, что, кажется, Занзибар где-то в Африке, — и это при том, что в свое время он двадцать лет прослужил губернатором Занзибара! Если Тименчик говорит, что не уверен, что цитата из «Записных книжек», значит, у него есть веские основания, типа того, что он только что написал работу о подтекстах этих «Книжек» или прочитал о них аспирантский курс, а возможно, и то, и другое».

Действительно, *understatement* воспринимается как типично английский *modus dicendi* и как сердцевина специфически английского юмора. В знаменитой книге британского юмориста венгерского происхождения Джорджа Микеша (George Mikes) “How to be an Alien” (1946) говорится: *The English have no soul; they have the understatement instead* — «У англичан души нет. Вместо души у них подтекст²». И далее: «Если европейский юноша хочет объясниться в любви, он становится перед девушкой на колени, говорит ей, что она самое прелестное, самое обворожительное существо на свете, что в ней есть что-то такое, что отличает еще лишь несколько сот тысяч других женщин, и что жизни без нее он себе совершенно не представляет. Чтобы доказать, что слов на ветер он не бросает, юноша иногда тут же, на месте стреляется. Таково самое урядное, повторяющееся изо дня в день объяснение в любви в темпераментных европейских странах. В Англии же молодой человек хлопает свою возлюбленную пониже спины и вопологоса роняет: «Я **вообще-то** против тебя ничего не имею». Если же он взболмел от страсти, то может добавить: «Ты мне **как бы** нравишься»³ (Джордж Микеш «Как быть иностранцем. Шекспир и я: Повести»; пер. А. Ливергант, Изд. Б. С. Г.-Пресс, 2007 г.)

Разумеется, то, что этот способ изъясняться считается типично английским, не означает, что по-русски соответствующий риторический прием невозможен. И по-русски можно сказать *Боюсь, что я не в Москве* или назвать *не очень удачным* провальное выступление, а жгучего брюнета — *не совсем блондином*. Этот прием представлен в приведенном фрагменте о Тименчике. Кроме того, стоит обратить внимание на то, что в цитате из Микеша *rather fancy* изящно переведено

² Любопытно, что переводчик Александр Ливергант переводит здесь *understatement* как *подтекст*, а не более привычно — например, как *сдержанность*.

³ The English have no soul; they have the understatement instead. If a continental youth wants to declare his love to a girl, he kneels down, tells her that she is the sweetest, the most charming and ravishing person in the world, that she has something in her, something peculiar and individual which only a few hundred thousand other women have and that he would be unable to live one more minute without her. Often, to give a little more emphasis to the statement, he shoots himself on the spot. This is a normal, tweek-day declaration of love in the more temperamental continental countries. In England the boy pats his adored one on the back and says softly: 'I don't object to you, you know.' If he is quite mad with passion, he may add: 'I **rather** fancy you, in fact.' (George Mikes “How to be an Alien”).

как *как бы нравишься*. Действительно, слово *как бы* в поздне-советское время, до того, как произошло его семантическое опустошение, как раз и использовалось в интеллигентской речи для сообщения дискурсу некоторой неуверенности и ненавязчивости, маркирующих его противопоставленность официозу. Снижение категоричности — один из стандартных приемов придания речи вежливости. Мы часто мягко говорим: *Мне кажется, вы не совсем правы*, когда на самом деле совершенно уверены, что собеседник в корне неправ. См. [Ларина 2009].

2. Understatement и сарказм

Кроме того, для русского языка весьма характерно сочетание приумножения, смягчения с сарказмом, как у Грибоедова:

- (1) *Фамусов Скажи, тебе понравилась она?
Обрыскал свет; не хочешь ли жениться?
Чацкий А вам на что?
Фамусов Меня не худо бы спроситься,
Ведь я ей **несколько сродни**;
По крайней мере, искони
Отцом недаром называли*
(А. С. Грибоедов. Горь от ума. 2, 2)

Понятно, что отец — это ближайший родственник, и занижающее выражение *несколько⁴ сродни* Фамусов использует здесь саркастически, чтобы поддеть Чацкого.

Оговоримся, что под сарказмом в настоящей работе имеется в виду использование языковых единиц в смысле, противоположном обычному, с риторической целью. Этот механизм лежит в основе не только собственно сарказма, но также и его более мягкого варианта — иронии.

Этот прием широко применяется в конфликтном диалоге; о понятии конфликтного диалога см. [Зализняк Анна 2004].

Приведем примеры:

- (2) *Я тебя, по-моему, не спрашивал!*
(3) *Я, кажется, велела помолчать!*
(4) *Я вроде не просил совета!*

Во всех этих случаях говорящий на самом деле просто констатирует факт, что он не спрашивал, велел, не просил и т. п., и добавление маркера мнения или неуверенности делает высказывание особенно раздраженным. Естественно, что в таких контекстах обычно не фигурируют маркеры неполной уверенности типа *наверное, скорее всего или вероятно*:

⁴ Так могут использоваться самые разные выражения со значением малой степени: *немного, чуть-чуть, слегка* — в том числе всевозможные сленговые выражения: *слегонца, полразика, на полшишечки* и т. п.

(5) *Я, вероятно, не спрашивал совета!*

Аналогично используются понижающие тимиологические⁵ показатели:

(6) — *Что-то случилось? — Да так, меня уволили;*

(7) — *Вообще-то мы ищем убийцу!*

Да так указывает на низкую важность, при этом, скорее всего, увольнение — важное для человека событие, и поэтому ответ звучит полемически. *Вообще-то* указывает на то, что какое-то обстоятельство или соображение не является решающим, однако поиски убийцы — это нечто очень важное, и *вообще-то* употреблено здесь в риторических целях.

Сравним два примера:

(8) — *Я не вовремя? — Ну, я вообще-то сплю, но если что-то срочное...*

(9) — *Ты чего звонишь в час ночи? Я вообще-то сплю!*

В первом примере *вообще-то* употреблено в прямом режиме: то, что я сплю, — это обстоятельство, которым можно пренебречь в экстренном случае. Во втором примере нейтральным было бы, например, *Я же сплю*, то есть подчеркивание неуместности столь позднего звонка, но саркастически используется *вообще-то*, понижающее важность аргумента. Это усиливает иллюзивный эффект: речевой акт упрека, который реализуется в высказывании, становится более выразительным, даже агрессивным.

3. Лексикализация

Процесс лексикализации языковых единиц в тех или иных сочетаниях или с теми или иными значениями проходит разные этапы. Нередко единица конвенциализуется в каком-то употреблении, но оно не полностью «отпочковалось» от основного употребления⁶. Возникает ситуация недоразвившейся полисемии; ср. [Урысон 1998]. Похожие единицы могут находиться в разных точках этой шкалы. Так, выражения *казалось бы* и *кажется* близки по смыслу, и оба могут использоваться саркастически. Однако фразы *Я, казалось бы, просил помолчать* и *Я, кажется, просил помолчать* различаются: фраза с *казалось бы* нестандартна и просто реализует риторический прием, а во фразе с *кажется* можно усмотреть зачатки уже наметившейся полисемии, и в подробном словаре стоило бы отметить соответствующий тип употребления этой единицы.

⁵ О тимиологических единицах в языке см., например, работы [Пеньковский 1995], [Трипольская 1992].

⁶ Ср. разную подачу разных лексем (значений) вокабулы и так называемых «употреблений» одной лексемы в «Активном словаре русского языка» [ADR 2014].

3.1. Вообще-то

У *вообще-то*⁷, как и у некоторых других подобных языковых выражений, саркастическое употребление конвенционализуется, образуя особый тип употребления:

- (10) *А я ещё же устала чуть-чуть / я такая думаю / кошмар / ну / дошла до дома / ну / Дима вышел уже / ждёт меня в подъезде / «что так долго?» Я говорю / «**Вообще-то** лифт не работает!»* [Телефонный разговор (2006)]
- (11) *Я / говорит / устал очень. От работы / от Москвы / от проблем... От тебя... Ты представляешь?! Я так рот-то и открыла! Устал он / видите ли! А я не устала? Я не устала? У меня работа стоячая! Он думает / это легко / стоять целыми днями? У меня ноги не казенные **вообще-то как бы...*** [Праздные разговоры (2006)]

Показательно, что *вообще-то* здесь сочетается с *как бы*. Не случайно и то, что выражение *вообще-то* наряду с *как бы* встретилось в приведенном русском переводе описания understatement из книги Микеша.

3.2. Между прочим

Рассмотрим еще одно языковое выражение, для которого характерна такая функция⁸. Сочетание *между прочим* чрезвычайно интересно не только своей семантикой, но и структурой многозначности. У него имеются недискурсивные употребления и связанные с ними дискурсивные. Посмотрим, какое место занимает среди них употребление в функции саркастического занижения.

Исходное значение *между прочим* — 'среди прочего, в том числе':

- (12) *Я <...> получил от редактора учтивое письмо, в котором, **между прочим**, было сказано, что мне в уме невозможно отказать, но в таланте должно, а что в литературе только талант и нужен* [И. С. Тургенев. Гамлет Щигровского уезда (1849)]

Речь здесь о том, что в письме было написано разное, в том числе высказана оценка ума и таланта адресата.

Следует отметить, что это значение несколько устарело и в современном языке в чистом виде встречается нечасто.

Если что-то упоминается среди прочего, не выделяется из остального, то этому, скорее всего, не придается особого значения. Поэтому на базе первого значения легко развивается второе:

- (13) *Чехов в отличие от Олеси делал записи **между прочим**, между художественными произведениями и без намерения сделать эти записи*

⁷ *Вообще-то* представлено как модификация *вообще* в [Путеводитель... 1993: 111], но нужный нам тип употребления там не рассматривается.

⁸ См. о нем [Инькова 2017], однако интересующие нас употребления в этой работе не затрагиваются.

художественными произведениями. [Аркадий Белинков. Сдача и гибель советского интеллигента. Юрий Олеша / Смерть поэта (1958–1968)]

Из приведенных значений очень естественно развиваются дискурсивные. В них *между прочим* маркирует ту часть текста, которую говорящий хочет подать как неосновную, необязательную. Обычно синонимом *между прочим* в таких употреблениях является *кстати*:

(14) *Думаешь, чего я сюда прибежала? Есть мысль: а что, если устроить маленькую домашнюю пьянку? У меня, между прочим, день рождения скоро. Приглашу ее, и мы начистоту поговорим с ней, а?* [Владимир Маканин. Человек свиты (1988)]

Ср. вполне возможное в этом контексте *У меня, кстати, день рождения скоро.*

Можно выделить несколько разновидностей дискурсивных употреблений *между прочим*.

Первый круг дискурсивных употреблений предполагает, что говорящий хочет начать разговор или заявить новую тему и ему необходимо как-то облегчить себе это начало:

(15) *Где ты живешь, между прочим? Я хотел бы дать твой адрес Андрею.* [Г. А. Газданов. Эвелина и ее друзья (1968)]

Второй круг связан с введением в речь побочной темы:

(16) *Его построил в середине прошлого века Александр Брюллов (брат художника, между прочим).* [Как весна, так все коты (2002) // «Домовой», 2002.03.04]

Следующий круг дискурсивных употреблений *между прочим* предполагает, что говорящий добавляет не просто что-то случайно вспомнившееся по ассоциации, но некоторую важную подробность, которая позволяет лучше понять ситуацию:

(17) *А вы мелко склочничаете, считаете копейки, которые я трачу на еду, между прочим, не из вашего кармана, да в придачу рассуждаете о бренности жизни.* [Александра Маринина. Не мешайте палачу (1996)]

Плохо, что они считают копейки, и еще хуже — что это копейки не из их кармана.

Очень типичный для *между прочим* круг употреблений — это случаи, где говорящий делает некоторое обобщение по поводу только что сказанного:

(18) *По фотографиям давал какие-то советы, вполне дельные — у него, между прочим, отличный вкус.* [Дина Рубина. Русская канарейка. Блудный сын (2014)]

В контекстах этого типа *между прочим* можно заменить на *кстати*, однако интереснее то, что контекстным синонимом *между прочим* будет и вообще: *Вообще верно говорят; Это вообще тенденция / Вообще это тенденция; ...вообще во всякий истории; У него вообще отличный вкус* и т. п.

Во всех этих типах дискурсивных употреблений *между прочим* подразумевает, что вводится некая дополнительная информация, не та, которую говорящий в первую очередь хочет донести.

Нас же интересуют случаи, когда *между прочим* используется как полемический прием, особенно в конфликтном диалоге. Ср.:

- (19) — *Цена изменилась, мы просто ценники не успели переписать!*
— *Между прочим, ценник является публичной офертой, и вы обязаны продать мне товар по этой цене!*

То, что ценник является публичной офертой, — это на самом деле совершенно не «между прочим», а как раз и есть самое главное, то, к чему апеллирует говорящий. Как будто такое «ослабляющее» слово ухудшает позицию говорящего, но на самом деле оно придает реплике некоторую задорность и напор.

Рассмотрим несколько подобных примеров:

- (20) *О чем мне с тобой говорить? Про «Спартак» — чемпион? Не интересуюсь. — Я сам, между прочим, не интересуюсь, а смотрю фигурное катание, — вдруг сказал Костя. [Алексей Слаповский. Любовь по-нашему // «Знамя», 2003]*

Собеседник возмущен презумпцией, что он интересуется футболом, и резко возражает. Интересно, что если убрать *между прочим*, реплика будет звучать, возможно, даже примирительно: — *Я сам не интересуюсь, а смотрю фигурное катание.*

В следующем примере *между прочим* очень характерно для скандальной официантки:

- (21) — *Пока это сготовят, ваш поезд уйдет! А с язвой, между прочим, по ресторанам не ходят! С язвой дома сидят!.. [Эльдар Рязанов, Эмиль Брагинский. Вокзал для двоих (1983)]*

В этом круге употреблений синонимом *между прочим* будет *вообще-то*. В таких употреблениях *между прочим* также содержится своего рода *understatement* в сочетании с сарказмом.

В последнее время вошли в моду и **новые языковые единицы**, которые по смыслу должны снижать напряженность ситуации, однако используются для полемического заострения: *на минуточку / на секундочку, если что [Это, если что, моя квартира!], а ничего, что...? [А ничего, что он мой начальник?]*.

3.3. На минуточку / на секундочку

Так, выражение *на минуточку* сейчас часто употребляется не только в значении 'ненадолго' [*взять ножницы не минуточку*], но и в особом дискурсивном значении, не имеющем отношения к указанию на время:

- (22) *Что значит / откуда я это знаю? Это на минуточку моя профессия!*
[Дмитрий Дьяченко и др. День радио, к/ф (2008)]

(23) *Роды были двойней, **на минуточку**. Естественно, что она была не обследована. А зачем? Рожать собиралась, как положено, — с «духовным» акушером, неким Славиком. В прошлом не то костоправом, не то народным целителем. [Татьяна Соломатина. Девять месяцев, или «Комедия женских положений» (2010)]*

Здесь сарказмом наполнен весь фрагмент: не только *на минуточку*, но и *естественно, а зачем?*, как положено.

Нельзя не упомянуть также замечательный «порошок», героями которого являются Дмитрий Менделеев и Александр Блок:

(24) *потом допишешь про аптеку
я **на минуточку** твой тесть
я тут спиртягу разбодяжил
потесть*

Аналогично используется выражение *на секундочку*:

(25) *Вадим Игоревич опустил голову и увидел перед собой мальчика лет десяти. Он потирал ушибленную коленку и виновато поглядывал на этого неуклюжего взрослого. — Внимательнее нужно быть, **на секундочку**, — процедил Вадим Игоревич сквозь зубы и, снова заложив руки за спину, продолжил свой путь.
— Такой взрослый парень, а все веришь в эти сказки. Нет никакого Деда Мороза. Его, **на секундочку**, не существует в природе.
[<https://fishki.net/2814582-novyj-god-na-sekundochku.html>]*

Особенность этих выражений в следующем. У них есть временное значение, и есть «саркастически понижающее». Но у них нет того употребления, по отношению к которому рассматриваемое употребление являлось бы саркастическим. Вне временного контекста они всегда употребляются саркастически.

При этом — в отличие от *на минуточку* и *на секундочку* — *на минуту*, *на минутку*, *на секунду*, *на миг*, *на мгновение* так не используются. Таким образом, здесь перед нами полностью лексикализованная фигура речи.

3.4. Если что

Следующее модное выражение из этой серии — *если что*⁹. В отличие от *на минуточку*, оно как раз может употребляться в «прямом» (понижающем) режиме: *Если что, звони*.

В последние годы это выражение стало активно использоваться в расширенном значении, приблизительно синонимично *кстати*:

(26) *Бюджет доверили считать мне. Я, **если что**, гуманитарий, и вообще не смыслю в сметах и цифрах. [<https://performance360.ru/epicfail/>]*

⁹ Сходным образом используется и оборот *на всякий случай*, который мы здесь за недостатком места не рассматриваем.

В подобных контекстах *если что* понимается примерно как 'если вдруг вы не знали' или 'если вам интересно'.

А кроме того, оно используется и для «саркастического понижения».

Так, в социальной сети человек жалуется, что необходимость вставать и здороваться с каждым сотрудником за руку очень отвлекает от работы, и собеседник советует ограничиваться словесным приветствием:

- (27) *На претензии спокойно отвечать «Я, если что, работаю» — сначала будут обижаться, возможно — дескать, «от коллектива отрываешься» — потом тупо перестанут обращать внимание.*
[https://pikabu.ru/story/privetstvie_na_rabote_nuzhna_pomoshch_4980291]

В другом случае автор жалуется, что из-за сбоя в компьютерной программе пришлось все переустанавливать:

- (28) *А Я. ЕСЛИ ЧТО, РАБОТАЮ БЕЗ ВЫХОДНЫХ ПО 12 ЧАСОВ В ДЕНЬ, ДАЖЕ ФРЕНДЛЕНТУ ПОЧИТАТЬ И НА КОММЕНТЫ ОТВЕТИТЬ НЕКОГДА.*
[<https://jo-pa57.livejournal.com/439349.html>]

В следующем примере человек пеняет в Твиттере своей собеседнице, что она не сразу отвечает, она оправдывается, что, мол, не успевает, но он не принимает оправданий:

- (29) — *ты меня вк игноришь!!!!!!!*
— *я тебе не успеваю отвечать потому что!!!!!!*
— *А ДОЛЖНА УСПЕВАТЬ **ВООБЩЕ-ТО ЕСЛИ ЧТО ТАК***
[<https://twitter.com/liyuom/status/910179754023833600>]

Здесь интересно, что пишущий использует для убедительности сразу три оборота, выражающие низкую тимеологическую оценку: *вообще-то если что так*.

3.5. *Ничего, что..?*

И, пожалуй, выражение с наибольшим «полюемическим задором» из этой серии — это оборот (*a*) *ничего, что..?* У него тоже есть «прямой» режим употребления:

- (30) — *Ничего, что я к вам спиной?*
(31) — *А ничего, что я в обуви? — Ничего, я еще пол не мыла.*

Однако в последние годы необычайно популярно его использование в саркастическом смысле. Причем поскольку в прямом режиме *ничего что..?* оформляет просьбу о разрешении или извинение, то есть речевые акты, предполагающие зависимость говорящего от адресата, то при саркастическом употреблении фразы с *ничего что* звучат издевательски. Например, адвокат Анна Ставицкая язвительно пишет в соцсети:

(32) *Сегодня Генпрокуратура РФ подала кассационное представление на приговор по делу Константина Котова и просила снизить наказание Котову до 1 года лишения свободы! <...> **А ничего что** прокурор в суде первой инстанции просил 4,5 года лишения свободы Котову, а в апелляции прокуратура тоже была согласна с наказанием!? Получается, что не правы были прокуроры, которые такое суровое наказание просили и их теперь тоже надо наказывать?!!!* [https://www.facebook.com/permalink.php?story_fbid=3401467319923990&id=100001822665357]

Приведем в качестве примера также запись, которая целиком построена на обороте *ничего, что?*. Сюжет состоит в том, что девушке показалась, что продавщица бутика была с ней недостаточно любезна:

(33) *И вот я стою в этом элитном магазине, а внутри всё разрывает смех!!! <...> Потому что, если говорить, как есть — **ничего, что** я свои первые «Своровски» купила в Париже, просто так — как память о Лувре? **А ничего, что** я была одета в натуральные замшевые сапоги со вставками змеиной кожи? **А ничего, что** я хожу только в натуральном Wilson? **А ничего, что** у меня всё золото украшено брелочками? **А ничего, что** у меня крутой телефон? **А ничего, что** я не ношу ни каких часов кроме «Swatch»? **А ни чего, что** я не работаю, а просто сижу дома и занимаюсь чем захочу? **А ничего, что** я стригусь у модных парикмахеров? **А ничего, что** у меня вся косметика и парфюмерия куплена в Sefora (Брокер — даже и рядом не стоял)? **А ничего, что** я живу в крутом районе? **А ни чего, что** у меня парень работает в Майкрасофт и на него молятся пол Украины Ит-шников? **А ничего, что** я еще тольком и сумки не распаковала после прогулок в Барселоне? **А ничего, что** еще утром я сдавала документы для поездки в Швейцарию на Новый год? и т.д... И после всего вышеперечисленного, мне какая то тупая продавщица будет своим жалким ртом, спрашивать — это я для себя цену узнаю??? Или хватит ли у меня денег???* [<http://kangelka.spaces.live.com/blog/cns!CB1A849DC712789E!1034.entry?sa=38114978>]

Очень часто это выражение записывается как *ниче что..?*:

(34) *Я настаиваю, чтобы звонки из вашего банка меня в ночное и любое другое время больше меня не беспокоили, вы там не офигели-ли, ночью звонить? **Ниче, что** людям на работу утром и они уже спят? После следующего звонка напишу жалобу в ФАС.* [<https://www.banki.ru/services/responses/bank/response/7620603/>]

(35) ***А ничё что** на том перекрёстке в принципе запрещены все левые повороты? Впрочем, о чём это я? Наплевательское и хамское отношение друг к другу на дорогах просто зашкаливает. А тут... Какой-то запрет на поворот...* [<https://ngs.ru/profile/168026/>]

4. Выводы

Итак, в современном русском языке, особенно в разговорной речи, чрезвычайно популярна риторическая фигура, представляющая собой сочетание understatement и сарказма (саркастическое занижение). У некоторых языковых единиц соответствующее значение частично или полностью лексикализировано: в ряде случаев образует особый тип употребления лексемы, а в случае *на минуточку* и *на секундочку* — отдельное значение.

Вообще сарказм трудно формализуем. Нередко это приводит даже к тому, что авторы саркастических высказываний попадают под следствие и суд, в том числе по «экстремистским» статьям. Поэтому обнаружение лингвистических манифестаций сарказма в перспективе может иметь не только теоретическое, но и прикладное значение.

References

1. ADR (2014) — The active dictionary of Russian [Aktivnyi slovar' russkogo yazyka], vol. 1–2. Apresjan Yu. D. (ed.), Moscow: Yazyki Slavyanskoi Kul'tury.
2. In'kova O. (2017) Kstati and mezhdru protchim: so much alike and yet different [Kstati i mezhdru prochim: takie blizkie i takie dalekie]. Vestnik Mosk. un-ta, №6, pp. 51–65.
3. Larina T. V. (2009) Category of politeness and style of communication. Comparison of the English and Russian linguocultural traditions [Kategoriya vezhливости i stil' kommunikacii: sopostavlenie anglijskih i russkih lingvokul'turnyh tradicij]. M.: Yazyki slavyanskikh kul'tur, 512 p.
4. Otkidych E. V. (2017) Mezhdru prochim // function words in lexicographical aspect [Mezhdru prochim // Sluzhebnye slova v leksikograficheskom aspekte] / Pod red. E. S. Sheremet'evoj, E. A. Starodumovoj, P. M. Tyurina. Vladivostok, pp. 276–285.
5. Pen'kovskij A. B. (1995) Timiological evaluations and their expression in intentional deviation from the truth [Timiologicheskie ocenki i ih vyrazhenie v celyah uklonyayushchegosya ot istiny umaleniya znachimosti]. Logicheskij analiz yazyka: Istina i istinnost' v kul'ture i yazyke / Otv. red. N. D. Arutyunova, N. K. Ryabceva. M.: Nauka, pp. 36–41.
6. Putevoditel' (1993) Baranov A. N., Plungyan V. A., Rakhilina E. V. Guide to discursive words of Russian. [Putevoditel' po diskursivnym slovam], 207 pp.
7. Tripol'skaya T. A. (1992) The meaning of importance [Modusnoe znachenie «vazhnosti» vyskazyvaniya i sposoby ego vyrazheniya]. Modal'nost' v eyo svyazyah s drugimi kategoriyami pod red. I. P. Mathanovoj. Novosibirsk: Izd-vo NGPI, pp. 27–33.
8. Uryson E. V. (1998). Unaccomplished polysemy and its types [“Nesostoyavshayasya polisemiya” i nekotorye ee tipy], Semiotika i informatika. Vyp. 36. M. pp. 226–262.

9. *Zaliznyak Anna A.* (2004) Reconstruction of the «true meaning» of utterances in a conflictual dialogue: a problem statement [Rekonstrukciya «istinnogo smysla» replik v konfliktnom dialoge: postanovka problem]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2004” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2004”] M., «Nauka». [Online]. Available: <http://www.dialog-21.ru/archive/2004/Zalizniak%20Anna.pdf>.

SENTIMENT FRAMES FOR ATTITUDE EXTRACTION IN RUSSIAN

Loukachevitch N. V. (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

Rusnachenko N. L. (kolyarus@yandex.ru)

Bauman Moscow State Technical University, Moscow, Russia

Texts can convey several types of inter-related information concerning opinions and attitudes. Such information includes the author's attitude towards mentioned entities, attitudes of the entities towards each other, positive and negative effects on the entities in the described situations. In this paper, we described the lexicon RuSentiFrames for Russian, where predicate words and expressions are collected and linked to so-called sentiment frames conveying several types of presupposed information on attitudes and effects. We applied the created frames in the task of extracting attitudes from a large news collection.

Keywords: sentiment analysis, connotation, distant supervision

DOI: 10.28995/2075-7182-2020-19-541-552

ОЦЕНОЧНЫЕ ФРЕЙМЫ ДЛЯ ИЗВЛЕЧЕНИЯ ОЦЕНОЧНЫХ ОТНОШЕНИЙ ИЗ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Лукашевич Н. В. (louk_nat@mail.ru)

МГУ имени М. В. Ломоносова, Москва, Россия

Русначенко Н. Л. (kolyarus@yandex.ru)

МГТУ имени Н. Э. Баумана, Москва, Россия

1. Introduction

Sentiment analysis of texts is one of the most important directions in natural language processing research. Numerous papers were devoted to problems of automatic classification of whole texts or text fragments to two, three or five classes according to the opinion found in this text. Currently, researchers began to consider

several types of interrelated phenomena that previously had considered as the same type of sentiment attitudes.

[[Mohammad 2016](#)] indicates the following different subtypes of phenomena interplaying with author sentiment: speaker emotion state vs. polarity of opinion mentioned in a sentence; success or failure vs. opinion about the success of failure; neutral reporting of a negative (war) or positive (celebration) event etc. Additionally, an author can mention some negative or positive relations between entities without conveying his/her own stance [[Loukachevitch, Rusnachenko, 2017](#)]. [[Feng et al. 2013](#)] consider polar connotations (sentiment associations) of objective words such as *unemployment*, *war* (negative connotations) or *human rights*, *sun* (positive connotations). These words do not express direct sentiments but may accompany or contradict attitudes conveyed in the text. [[Choi et al. 2014](#)] consider the interaction between described good (bad) effects and conveyed polarities.

Thus, we can see a scope of various phenomena related to sentiment analysis, but most existing sentiment vocabularies have a simple structure as lists of words or expressions with positive or negative sentiment scores. For example, in the news title “White House Blocked 2018 Statement Condemning Russia”, both named entities ‘White House’ and ‘Russia’ are located in the context of negative words *block* and *condemn* (according to the MPQA lexicon [[Wilson et al., 2005](#)]) but we do not infer any author’s stance to White House or Russia. Besides, we can suppose some positive effects from the described situation to Russia and no sentiments or effects on White House.

At the same time, similar language means can be used to convey implicit attitudes of the author [[Liu, 2012](#)]; [[Nozza et al., 2017](#)]. For example, in the sentence “He approved the bombing of civilians” the author of the text is negative towards the subject. Here the following issues interplay: a positive attitude to civilians presupposed for most people, a negative action towards civilians and approving this event by the subject. The interaction of the above-mentioned phenomena usually correlates with specific sentiment predicates [[Rashkin et al., 2016](#)]; [[Klenner et al., 2017](#)], which can have some presupposed information about attitudes between participants of the situation described by the predicate or effects on the participants as a result of the situation.

In this paper we consider a specific lexicon called RuSentiFrames for Russian, where predicate words and expressions are collected and linked to so-called sentiment frames conveying several types of presupposed information on attitudes and effects. We give the detailed description of RuSentiFrames and describe methods for evaluation of the created resource.

The structure of the paper is as follows. **Section 2** describes related work. In **Section 3** we present the structure of the RuSentiFrames lexicon, the principles of describing its lexical entries. **Section 4** is devoted to evaluation of the RuSentiFrames lexicon.

2. Related work

Most sentiment vocabularies are presented as lists of words and expressions with scores of their sentiment [[Wilson et al. 2005](#)]. Some vocabularies provide also additional characteristics of the word sentiment called as ‘strength’. Also sentiment scores

can be assigned to specific senses of ambiguous words [Baccianella et al., 2010]; [Loukachevitch and Levchik, 2016]. For more accurate extraction of sentiment attitudes cited or expressed in texts, it is not enough to have a simple sentiment list with sentiment scores assigned to words and expressions [Neviarouskaya, 2009]; [Deng and Wiebe, 2014].

In [Rashkin et al., 2016] the authors stressed that it is important to extract implied sentiments and proposed the approach to description of so-called connotation frames for transitive verbs. The description includes three participants' roles: agent, theme, and writer. The frame includes the attitude polarity of participants to each other (positive, negative, or neutral), the effect of the situation to agent or theme, mental states and values of the participants. For experiments, 1,000 most frequent English verbs were extracted from a corpus. These verbs were provided with five example sentences constructed from most frequently seen Subject-Verb-Object triples in Google syntactic ngrams and were annotated by crowdsourcers. The obtained values were averaged.

[Klenner et al. 2014], [2016] described a verb resource for German containing verb polarity frames. Each frame consists of the subcategorization frame and the polarity (positive or negative) effects associated with the roles. Also so-called verb signature is assigned. The verb signature indicates the factuality of roles in dependence on various factors (such as negation, mood, etc.). By 2017, 1,500 verb-polarity frames for 1,100 verbs were described. Some nominalizations (for example, *destruction*) have been also considered as frame entries [Klenner et al., 2017].

[Deng et al. 2014] considered events that positively or negatively affect entities (goodFor/badFor). For example, lowering something is bad for this something, but creating something is good for this something. This paper also describes the sentiment inference when the sentiment is conveyed towards a bad or good event.

Several Russian sentiment lexicons of sentiment words with scores have been published. In [Chetviorkin and Loukachevitch, 2012], automatically generated Russian sentiment lexicon in the domain of products and services (ProductSentiRus) is described. The ProductSentiRus is obtained by application of a supervised model to user's review collections in several domains. It is presented as a list of 5,000 words ordered by the decreased probability of their sentiment orientation without any positive or negative labels.

The general Russian lexicon of sentiment words and expressions, RuSentiLex, was created in a semi-automatic way [Loukachevitch and Levchik, 2016]. In structure, RuSentiLex is also a list of words and expressions having several attributes. The entries of the RuSentiLex lexicon are classified according to four sentiment categories (positive, negative, neutral, or positive/negative) and three sources of sentiment (opinion, emotion, or fact). The words in the lexicon having different sentiment orientations in different senses are linked to appropriate concepts of the Russian thesaurus RuThes [Loukachevitch and Dobrov, 2014], which can help disambiguate sentiment ambiguity in specific domains or contexts.

Russian Sentiment Lexicon Linis Crowd has been created via crowdsourcing [Koltsova et al. 2016]. The lexicon is aimed at detecting sentiment in user-generated content (blogs, social media) related to social and political issues. Each word was

assessed by at least three volunteers in the context of three different texts and scored from -2 (negative) to +2 (positive).

Several international lexicons were automatically constructed for Russian. The Chen-Skienna lexicon (2,876 words) [Chen and Skienna, 2014] was generated for 136 languages via graph propagation from seed words, including Russian. [Mohammad and Turney 2013] generated the Russian variant of the EmoLex lexicon with automatic translation from the English lexicon obtained by crowdsourcing (4,412 Russian words). [Kotelnikov et al. 2018] studied available Russian sentiment lexicons and found that all the lexicons have relatively small intersection with each other. They compared the above-mentioned Russian lexicons as features in machine-learning text categorization of user's reviews in several domains using the SVM method. It was found that the best results of classification using a single lexicon in all domains were obtained with ProductSentiRus [Chetviorkin and Loukachevitch 2012]. The union of all lexicons gives slightly better results.

3. Sentiment Frames for Sentiment Analysis in Russian

Russian Sentiment Lexicon RuSentiFrames describes sentiments and connotations conveyed with a predicate word in either verbal or nominal form¹.

3.1. General Structure of Sentiment Frames

In this study, *sentiment frame* is a set of positive or negative associations (connotations) related to a predicate word or expression. A predicate usually describes a situation with some participants. The types of connotations that are conveyed in sentiment frames are as follows:

- attitude of the author of the text towards mentioned participants,
- positive or negative sentiment between participants,
- positive or negative effects on participants,
- positive or negative mental states of participants related to the described situation.

To describe participants of a situation, we should designate predicate-specific roles. There are several approaches to sets of semantic roles such as universal sets [Jackendoff, 1992], frame elements as specific roles for each frame as in FrameNet [Fillmore and Baker, 2001], or enumerated roles as in PropBank [Palmer et al., 2005] and AMR representation [Banarescu et al., 2013]. It is very difficult to choose an appropriate universal set of semantic roles. Therefore we accepted the approach of PropBank. In this approach, individual verb's semantic arguments are numbered, beginning with zero. For a particular verb, Arg0 is generally the argument exhibiting features of a Prototypical Agent [Dowty, 1991], while Arg1 is a Prototypical Patient or Theme.

All assertions are provided with the score of confidence, which currently has two values: 1, if we believe that this assertion is true almost always, or 0.7, if we consider

¹ <https://github.com/nicolay-r/RuSentiFrames/tree/v2.0>

the assertion as default. We do not describe assertions about neutral sentiment, effect and state of participants.

Figure 1 presents the attitudes and effects for the *condemn* frame:

```

frame: осудить (to condemn)
  "roles": {
    "A0": "who condemns",
    "A1": "who is condemned",
    "A2": "grounds for condemnation",
    "A3": "punishment"},
  "polarity":
    [[ "A0", " A1", "neg", 1.0],
    [ "A0", " A2", "neg", 1.0],
    [ "A0", " A3", "neg", 1.0],
    [ "A1", " A0", "neg", 1.0],
    [ "A1", " A3", "neg", 1.0]],
  "effect": [[ "A1", "-", 1.0]],
  "state":  [[ "A1", "neg", 1.0]]

```

Fig. 1: Frame example

This frame means that a condemner A0 is negative to a condemned person A1, to the grounds of condemnation A2 and knows that the punishment A3 is also something negative. A1 is negative to A0 (for the condemnation) and to punishment. The effect on A1 and the state of A1 are negative. We denote effects with signs “-” or “+” to highlight that this information is different from sentiment attitudes. For the example verb, we cannot guess the author’s opinions on the described situation and their participants therefore the corresponding assertions are absent in the frame.

If words or expressions are semantically related and have the same roles and associated connotations, we assign them to the same frame. Currently, we do not associate roles with syntactic means of their expression planning to gather this information in a semi-automatic manner.

3.2. Procedure of Frame Description

Experts describe frames using the following steps:

- choosing a target word with positive or negative connotations,
- search for semantically related words (synonyms, hyponyms) and expressions with similar connotations,
- introducing the main roles of a situation described with the target word,
- description of connotations of relations between participants, an author to participants, effects and states. The expert should check the usage of the target words analyzing sentence examples in contemporary Russian texts.

In complicated cases, the following additional guidelines are applied.

In description of connotations, social and human rights values should be accounted for, if relevant. For instance, this means that the prior author's attitude toward the agent (A0) in the killing event is negative. In examples such as “The police killed the terrorist”, possible positive polarity of the sentence can be felt. But we suppose that such a polarity arises from the negative word “terrorist” and a common-sense rule that if a negative action is directed to a negative entity then the whole situation is considered as positive [Deng and Wiebe, 2014]. In any case, in the above-mentioned example we do not deal with the prior connotations but with their transformations in the context of negative participants. Such rules are planned to add to be utilized together with the RuSentiFrames lexicon.

Sometimes it is very useful to search for examples of usages of a target word with positive and negative participants and check how polarities change. For example, in Russian there is verb *укокошить* (finish off as to kill), which means a colloquial expression for killing with possible negative polarity to the object of the situation (A1). We can try to search for or construct sentences with positive words (such as *герой* (hero)) as the A1 role: “*укокошить героя*” (? to finish the hero). It is interesting that in such a context for Russian no negativity of the situation is conveyed. The resulted polarity looks like a negative irony towards to *hero*, the positiveness of *hero* is decreased. Thus, we can conclude that the connotations with the A1 role of *укокошить* are really negative.

3.3. RuSentiFrames Entries: Single Words and Expressions

We formed the list of words for the description in RuSentiFrames in the following way.

At first, we included predicate words and expressions from the RuSentiLex lexicon [Loukachevitch and Levchik, 2016], Besides, we looked through the word frequency list obtained from a news text collection to find words conveying good or bad effects: *создать* (to create), *понижить* (to lower), *рост* (growth), *увеличивать* (to enhance), etc., because these words are usually not included into sentiment lists. At last, we analyzed news reports about war actions collecting words and expressions with positive and negative effects.

The created frames are associated not only with a single entry but with a “family” of related words and expressions, which have the same attitudes. As a result, the following lexical units can be associated with a sentiment frame:

- single words: mainly verbs and nouns,
- idioms: *вешать лапшу на уши* (to hang noodles on the ears—to lie), *взять за горло* (to take by the throat—to press);
- light verb constructions: *нанести вред* (inflict harm), *нанести обиду* (cause offence), *нанести поражение* (inflict a defeat), etc.;
- verbs (or nouns) with prepositions. Sentiment attitudes conveyed by some words can significantly depend on a preposition used with this word: for example, *выступить против* (to speak against—to oppose), *завязывать с* (to stop doing something);
- compositional expressions synonymous to a lexical unit or multiword expression, which has the sentiment frame.

For example, the sentiment frame called *Запретить* (to forbid) contains 53 text entries including such expressions as *налагать запрет* (to impose a ban), *наложение запрета* (imposition of a ban), *закрывать доступ* (to close access), *закрытие доступа* (closing access), *прекращение доступа* (termination of access), *прекратить доступ* (to terminate access), *налагать вето* (to impose veto), etc.

Some frame entries are ambiguous. In such cases, they are attached to different frames, for example, Russian verb *выгореть* can mean 'to have success' and 'to be destroyed by fire'. Currently, it is difficult to automatically disambiguate such words but they should be assigned to relevant frames for future studies.

Currently, RuSentiFrames contains 311 frames with more than 7K associated frame entries. **Table 1** shows the distribution of the RuSentiFrames entries according to parts of speech and other characteristics.

Table 1: Quantitative characteristics of the RuSentiFrames entries

Type of lexical unit	Number
Verbs	3,239
Nouns	986
Phrases	2,551
Other	12
Unique entries	6,788
Total entries	7,034

3.4. Classes of Frames

Most frames contain several types of slots, including the attitudes between participants, the authors' attitude, and effects on the participants. Some frames convey only information about relations between participants, but the information about the author's attitudes towards the participants and the event is absent. For example, for verbs *надеяться* (to rely on) or *создать* (to create), the first participant (A0) is positive towards the second participant (A1).

Some words are very similar in sense but differ in the author's attitude to a participant. For example, Russian verbs *наказать* and *карать* are very similar in sense (to punish). But for verb *наказывать*, the author seems neutral to the situation in default, but for *карать*, the author seems positive to the agent of the situation (A0) and negative to A1 (who is punished).

Table 2 describes the distribution of frame entries according to sentiments between main participants of the situation and from the author to the participants. **Table 3** describes the distribution of the RuSentiFrames entries according to effects on main participants A0 and A1. Some frames contain up to 4 roles associated with some attitudes (from the author or from the other participants) or effects, for example, frames *осудить* (to condemn), *выиграть* (to win), *выгнать* (to kick out), etc.

Table 2: The distribution of RuSentiFrames text entries according to attitudes

Relations	Sentiment	Number
A0 to A1	Pos	2,558
A0 to A1	Neg	3,289
Author to A0	Pos	170
Author to A0	Neg	1,581
Author to A1	Pos	92
Author to A1	Neg	249

Table 3: The distribution of RuSentiFrames text entries according to effects on main participants

Effect	Sentiment	Number
A0	Pos	1,008
A0	Neg	733
A1	Pos	2,355
A1	Neg	3,504

If to compare with other sentiment-oriented resources we can say that RuSentiFrames is the only structured resource for Russian, all other existing sentiment lexicons are lists of words with attributes. If to compare with connotation frames [Rashkin et al., 2016], they took only 1,000 most frequent transitive verbs. But we created a much more structured resource: more than 6 thousand words and expressions, up to four roles, the grouping of words and expressions with similar connotations to the same frames, more elaborate analysis of complex cases.

4. Coverage and Evaluation of RuSentiFrames

To check the agreement in the description of the created frames, two experiments were carried out. For each experiment, 200 different words (100 for each experiment) were randomly selected.

In the first experiment, two experts described frames for selected words in parallel using their intuition and text examples,

In the second experiment, one expert created frames and gave only roles (without connotations) to an annotator. The annotator gathered 10 random non-duplicate sentences for each word from different topics of the current news flow. The task of the annotator was to assign positive or negative scores to each role of the word mentioned in a sentence under analysis. The obtained scores were averaged. The average scores and connotations were compared with the original frame of the word.

We cannot estimate inter-annotator agreement using usual techniques [Bobichev and Sokolova, 2017], because any expert can add or miss some connotations, or do not reveal them in texts. We can consider two sets of connotations for each of the experts (E_1 and E_2). We calculate the intersection between two these sets $E_1 \cap E_2$, which include the same dimensions with the same positive or negative scores. Then we consider ratios:

$$R_1 = \frac{E_1 \cap E_2}{E_1} \quad R_2 = \frac{E_1 \cap E_2}{E_2}$$

Table 4. Agreement between different experts in creating and annotating frames

Measure	Expert1 vs. Expert2	Expert to Annotator
R_1	0.81	0.82
R_2	0.72	0.75
Harmonic mean	0.76	0.78

At last, we calculate the harmonic mean (HM) between two ratios to average them (like for F-measure). The results presented in Table 4 show that there is a considerable share of core connotations, for which opinions of experts (annotators) coincide.

To evaluate RuSentiFrames in an analytical task, we extracted frequent attitudes between named entities in the news flow and after that we could estimate if the extracted attitudes correlate with factual attitudes. We exploited a news corpus of 2017 consisting of 2.5 M news articles. Named entities were extracted with DeepPavlov library². We considered sentences where at least two named entities and at least a single frame entry (internal frame entry) between them were mentioned. The corresponding frames should have positive or negative labels for A0 to A1 attitudes.

We assigned the positive sentiment score when all the polarities of the internal frame entries had the positive sentiment. Otherwise, the negative sentiment score was assigned. We also consider the frame entry polarity as inverted, when it was used with negation. **Table 5** presents the most negative attitudes found in the corpus. **Table 6** shows the most positive attitudes from the same corpus.

Table 5. The most negative attitudes found in the 2017 news corpus

A0	A1	Frequency of co-occurrence in frames	Positive	Negative
Нагорный Карабах (Nagorno-Karabakh)	Азербайджан (Azerbaijan)	123	0 (0%)	123 (100%)
Израиль (Israel)	Дамаск (Damascus)	41	0 (0%)	41 (100%)
Пентагон (Pentagon)	Аль-Каеда (Al Qaeda)	36	0 (0%)	36 (100%)
Россия (Russia)	ИГИЛ (ISIL)	245	19 (7.8%)	226 (92.2%)
Киев (Kiev)	Госдума (State Duma)	158	3 (1.9%)	155 (98.1%)
Турция (Turkey)	ИГИЛ (ISIL)	142	6 (4.2%)	138 (95.8%)
Азербайджан (Azerbaijan)	Армения	96	4 (4.2%)	92 (92.8%)

² <http://docs.deeppavlov.ai/en/master/features/models/ner.html>

According to the frame-based analysis, the USA-Russia relations were mainly indicated as negative (65%), the USA-Ukraine relations are mainly positive (64%), the EU-Russia relations are mainly negative (78%), the Great Britain-Russia relationships are mainly negative (59.2%). Thus we can see that the extracted attitudes correspond to factual attitudes between entities. In future, the extracted attitudes will be analyzed in a more detailed way to improve frame descriptions.

Table 6. The most positive attitudes found in the 2017 news corpus

A0	A1	Frequency of co-occurrence in frames	Positive	Negative
Шойгу (Shoigu)	Путин (Putin)	44	44 (100%)	0 (0%)
НАТО (NATO)	Эстония (Estonia)	30	30 (100%)	0 (0%)
Украина (Ukraine)	МВФ (IMF)	185	172 (93%)	13 (7%)
Порошенко (Poroshenko)	НАТО (NATO)	169	165 (97.6%)	4 (2.4%)
Канада (Canada)	Украина (Ukraine)	141	135 (95.7%)	6 (4.3%)
Украина (Ukraine)	НАТО (NATO)	103	96 (93.2%)	7 (6.8%)
США (USA)	Черногория (Montenegro)	100	99 (99%)	1 (1%)

The created lexicon has been also already used for creating a training collection in the distant supervision framework [Rusnachenko et al., 2019].

5. Conclusion

Texts can convey several types of inter-related information concerning opinions and attitudes. Such information can include the author's attitude towards mentioned entities, attitudes of the entities towards each other, positive and negative effects on the entities in the described situations. The effect extraction is often important for the attitude analysis because the positive attitude towards negative situation for an entity is usually means the negative attitude towards the entity, and vice versa.

In this paper, we described the lexicon RuSentiFrames for Russian, where predicate words and expressions are collected and linked to so-called sentiment frames conveying several types of presupposed information on attitudes and effects. We applied the created frames in the task of extracting attitudes from a large news collection.

6. Acknowledgements

The reported study is partially funded by RFBR, research project № 20-07-01059.

References

1. *Baccianella S., Esuli A., Sebastiani F.* (2010), Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. LREC-2010.—V. 10, pp. 2200–2204.
2. *Baker, C., Fillmore Ch., Lowe J.* (1998), The Berkeley FrameNetproject. In: Proceedings of COLING/ACL., pp. 86–90, Montreal.
3. *Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Schneider, N.* (2013), Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 178–186.
4. *Bobicev, V., Sokolova, M.* (2017), Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. In Proceedings of RANLP-2017 conference, pp. 97–102.
5. *Chetviorkin, I., Loukachevitch, N.* (2012), Extraction of Russian Sentiment Lexicon for Product Meta-Domain. In COLING-2012, pp. 593–610.
6. *Chen Y., Skiena S.* (2014), Building Sentiment Lexicons for All Major Languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 383–389.
7. *Choi, Y., Deng, L., Wiebe, J.* (2014), Lexical acquisition for opinion inference: A sense-level lexicon of benefactive and malefactive events. In Proc. of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Baltimore, Maryland. Association for Computational Linguistics.
8. *Choi E., Rashkin H., Zettlemoyer L., Choi Y.* (2016), Document-level Sentiment Inference with Social, Faction, and Discourse Context. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL-2016, pp. 333–343.
9. *Deng L., Wiebe J.* (2014), Sentiment propagation via implicature constraints. Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2014).
10. *Dowty D.* (1991), Thematic proto-roles and argument selection. *Language*, 67(3): 547–619.
11. *Feng S., Kang J. S., Kuznetsova P., Choi Y.* (2013), Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning. In *ACL (1)*, pp. 1774–1784.
12. *Fillmore Ch., Baker, C.* (2001), Frame semantics for text understanding. In Proceedings of NAACL WordNet and Other Lexical Resources Workshop.
13. *Jackendoff R.* (1992), *Semantic structures*. MIT press, 1992.
14. *Klenner M., Amsler M., Hollenstein N.* (2014), Verb polarity frames: a new resource and its application in target-specific polarity classification. In G. Faaß (Ed.), *KONVENS*, pp. 106–115.

15. *Klenner M, Amsler M.*, (2016), Sentiframes: A resource for verb-centered German sentiment inference. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), pp. 2888–2891.
16. *Klenner M., Tuggener D., Clematide S.* (2017), Stance detection in Facebook posts of a German right-wing party. Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pp. 31–40.
17. *Koltsova O. Y., Alexeeva S. V., Kolcov S. N.* (2016), An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media. Computational Linguistics and Intellectual Technologies, pp. 277–287.
18. *Kotelnikov E., Peskischeva T., Kotelnikova A., Razova E.* (2018), Comparative Study of Publicly Available Russian Sentiment Lexicons. In Conference on Artificial Intelligence and Natural Language, pp. 139–151.
19. *Liu B.* (2012), Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers.
20. *Loukachevitch N., Dobrov B.* (2014), RuThes linguistic ontology vs. Russian word-nets. In Proceedings of the Seventh Global Wordnet Conference, pp. 154–162.
21. *Loukachevitch N., Levchik A.* (2016), Creating a General Russian Sentiment Lexicon. In Proceedings of Language Resources and Evaluation Conference LREC-2016, 2016.
22. *Loukachevitch N., Rusnachenko N.* (2018), Extracting sentiment attitudes from analytical texts. arXiv preprint arXiv:1808.08932.
23. *Mohammad S. M.* (2016), A practical guide to sentiment annotation: Challenges and solutions. Proceedings of NAACL-HLT, pp. 174–179.
24. *Mohammad S. M., Turney D. P.* (2013), Crowdsourcing a word-emotion association lexicon. Computational Intelligence 29(3): 436–465.
25. *Neviarouskaya A., Prendinger H., Ishizuka M.* (2009), Semantically distinct verb classes involved in sentiment analysis. IADIS AC (1), pp. 27–35.
26. *Nozza D., Fersini E., Messina E.* (2017), A Multi-View Sentiment Corpus. In Proceedings of EACL-2017, pp. 273–280.
27. *Palmer M., Gildea D., Kingsbury P.* (2005), The proposition bank: An annotated corpus of semantic roles. Computational linguistics: 71–106.
28. *Rashkin H., Singh S., Choi Y.* (2016), Connotation Frames: A Data driven Investigation. Proceedings of Association for Computational Linguistics Conference ACL-2016, pp. 311–322.
29. *Rusnachenko N., Loukachevitch, N., Tutubalina, E.* (2019), Distant supervision for sentiment attitude extraction. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 1022–1030.
30. *Wilson T., Wiebe J., Hoffmann P.* (2005), Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the conference on human language technology and empirical methods in natural language processing, pp. 347–354.

GRAMEVAL 2020 SHARED TASK: RUSSIAN FULL MORPHOLOGY AND UNIVERSAL DEPENDENCIES PARSING¹

Lyashevskaya O. N. (olesar@yandex.ru)

National Research University Higher School of Economics;
V. V. Vinogradov Russian Language Institute of RAS,
Moscow, Russia

Shavrina T. O. (rybolos@gmail.com)

National Research University Higher School of Economics;
Sberbank, Moscow, Russia

Trofimov I. V. (itrofimov@gmail.com),

Vlasova N. A. (nathalie.vlassova@gmail.com)

A. K. Ailamazyan Program Systems Institute of RAS,
Pereslavl-Zalessky, Russia

The paper presents the results of GramEval 2020, a shared task on Russian morphological and syntactic processing. The objective is to process Russian texts starting from provided tokens to parts of speech (pos), grammatical features, lemmas, and labeled dependency trees. To encourage the multi-domain processing, five genres of Modern Russian are selected as test data: news, social media and electronic communication, wiki-texts, fiction, poetry; Middle Russian texts are used as the sixth test set. The data annotation follows the Universal Dependencies scheme. Unlike in many similar tasks, the collection of existing resources, the annotation of which is not perfectly harmonized, is provided for training, so the variability in annotations is a further source of difficulties. The main metric is the average accuracy of pos, features, and lemma tagging, and LAS.

In this report, the organizers of GramEval 2020 overview the task, training and test data, evaluation methodology, submission routine, and participating systems. The approaches proposed by the participating systems and their results are reported and analyzed.

Key words: morphological tagging, dependency parsing, lemmatization, NLP evaluation, GramEval shared task, Russian

DOI: 10.28995/2075-7182-2020-19-553-569

¹ The publication was partly prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2020 and within the framework of the Russian Academic Excellence Project «5-100».

GRAM EVAL 2020: ДОРОЖКА ПО АВТОМАТИЧЕСКОМУ МОРФОЛОГИЧЕСКОМУ И СИНТАКСИЧЕСКОМУ АНАЛИЗУ РУССКИХ ТЕКСТОВ

Ляшевская О. Н. (olesar@yandex.ru)

НИУ Высшая Школа Экономики; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Шаврина Т. О. (rybolos@gmail.com)

НИУ Высшая Школа Экономики; Сбербанк, Москва, Россия

Трофимов И. В. (itrofimov@gmail.com),

Власова Н. А. (nathalie.vlassova@gmail.com)

Институт программных систем им. А. К. Айламазяна РАН, Переславль-Залесский, Россия

GramEval 2020 — дорожка по оценке методов и технических решений для полного морфологического и синтаксического анализа текстов на русском языке. В 2020 году доминантой была выбрана жанровая репрезентативность текстового материала. Для оценки подходов к автоматическому анализу текста был подготовлен тестовый набор данных, охватывающий пять жанров современного языка: новости, сообщения из социальных сетей и электронную коммуникацию, энциклопедические статьи, художественную литературу, поэзию, а также исторические тексты 17 века.

Текстовый материал для обучения и тестирования предоставлялся в формате Универсальных Зависимостей (Universal Dependencies) версии 2.5. Входной формат содержал информацию о границах предложений и токенов. Задачей систем-участников было определить часть речи, грамматические признаки и лемму каждого токена, а также построить дерево зависимостей каждого предложения с типизацией синтаксических отношений.

В ходе мероприятия участники имели возможность получать оценки качества своих решений благодаря платформе CodaLab. Автоматически предоставлялась детализация оценок по уровням разметки и текстовым регистрам, информация о частотных ошибках. Окончательный рейтинг систем составлялся на основе четырёх показателей: качества определения части речи, грамматических признаков, леммы и построения дерева зависимостей (LAS).

В данной статье организаторы GramEval 2020 рассматривают основные вопросы, связанные с организацией дорожки, а также полученные участниками результаты. Затрагиваются темы методологии оценки, подготовки обучающих и тестовых данных. Приводится краткое описание подходов участников и анализ допущенных ошибок.

Ключевые слова: морфологический анализ, синтаксический парсинг, парсинг зависимостей, лемматизация, оценка систем автоматической обработки текста, дорожка GramEval, русский язык

1. Introduction

Russian grammar has a rich history of theoretical and applied modelling. Starting with the work of A. Zaliznyak [12], the grammatical description has reached a new level, making it possible to build automatic systems of morphological analysis.

Since then, technologies in Russian NLP have made significant advances thanks to data from search engines [7], [1], as well as to shared tasks based on texts from various sources. Since 2010, automatic morphological tagging has become a traditional task for Russian and international researchers.

In 2010, for the first time, a shared task was held for automatic Russian part-of-speech tagging, lemmatization, and morphological analysis, including the subtask of annotating the rare words [5]. The participants achieved 98% accuracy on lemmatization and 97.3% accuracy on the part-of-speech tags.

At the MorphoRuEval 2017 shared task [8], a 97.11% accuracy in all morphological features and 96.91% accuracy in lemmatization were achieved on a balanced set of data from various sources (news, social networks, fiction, etc.).

From 2016 to 2019, morphology also became the main focus of the multilingual competition SIGMORPHON, where for the Russian language [4] a leading result of 94.4% accuracy on word inflexion in context was obtained.

Syntactic parsing was the focus of the Ru-Eval 2012 shared task [11]. The organizers conducted a survey of existing automatic approaches and resources and provided data in a conditional dependency format. In 2017, with the advent the Universal Dependencies (UD) initiative [6], shared tasks on multilingual parsing, including Russian, became possible, combining academic and industrial development systems under a common track. CoNLL shared tasks 2017–2018 [13], [14] has set the task of complete grammatical annotation, from raw text to syntax: for the Russian language, the quality of 92.48% accuracy LAS (labelled attachment score) on the materials of UD-SynTagRus and 72.24% accuracy LAS on the materials of social media was achieved. It became apparent that the quality of annotation should be evaluated on balanced datasets representing various styles and registers of writing.

In the above works, morphology and syntax are considered as independent tasks and evaluated separately; in most cases, systems that solve these problems are designed in such a way that they mark data independently at 3 levels—1) morphology 2) lemmatization 3) syntax, or at 2 independent levels—1) morphology and lemmatization and 2) syntax. Meanwhile, the relation among all three levels of the grammatical annotation is obvious: for example, an error in determining part of speech can lead to a lemmatization error and/or to an incorrect identification of syntactic relation.

We believe that the moment has come when the simultaneous intersection of the following factors allows us to create benchmark competitions in the general grammatical annotation of Russian texts, in which the overall level of annotation would be simultaneously assessed by various sources of Russian texts in all their diversity:

1. The development of deep learning, including universal language models, capable, according to some studies, of independently learning ideas about the semantic, lexical, and syntactic levels of the language [3],
2. Accumulation of big data from various sources,

3. The presence of a standardized format for morphological and syntactic annotation—UD 2.0.

The results of this initiative are presented in this article. Continuing the tradition of the previous independent shared tasks for Russian, we propose the new format of the NLP competitions—evaluating the joint models by their generalizing ability on the whole variety of language data of differing periods and sources.

2. Data

The data was provided in the UD 2.0, CONLL-U format, with respect to some variability in various data sources and their annotation methods, which will be described later.

The task of the GramEval 2020 organizers was to provide the most diverse training and test samples, taking into account the benefits of parsing quality improvements for industry, NLP research, digital humanities and theoretical linguistic research. For this reason, the main training sample with manual annotation included the most normative segments—news and fiction—as well as texts of social media, wiki, poetry, and texts of the 17th. century. Poetry was considered, since lexicon, morphemic and syntactic patterns, and word order is considered more variable in verses than in prosaic texts. As for the 17th. century data, the native speakers of Russian have almost no difficulties understanding such texts. Since it is assumed that modern processing systems are a closer match to human performance, it was interesting to take a diachronic look at the Russian NLP evaluation. An equally important factor was the availability of materials for all six registers in the UD format for training.

2.1. Training data

As training data, existing open datasets were collected from various sources: UD repository, MorphoRuEval, and RNC historical corpora collections.

- UD **SynTagRus** v2.5 (1.1M tokens, fiction, news, wiki, nonfiction). Annotation: automatic (ETAP3), human correction in native SynTagRus, then re-tokenized and converted automatically to UD 2.x. Enhanced dependencies removed. Since the treebank was not fully valid for the UD v2.5 scheme, a version with semi-manual corrections was also provided.
- UD Russian **GSD** v2.5 (96K tokens, wiki). Annotation: automatic (Google Stanford Dependencies) converted and manually checked.
- UD Russian **Taiga**: samples extracted from the Taiga Corpus and MorphoRuEval-2017 text collections (mostly social media and poetry, 39K tokens). Annotation: manual.
- **MorphoRuEval** test 2017: news (Lenta.ru, 5K tokens), fiction (magazines.gorky.media, 7K tokens) and social media (VK, 5K tokens). Morphological annotation done during the previous shared task annotations was manually changed to get better agreement with the current UD standards; syntax was annotated manually from scratch.

- **RNC 17th c.:** texts from the Middle Russian corpus (business & law, letters, Church Slavic, hybrid texts, 39K tokens). Annotation: manual, no lemmatization. In addition, 4M tokens were provided with manual morphological and automatic syntactic annotation.

2.2. Supporting data

Additional data were provided ‘as is’ with fully automatic annotation:

- Twitter: UDPipe pipeline (tokenization, morphology, syntax). Corpus of Russian tweets with sentiment annotation from <http://study.mokoron.com>.
- Wikipedia: UDPipe pipeline (tokenization, morphology, syntax). The actual dump of Russian Wikipedia, first 100,000 articles
- Comments from Russian Youtube Trends, April 2019. UDPipe pipeline (tokenization, morphology, syntax).
- Lenta.ru news: symbol unification + UDPipe pipeline (tokenization, morphology, syntax). Lenta Ru news, up to 2018.
- Stihi.ru poetry: symbol unification + UDPipe pipeline (tokenization, morphology, syntax).
- Proza.ru fiction: symbol unification + UDPipe pipeline (tokenization, morphology, syntax).
- Fiction Magazines (Taiga): UDPipe pipeline

2.3. Development and test data

The shared task included two stages: public and private test. At each stage, gold data was used, prepared specifically for the shared task. All in all, 7 annotators took part in data labeling at different linguistic levels. After that each sentence was verified by two supervisors. One of them, a contributor to Russian treebanks in the UD repository, checked through all data sets, for better consistency of the annotations. The size of the development and test sets is given in **Table 1**.

Table 1: The number of tokens and sources of the development and test sets

Register	Dev set	Dev source	Test set	Test source
news	1K	MorphoRuEval2017	1K	MorphoRuEval2017
social	1K	MorphoRuEval2017	1K	MorphoRuEval2017 + Taiga
wiki	1K	Russian GSD	1K	Wikipedia
fiction	1K	SynTagRus	1K	Taiga + RNC
poetry	1K	Taiga	1K	Taiga + RNC
17 cent	1K	Middle Russian-RNC	1K	Middle Russian-RNC

During the public test, the participants downloaded files for each of the 6 text registers, in a vertical format. The participants processed the input data with their systems and submitted the results to the leaderboard, obtaining detailed results for each source, error statistics, and could compare their result with the gold data.

During the private test phase, the participants were asked to download one large input file which included 10% test data for all six registers. The gold annotation was kept unavailable to the participants. The private test included news from UD MorphoRuEval2017, social media from UD MorphoRuEval2017 and Taiga, wiki from Wikipedia, fiction from Taiga and RNC, poetry from Taiga and RNC, 17th c. from UD MidRussian RNC.

3. Evaluation metrics

Evaluation procedure is based on the calculation of quality measures in the tasks of pos-tagging (qPos), morphological features tagging (qFeat), lemmatization (qLemma), and dependency parsing (qLas). The arithmetic mean of these values was used as participant's score (1) on a test set.

$$Score = mean(qPos, qFeat, qLemma, qLas) \quad (1)$$

Since separate test sets were created for each register, the composite participant's score (2) was calculated as an arithmetic mean for all registers.

$$Overall\ Score = mean(Score_{news}, Score_{wiki}, Score_{social}, Score_{fiction}, Score_{poetry}, Score_{17\ c.}) \quad (2)$$

$$Overall\ Score = mean(news\ score, wiki\ score, social\ score, fiction\ score, poetry\ score, 17\ c.\ score) \quad (3)$$

3.1. Pos-tagging, morphological features, lemmatization and syntax

Four main metrics—pos accuracy, other morphological features recall (macro-average over tokens), lemmatization accuracy and labeled attachment score (LAS) are measured the same way:

- Metrics are measured for each text source (news, poetry, etc), comparing participant submission results and gold markup:
 - Each predicted token annotation is being compared to the gold one:
 - Whether pos is the same as the gold one or not (POS: 1 or 0)
 - Sum all the matching features is divided by the number of the gold features (FEAT: continuous from 1 to 0)
 - Whether the lemma is the same as the gold one or not (LEMMA: 1 or 0)
 - Whether the syntactic head is the same, and if yes, is the relation correct (LAS: 1 or 0)
 - Sums of POS, FEAT, LEMMA, LAS points are being divided by the number of tokens in the text source—we get qPos, qFeat, qLemma, qLas quality
- All the quality on each source is being averaged (summed and divided by number of sources) to get overall quality.

Besides, when comparing lemmatization, letter capitalization and *e/ë* choice is not considered different. When evaluating LAS, full dependency relation (with tags after “:”) was considered.

3.2. Additional metrics

In order to achieve compatibility with the universal standard and the experience of international community, additional metrics not included in the leaderboard were calculated: F1 metrics for pos, features, and dependency relations, lemmatization, as well as UAS, MLAS, BLEX metrics according to the CoNLL method². These metrics were included to avoid situations in which systems could get high accuracy due to the rule-based evaluation hacking, for example, excessive addition of extra tags, etc., as well as for comparison with the results for the Russian language obtained in the previous shared tasks [13], [14].

3.3. Token alignment evaluation

For convenience of participants, the token alignment score was computed. It allows them to control whether the tokenization is corrupted or not. Every sentence from submission was compared to the corresponding gold one. Each token in a sentence was compared with the gold one. If the tokens were considered equal (see above), token alignment sum was incremented; sentence alignment score was a token alignment sum divided by the number of tokens. The participants were given the final alignment score of mean scores of every sentence.

During the private test phase, all systems had their alignment score of 100%.

4. System submission platform and routine

The competition was held on the CodaLab platform³, which allowed participants to choose the best parameters of their systems and analyze their performance on various text sources that were known in advance.

Each participant could make up to 100 submissions per day.

Participants were also allowed to use any external data for training their systems, including non-open sources. But the resulting system itself and the models should be open source and published on the Github.

Starting from 18 initial teams, 4 systems have reached the final test phase, one of which have provided two final submissions. The authors of the final systems represent different countries (Russia, France), tech companies (Yandex, ABBYY, MTS) and universities (CEA-LIST: Laboratory for Integration of Systems and Technology, Moscow State University, Moscow Institute of Physics and Technology).

² <https://universaldependencies.org/conll18/evaluation.html>

³ <https://competitions.codalab.org/competitions/22902>

5. Baseline system

5.1. rnnmorph + UDPipe

The starting point for the competition was a hybrid system assembled from the RNNMorph morphological analyzer [2] and the Parsito syntax module [10] (from UDPipe [9]). As of 2017, RNNMorph was a top-notch solution for the morphological analysis of the Russian language [8]. The choice of UDPipe is due to the popularity of this system and the positive experience of using it as a baseline system at CoNLL competitions [13], [14].

GramEval participants were given access to the source codes of the hybrid system with its default settings (without pretraining on the competition data). Besides, at each phase of the competition, the results of the baseline assessment were published.

5.2. Other milestones

In addition to the baseline system, we have trained and evaluated several well-known systems for morphological analysis and dependency parsing. We believe this is of help to the participants to better understand where their solutions fit in with other top ranked systems. This will also show how the quality of analysis has changed over recent years.

We selected MaltParser, SyntaxNet, UDPipe, StanfordNLP, TurkuNLP and rnnmorph for this evaluation. All of them were trained on the same training set. We followed the default settings, where possible, assuming that the developers had determined them in a rational way. Detailed information about the training setup can be found on the GramEval-2020 website⁴. Note that MaltParser, SyntaxNet and rnnmorph generate only part of the markup; these systems were evaluated within their competence. Note also that MaltParser itself cannot generate morphological features that it needs for parsing. Therefore, the morphological layers for MaltParser were generated by UDPipe.

6. Results

Table 2 presents the official leaderboard and **tables 3–6** detail the quality of morphological analysis, lemmatization, and parsing for each register, respectively. Besides that, the latter tables show the results of additional baseline systems (indicated by italics).

Table 2: GramEval official leaderboard—Overall score

System	Overall Score
qbic	0.91609
ADVance	0.90762
lima	0.87870
vocative	0.85198
baseline	0.80377

⁴ Default models and parameters for each module.

Table 3: Scores: parts of speech

	fiction	news	poetry	social	wiki	17 cent
qbic	0.980	0.966	0.969	0.947	0.927	0.963
ADVance	0.980	0.965	0.960	0.937	0.921	0.960
lima	0.976	0.971	0.957	0.937	0.925	0.935
vocative	0.975	0.965	0.929	0.917	0.909	0.870
<i>Turku</i>	0.970	0.964	0.951	0.926	0.902	0.870
<i>Stanford</i>	0.974	0.964	0.944	0.913	0.924	0.896
<i>UDPipe</i>	0.975	0.967	0.927	0.916	0.906	0.868
<i>SyntaxNet</i>	0.953	0.952	0.906	0.884	0.904	0.866
<i>rnnmorph</i>	0.970	0.949	0.946	0.928	0.922	0.894

Table 4: Scores: grammatical features

	fiction	news	poetry	social	wiki	17 cent
qbic	0.987	0.981	0.967	0.947	0.944	0.929
ADVance	0.986	0.981	0.960	0.959	0.928	0.929
lima	0.979	0.966	0.956	0.953	0.967	0.896
vocative	0.948	0.944	0.898	0.900	0.904	0.793
<i>Turku</i>	0.952	0.962	0.921	0.918	0.921	0.831
<i>Stanford</i>	0.949	0.957	0.914	0.904	0.923	0.841
<i>UDPipe</i>	0.946	0.946	0.899	0.899	0.902	0.791
<i>SyntaxNet</i>	0.934	0.926	0.886	0.887	0.872	0.801
<i>rnnmorph</i>	0.878	0.858	0.857	0.852	0.838	0.825

Table 5: Scores: lemmatization

	fiction	news	poetry	social	wiki	17 cent
qbic	0.980	0.982	0.953	0.960	0.936	0.783
ADVance	0.977	0.981	0.952	0.954	0.922	0.797
lima	0.937	0.950	0.913	0.953	0.923	0.610
vocative	0.961	0.955	0.939	0.955	0.915	0.582
<i>Turku</i>	0.974	0.976	0.949	0.956	0.928	0.584
<i>Stanford</i>	0.973	0.959	0.926	0.952	0.922	0.571
<i>UDPipe</i>	0.963	0.957	0.912	0.941	0.934	0.579
<i>rnnmorph</i>	0.950	0.907	0.918	0.928	0.904	0.588
<i>rnnmorph</i>	0.878	0.858	0.857	0.852	0.838	0.825

Table 6: Scores: LAS

	fiction	news	poetry	social	wiki	17 cent
qbic	0.896	0.912	0.814	0.807	0.781	0.665
ADVance	0.869	0.911	0.780	0.784	0.760	0.618
lima	0.850	0.843	0.725	0.713	0.697	0.546
vocative	0.826	0.834	0.660	0.659	0.694	0.500
<i>Turku</i>	0.859	0.877	0.731	0.733	0.711	0.502
<i>Stanford</i>	0.854	0.873	0.709	0.706	0.703	0.509
<i>UDPipe</i>	0.811	0.817	0.666	0.644	0.668	0.462
<i>SyntaxNet</i>	0.808	0.802	0.6	0.614	0.645	0.446
<i>MaltParser</i>	0.599	0.553	0.404	0.476	0.436	0.340

All final submissions outperformed the baseline approach in morphology, lemmatization, and parsing. As for the systems’ performance by register, three groups can be distinguished:

- fiction and news are the easiest to process (in general, > 95% in pos, features, and lemmas, > 83% in syntax);
- social media, poetry, and wiki texts are more challenging to process (in general, -2% to -7% drop at the lexico-grammatical levels and more significant drop at the level of syntax, see below);
- the performance on the diachronic 17th c. data is low, especially at the level of syntax. Note, however, that the top-2 systems achieve the 96% quality in post-tagging and the 93% quality in feature tagging.

As expected, the performance drop is more pronounced in syntax, features, and lemma processing.

In the task of parsing, for all systems there is a significant (> 8%) difference in the quality of analysis on fiction + news, on the one hand, and poetry + social + wiki, on the other hand, see [Table 6](#) and [Figure 1](#). However, only two participants managed to surpass the best of additional baseline systems in all registers.

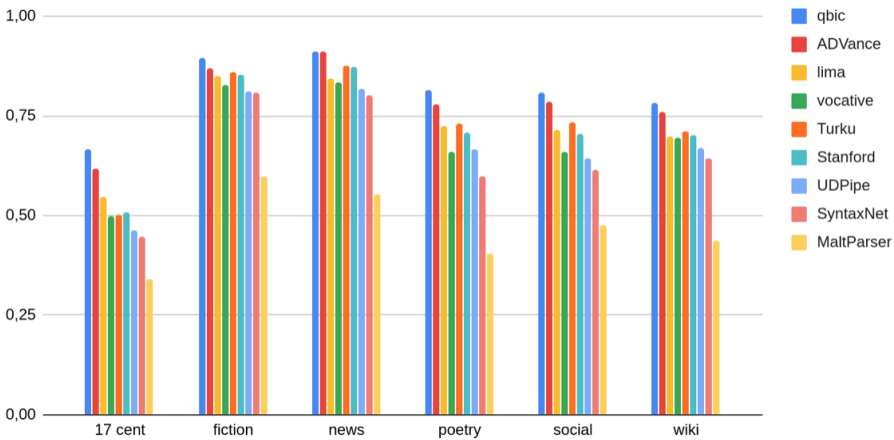


Figure 1: LAS across registers

7. Approaches adopted by the systems

The participants’ approaches represent a fairly wide variety of modern neural network approaches—universal BERT transformers, recurrent neural architectures—LSTM and CRF-LSTM, feedforward layers, word and char embedding sources—BERT, word2vec and fasttext—therefore, we can say that they are quite indicative from the point of view of the current level of technology [15]–[18].

Despite the general statement of the problem, some efforts were also spent on fitting the scores of systems on specific data—two participants use rule-based

approaches to adapt the outputs of the systems for 17th c. data, and also use classifiers to detect the most outlying training sources—social media, poetry and historical data. The resulting architectures are described in [Table 7](#).

It is noteworthy that the highest quality was shown by a system (qbic) annotating morphology, syntax, and lemmatization independently. Qualitative differences in the systems’ performance are discussed in the next section.

Table 7: Table 7: Architectures of the GramEval 2020 participating systems

Team	Data	Architecture	Embeddings
qbic (1)	All GramEval data except SynTagRus	End-to-End parser: features, lemmas, and dependencies are predicted by joint BERT model with independent modules. Encoder is a single-layer LSTM, decoders are simple feedforward models for predicting lemmas and features, as well as a biaffine attention model for dependencies and their labels	Pretrained RuBERT
ADVance (2)	All GramEval data + poetry Taiga corpus for embedding training	Classifier of 4 main data sources—normative fiction, 17 c., poetry, social. + Morphotagger and parser on BERT, pretrained on SynTagRus 2.5 + 17 c. lemmer on rules	4 separately trained BERTs on GramEval data
lima (3)	All GramEval data	Original implementation of Dozat & Manning: embedding layer + LSTM layer + feedforward layer. Differs from the original models in that morphology and syntax are trained simultaneously in multitask learning mode	Pretrained FastText
vocative (4)	GramEval2020 data with rule-based parser validation for extracting good training samples for pos-tagging and parsing. + clean GramEval data for UDPipe training + own treebank data for pos tagging training	Ensemble model: 1) dictionary-based lemmatizer 2) LSTM-CRF pos tagger, considering the context and features + pure CRF pos tagger for sentences longer than 30 words + Russian UDPipe for pos and features 3) parser: UDPipe trained on GramEval data 4) Rule-based correction for 17 c. data	Pretrained word2vec wordchar2vector

8. Analysis of submitted annotations

Table 8 and 9 outline the systems’ agreement in full morphology and dependency markup, respectively. In [Table 8](#), Accuracy / Cohen’s kappa for the combination of pos

and feature tags are shown. To calculate Cohen’s kappa, the list of categories was determined on the basis of all the observed responses. In **Table 9**, Accuracy is shown calculated for the combination of syntactic head and dependency relation. Both relation types and subtypes (labels before and after the colon) were considered.

Table 8: Agreement in pos and feature markup

	qbic	ADVance	lima	vocative
gold	0.874 / 0.867	0.803 / 0.791	0.811 / 0.801	0.765 / 0.752
qbic		0.844 / 0.835	0.830 / 0.820	0.781 / 0.769
ADVance			0.767 / 0.754	0.722 / 0.706
lima				0.784 / 0.771

Table 9: Agreement in the syntactic head and dependency relation markup

	qbic	ADVance	lima	vocative
GOLD	0.813	0.788	0.729	0.697
qbic		0.826	0.769	0.720
ADVance			0.765	0.711
lima				0.706

One can note a greater cross-system agreement than that between systems and the gold markup. Indirectly, this suggests that systems make similar mistakes.

The output data of the competing systems show that the errors in morphological analysis are mostly the same as in previous competitions for the Russian language. The errors in lemmatization, pos-tagging and morphological features most often correlate.

In quantitative terms, most errors are associated with uppercase uses and non-standard spellings. Erroneous pos-tagging and morphological features arise in all the outputs at the beginning of the sentence, at the beginning of the line in the poetry, in proper names that share ambiguity with common nouns (*Наука, Тузр*). Furthermore, the competing systems encounter difficulties while analyzing words with spelling errors, author spelling, hashtags (typical for social networks), abbreviations and acronyms, for example, in Wikipedia references.

In lemmatization, systems found it difficult to analyze words with a rare inflectional model (*распростертый, огороженный, объемлет, ищет, горю* etc.) and pluralia tantum nouns or plural homonyms (*ножницы, окова-оковы, мозг-мозги*). Difficulties in resolving homonymy remain in pos-tagging, cf. *быть* VERB vs AUX, *что* PRON vs. SCONJ vs. PART vs. ADV, *и* CCONJ vs PART, ADV vs ADJ, DET vs PRON, uses of words like *мунa* (NOUN vs ADP vs PART), *походу* (NOUN vs ADV), *смотря, значит* (VERB vs ADV). Such errors may also be triggered by low quality markup in training sets since frequent homonyms are difficult to be spot-checked manually. In addition, from the point of view of linguistic theory and existing corpus practices, there are well known cases that can be approached differently and thus tagged inconsistently in various training data. These include:

- participles vs. verbal adjectives (*волнующий, образованный, греющий, начитан, обязан*, etc.),
- words such as *нельзя, надо, пора* tagged as VERB vs ADV vs NOUN,
- inconsistency in lemmatizing the nouns ending with *-ие/-ье* (e.g. *безумие—безумье*) and adjectives ending with *-ой, -ый* (*грунтовый—грунтовой*).

In general, we observe that the systems do well with the morphological feature ambiguity. There are relatively few errors due to the paradigm syncretism (eg. *события* Case=Acc Number=Plus vs. *события* Case=Gen Number=Sing). The most common errors in morphological features are as follows:

- **animacy** in adjectives, pronouns and numerals (systems add **animacy** not only for those word forms where difficulties in analysis are possible); this feature is often mistakenly identified if the word is uppercase and / or comes first in the sentence;
- features of the verb *быть* (all systems add **aspect**);
- **gender**: the competing systems attribute **gender** to adjectives in the plural, make mistakes in determining the gender of proper names, there are also errors with the gender of common nouns in indirect cases (even though the lemma is defined correctly);
- **case**: some systems systematically add case to short adjectives and participles in the predicative position;
- **aspect**. Errors arise in biaspectual verbs (*подвизается, мигует*, etc.);
- **voice**: all systems mark finite verb forms ended with *-ся* as passive (Pass) rather than middle (Mid);
- **degree** in adjectives and adverbs: the participants often do not tag superlative (Sup) and comparative (Cmp) degree.

Table 10 presents top-20 mismatches in the dependency relation labeling⁵, with occurrences (N) calculated over all systems. For the most part, these are mismatches between flat syntactic relations, clause and phrase joining relations, verb-argument relations, and modifier relations mixed with either other modifiers or argument relations. The common source of errors is register-specific tokens and constructions such as ‘=’, ‘*’, ‘***’, ‘" punctuation marks in wiki and social media texts, attachment of interjunctions (**discourse**) incorrectly predicted as **parataxis**, the **list** relations common to the wiki biographies also predicted as **parataxis**. It can be seen that the best system is accurate in predicting the **punct** relations, which is problematic to the other three systems. At the same time, it is low-sensitive to the **fixed**, **discourse**, **parataxis**, **vocative**, **compound**, and **list** relations.

All systems tend to mix indirect objects (**iobj**) in the instrumental case with oblique (**obl**). This and a large number of other core argument relation mismatches can probably be attributed to argument relations incorrectly represented in UD-Syntagrus, the main training dataset for Russian parsers. The alluvial confusion plot (**Figure 2**) demonstrates that in many cases, the systems make errors in argument relations

⁵ For the definition of relations see the UD site <https://universaldependencies.org/u/dep/index.html>.

differently. However, we find that, relative to the 2012 shared task, the core argument labeling has improved significantly and such errors are rare in all systems.

The 17th c. dataset, on which all systems demonstrated low scores, require separate attention. In these texts, archaic endings and orthography are main factors that noticeably affect the system’s performance in pos and lemma analysis. No more than one system was able to analyse words such as *такова, всяково, теперь, кроме, окяно, козною* for pos and *двесте, детеи, земнаго, плаваюча, звер, итить, шездесят* for lemmas correctly. Register-specific training efforts were justified when analysing the close-class words such as *аз, and который*. Unlike in other registers, nominal phrases were challenging for parsing due to time-specific syntactic patterns in named entities (e.g. mixed genitive-possessive construction in *на Романове отца их поместье*, compound in *Иль мурзюю*) and the long chains of genitive groups with inverse word order.

Table 10: Most frequent mismatches in dependency relations

N	gold	predicted	N	gold	predicted
74	punct	discourse	19	obj	nsubj
49	parataxis	appos	18	amod	nummod
42	iobj	obl	18	punct	parataxis
40	list	parataxis	17	obj	obl
38	parataxis	conj	16	conj	parataxis
32	discourse	parataxis	15	appos	nmod
24	amod	appos	15	discourse	advmod
24	nmod	appos	15	mark	advmod
23	obl	nmod	15	xcomp	obl
20	nsubj	obj	14	appos	parataxis

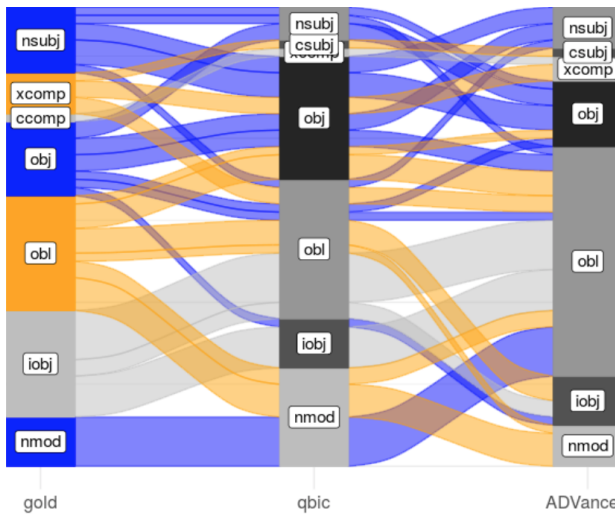


Figure 2: Core argument and nominal modifier relations incorrectly predicted by either of top-2 systems

9. Conclusion

During GramEval shared task we have introduced a new approach to full morphology and dependency parsing evaluation for Russian:

- testing and training procedures were carried out on greater variability of text sources—considering temporary, stylistic and genre variation
- the public and private test phases were organized on an open platform, expanding the capabilities of participants and allowing them to become more familiar with the overall performance of their systems on different data;
- new training data was prepared, both with automatic annotation and with both automatic and expert assessment of the data;
- the competition guidelines provide compatibility with the UD standard, as well as at the level of additional metrics—compatibility with the CoNLL competitions;
- as the result of the competition, a comparison of different parsing strategies was obtained, and a new state-of-the-art method for full Russian morphological parsing of Russian.

The competition leaderboard is now permanent at the CodaLab, and we welcome researchers and developers to submit their systems to the leaderboard and compare their results with other approaches.

All materials of GramEval 2020 including supplementary tables and figures for this paper are available at the shared task repository⁶. As the collection represents the vast variety of genres, registers, corpora, annotation practices, with new development and test data checked manually, we hope that the output GramEval 2020 will stay practical and relevant for the NLP community.

10. Acknowledgements

The authors would like to thank the participants of GramEval 2020 for helpful comments and valuable suggestions. We are grateful to all members of the community who contributed to the data preparation and evaluation, scripts, and setting up the GramEval 2020 CodaLab page, and especially to Vera Davydova, Aleksey Dorkin, Maria Ermolova, Aleksandra Konovalova, Kristina Litvintseva, Elizaveta Nosova, Anna Safaryan, Dmitry Sichinava, and Elena Suleymanova.

⁶ <https://github.com/dialogue-evaluation/GramEval2020>

References

1. *Anastasyev D.* (2020) Exploring pretrained models for joint morpho-syntactic parsing for Russian. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2020*, Vol. 19.
2. *Anastasyev D., Gusev I., Indenbom E.* (2018) Improving part-of-speech tagging via multi-task learning and character-level word representations. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2018*, Vol. 17, pp. 14–27.
3. *Bocharov V. V., de Chalendar G.* (2020) The Russian language pipeline in the LIMA multilingual analyzer. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2020*, Vol. 19.
4. *Clark K., Khandelwal U., Levy O., and Manning C. D.* (2019) What Does BERT Look At? An Analysis of BERT's Attention. In: arXiv:1906.04341.
5. *Cotterell R., Kirov Ch., Sylak-Glassman J., et al.* (2018) The CoNLL–SIGMORPHON 2018 shared task: Universal morphological inflection. In *Proceedings of CoNLL–SIGMORPHON 2018*.
6. *Dereza O. V., Kayutenko D. A., Fenogenova A. S.* (2016) Automatic morphological analysis for Russian: A comparative study. In *Proceedings of the International Conference Dialogue 2016. Computational linguistics and intellectual technologies. Student session (online publication)*. Retrieved from: <http://www.dialog-21.ru/media/3473/dereza.pdf>.
7. *Grashchenkov P. V., Koziev I.* (2020) POS-tagger Dataset Augmentation by Ensemble Parsers. Unpublished Ms. Moscow.
8. *Lyashevskaya O., Astafeva I., Bonch-Osmolovskaya A., et al.* (2010) NLP evaluation: Russian morphological parsers [Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka]. In: *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2010*, Vol. 9 (16), pp. 318–326.
9. *Nivre J., De Marneffe M. C., Ginter F., et al.* (2016) Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666.
10. *Segalovich I.* (2003) A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, In *Proceedings of MLMTA-2003*, Las Vegas, Nevada, USA.
11. *Sokirko A. V.* (2001) *Semantic Dictionaries in the Natural Language Processing: Based on the DIALING system [Semanticheskie slovari v avtomaticheskoy obrabotke teksta: Po materialam sistemy DIALING]*. Cand. Tech. Sc. Dissertation. Moscow.
12. *Sorokin A., Shavrina T., Lyashevskaya O., et al.* (2017) MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2017*. Vol. 16–1, pp. 297–313.
13. *Sorokin A. A., Smurov I., Kirjanov D.* (2020) Tagging and parsing of multidomain collections. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2020*, Vol. 19.

14. *Straka, M., Hajič, J., Straková, J., and Hajič jr., J. (2015) Parsing universal dependency treebanks using neural networks and search-based oracle. In Proceedings of 14th International Workshop on Treebanks and Linguistic Theories (TLT 2015), Warszawa, Poland.*
15. *Straka M., Hajič J., Straková J. (2016) UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.*
16. *Toldova S., Sokolova E., Astafiyeva I., et al. (2012) NLP evaluation 2011–2012: Russian syntactic parsers [Ocenka metodov avtomaticheskogo analiza teksta 2011–2012: Sintaksicheskie parsery russkogo jazyka]. In Computational linguistics and intellectual technologies. Proceedings of International Conference Dialogue 2012. Vol. 11 (18), pp. 797–809.*
17. *Zaliznyak, A. A. (1977) Grammatical Dictionary of Russian [Grammaticheskij slovar' russkogo jazyka]. Moscow.*
18. *Zeman D., Hajič J., Popel M., et al. (2018) CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 1–21.*
19. *Zeman D., Popel M., Straka M., et al. (2017) Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1–19, Vancouver, Canada.*

Appendix

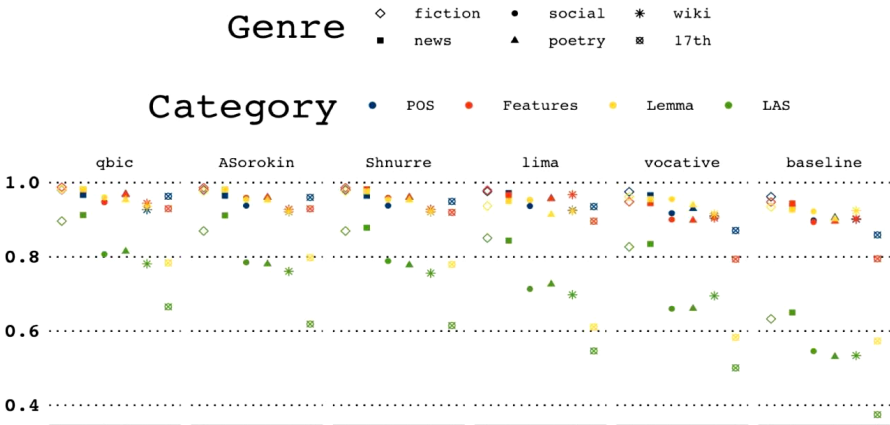


Figure 3: Systems' scores by register

SUMMARY CONSTRUCTION STRATEGIES FOR HEADLINE GENERATION IN THE RUSSIAN

Malykh V. (valentin.malykh@phystech.edu)

Kazan Federal University, Kazan, Russia

Cherniavskii D., Valukov A.

Moscow Institute of Physics and Technology, Moscow, Russia

Key words: headline generation, summarization, Russian

DOI: 10.28995/2075-7182-2020-19-570-578

СТРАТЕГИИ СОСТАВЛЕНИЯ РЕФЕРАТОВ В ЗАДАЧЕ ГЕНЕРАЦИИ ЗАГОЛОВКА ДЛЯ РУССКОГО ЯЗЫКА

Малых В. (valentin.malykh@phystech.edu)

Казанский (Приволжский федеральный
университет, Казань, Россия

Чернявский Д., Валюков А.

Московский физико-технический институт
(научно-исследовательский университет), Москва, Россия

1. Introduction

In the modern world, texts are plenty in the everyday life of a person—the news articles, blogs, social networks. These texts could be long, for example, the typical length of a New York Times news article is more than 700 words [13]. The reading process could take significant time for even one article, so this raises a question of shortening this time. To handle the mentioned issue there were proposed techniques of extractive and later abstractive text summarization, i.e. the generation of a short text summary using longer original text.

There is an issue with most of abstractive and some of extractive summary generation strategies, they all need a training set, which could take time and labour to create, like CNN/DailyMail dataset initially presented in [5] and compiled for

text summarization task in [10]. To overcome this issue there was presented a separate task of headline generation for news documents. Since the news documents are plenty, and they could be used with ease.

The headline generation task could be considered as a two-stage task. On the first stage, a summary of the article body is constructed and on the second stage, the headline is generated using the constructed summary. In this work, we concentrated on a headline generation task for the Russian language in an aspect of comparison summary construction techniques.

This work is composed as follows: related work, dataset and metrics description, base models description, summary strategies, experiments, and conclusion.

2. Related Work

There were already successive attempts in headline generation for different languages. For the English language, there are several works. The authors of [12] were to the best of our knowledge the first to apply neural networks to headline generation. In more recent work of Hayashi et al. [4], an encoder-decoder approach was presented, where the first sentence was reformulated to a headline. The related approach was presented in [11], where the approach of the first sentence was expanded with a so-called topic sentence. The topic sentence is chosen to be the first sentence containing the most important information from a news article (so-called 5W1H information, where 5W1H stands for who, what, where, when, why, how). This approach has a limitation that these sentences should be marked up beforehand. Tan et al. [18] present an encoder-decoder approach based on a pre-generated summary of the article. The summary is generated using a statistical summarization approach.

For the Russian language, there are a few works on this topic. In the work of [2] there were presented universal Transformer model, which used whole article body as input to generate a headline. The other works [3], [14], [16], which were resulted the shared task on headline generation, described in [8]. Sokolov and Stepanov in [14], [16] have used copy mechanism in encoder-decoder models to improve quality of the generation, while Gusev in [3] invoked phrase-based attention mechanism to improve the Transformer model itself. It need to be mentioned that all the previous works were using whole article body to generate headline hypotheses.

3. Dataset & Metrics

In this work, we decided to explore different summary construction strategies for Russian language dataset. There is only one dataset of significant size for Russian. It is “Rossiya Segodnya” News Dataset described in [2]. There are 1,003,869 news articles in the corpus with a mean title length 9.5 words, mean text length of 315.6 words, and mean 15.0 sentences. Following [2] we divided the dataset into three parts: 10,000 news documents were withheld as a validation set, 20,000 ones as a test set, and the rest was considered as a train set.

We are using ROUGE metric family, presented in [7]. Essentially, the ROUGE metric is counting common token sequences in ground truth and hypothesis sequences. There are three main variants: ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 and ROUGE-2 are using unigrams and bigrams respectively to compute a score. ROUGE-L is using longest common subsequence for a reference and a hypothesis to compute the score. Here are the formulae for ROUGE metrics from original paper [7]:

$$\text{ROUGE-N} = \frac{\sum_{r \in \{\text{references}\}} \sum_{w \in r} \text{Match}(w)}{\sum_{r \in \{\text{references}\}} \sum_{w \in r} \text{Count}(w)}, \quad (1)$$

where N stands for the length of a n-gram w , **Match** is the maximum number of n-grams co-occurring in a candidate summary (hypothesis) and in a set of reference summaries, and **Count** is a number of all n-grams in references' set.

In particular, the ROUGE-N formulae mentioned above are describing how much the hypothesis is capturing the reference summary and is often referred as the recall variant of ROUGE-N metrics, or simply ROUGE-N-Recall. As there are no control over the length of the hypothesis, so it can capture almost all of the reference summary while being excessively long. This issue is solved by the precision modification of ROUGE-N metrics that has the same formulae but the **Count** variable is now referred to the number of all n-grams in hypothesis' set. The ROUGE-N-F1 score is calculated as classical F_1 measure with ROUGE-N-Precision and ROUGE-N-Recall using harmonic mean.

In addition to ROUGE, we decided to use an extraction score, presented in [1]. The extraction score is a metric of extractiveness of a summary. It searches for the long substrings from a source text in the summary. Extraction score is defined as follows:

$$\text{ext_score}(S) = \sum_{s \in P(ACS_s)} s \times \left(e^{s-1} - \frac{1-s}{e} \right), \quad (2)$$

where S is a summary, a ACS_s is the set of all long non-overlapping common sequences between S and the document, $P(ACS_s)$ is a set, where each element is the length of a common sequence divided by the length of the summary.

4. Base Models

We have conducted experiments with two basic models which follow Encoder-Decoder approach presented in [17]. One is a recurrent neural network and another is Transformer network, described in [19], which are described below in more details.

4.1. Recurrent Model

The Encoder and the Decoder are both complied of two-layer bidirectional Long Short-Term Memory (LSTM) [6] recurrent networks. The encoder network recursively receives news body words (in both direction) as input and produces hidden states, one for each input word (since there are two reading directions, there are two

hidden states for each word, these hidden states are concatenated to produce a whole hidden state for a word). Afterwards, it passes its final hidden state as the initial hidden state to the decoder network, and for each decoder step, an attention distribution is calculated upon the encoder hidden states. This distribution is used to predict the next word of the news headline. At the beginning of prediction, the decoder network receives a special <START> token, and later it uses the previously generated word as an input. This procedure is illustrated in Fig. 1.

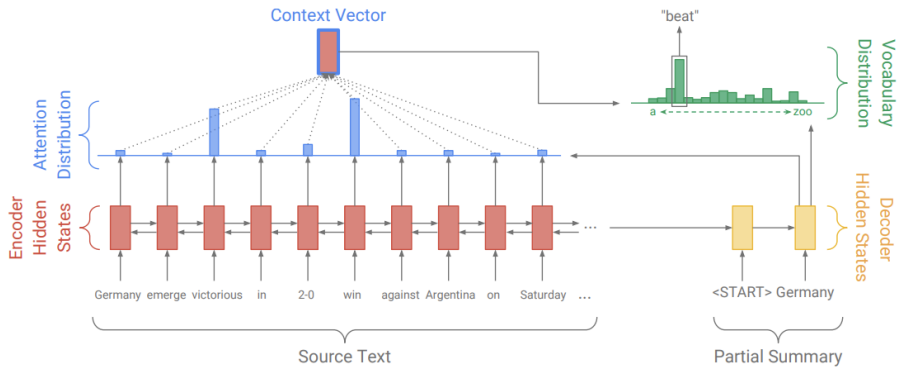


Figure 1: Seq2seq with attention from [14]

4.2. Transformer

In this case, the network receives the news body text not recursively but as a whole. The encoder and the decoder network both consist of 6 identical layers, described below. Firstly, the positional encoding is added to the input embeddings to ensure the difference of embeddings in a different part of the text. Then the multi-head self-attention is calculated upon the input. This attention is then added to the input of the multi-head module and a layer normalization is applied. After it is passed to the Feed Forward network. Final encoder layer output is then passed to each of layer the decoder network to the encoder-decoder attention block, which comes after the self-attention block. Complete architecture is shown in Fig. 2.

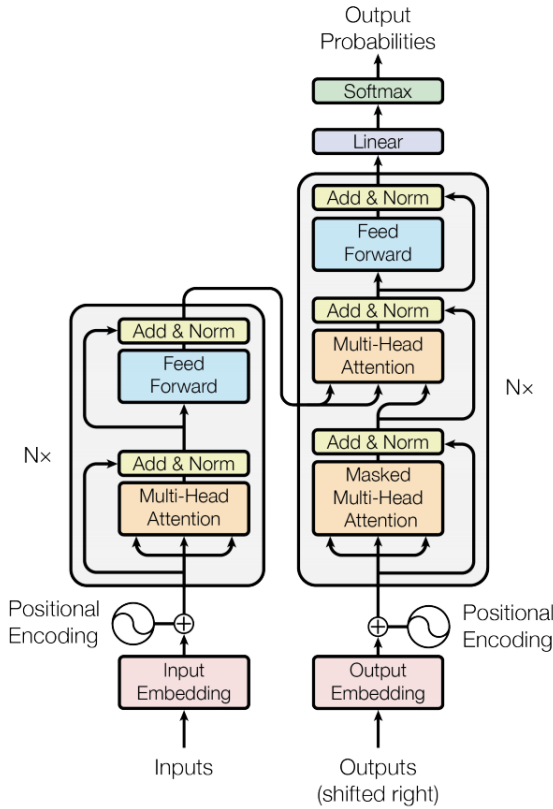


Figure 2: Transformer architecture from [19]

5. Summary Construction Strategies

In this section, we describe different strategies of summary construction for headline generation. The constructed summaries are used as input for encoder-decoder models.

Whole body. This is the simplest strategy, i.e. to use the whole article body text to generate a headline. This approach is the most common in the literature, e.g. [12]. This approach mentioned as “full” in **Tab. 1**.

First Sentence. The first sentence could be a strong hypothesis for the headline, as shown in [2]. We should mention that in “Rossiya Segodnya” news documents the first sentence is a formal statement of date and place, for example:

МОСКВА, 21 августа 2015.

We skip this formal intro and use the next informative sentence as the first sentence in our setup. This approach mentioned as “1sent” in **Tab. 1**.

Three First Sentences. The first sentence baseline although showed good performance could not contain enough information for headline generation. For example, in paper [18] authors stated that the first sentence is not informative enough. So we decided to use the first three sentences as a summary. Again, the formal intro is skipped in this setup. This approach mentioned as “3sent” in Tab. 1.

Unsupervised Summary. Work [18] showed that unsupervised summary could be a good hypothesis for a headline. In our work we have used classic TextRank algorithm, described in publication [9], to generate a summary from an article body. This approach mentioned as “*unsup*” in Tab. 1.

NER Summary. We propose a novel approach to construct a summary using named entity recognition (NER). We use PullEnti pre-trained model [15] to mark up the “Rossiya Segodnya” corpus. The mark up contains mentions of persons, organizations and locations. To construct the summary we extracted the sentences which contain at least one named entity. This approach mentioned as “*NER*” in Tab. 1.

Table 1: ROUGE-1,2,L scores for Recall (r) and F-measure (f) variations; also extraction score¹

Model Metric	R-1-f	R-1-r	R-2-f	R-2-r	R-L-f	R-L-r	ext. score
1sent	23.395	44.055	10.302	20.716	16.291	40.390	0.427
3sent	15.235	53.039	5.836	24.089	8.698	49.656	0.477
unsup	14.095	48.003	5.110	20.286	8.507	44.772	0.367
NER	12.499	36.168	4.124	13.797	7.797	33.362	0.241
Seq2seq+1sent	39.866	38.671	23.111	22.480	37.058	36.758	0.551
Seq2seq+3sent	42.545	41.584	25.131	24.668	39.613	39.539	0.627
Seq2seq+full	41.927	40.641	24.639	23.944	39.002	38.663	0.582
Seq2seq+unsup	36.147	35.093	19.643	19.134	33.448	33.223	0.425
Seq2seq+NER	25.556	24.104	13.142	12.547	23.287	22.884	0.269
Transformer+1sent	41.075	40.557	24.593	24.372	38.319	38.488	0.719
Transformer+3sent	42.922	41.863	25.476	24.908	39.996	39.784	0.673
Transformer+full	39.627	37.945	21.153	20.328	36.525	35.852	0.423
Transformer+unsup	34.090	32.764	17.583	16.967	31.422	30.936	0.363
Transformer+NER	28.501	27.688	14.705	14.387	26.298	26.142	0.379
Gavrilov et al. [2]	39.75	37.62	22.15	21.04	36.81	35.91	—
Sokolov [14]	42.96	—	25.43	—	40.02	—	—
Stepanov [16]	25.23	25.79	10.33	10.60	22.82	24.08	—
Gusev [3]	41.61	40.33	24.46	23.76	38.85	38.51	—

¹ For the ROUGE scores higher is better, for extraction score lower is better. **Bold** marks up the best result, while *Italic-Bold* marks up the second best result. All the metrics are computed on test set.

6. Experiments

As a Transformer model we have trained a basic 6-layer Transformer architecture model with 8 heads. The dimension of fully connected layer was 2,048. We trained it with a batch size of 4,096 for 110k training steps. For Seq2seq model we took a default 2-layer LSTM with 500 hidden units on both of the encoder and decoder. To train these models we used single nVIDIA Titan X (Pascal) GPU with 12Gb of RAM.

6.1. Results

The evaluation scores for both architectures and all summary construction summaries are presented in **Tab. 1**. The best results by recall in ROUGE-1 and ROUGE-L are showed by the 3sent baseline. This fact could be considered trivial. But interestingly, the Transformer approach over 3sent summary achieves best results by ROUGE-2 metric, including recall score. In addition this model shows the second best results in ROUGE-1 and ROUGE-L both recall and F-measure. We interpret this as Transformer is being very extractive—its variants are consistently more extractive than other two approaches, and 3sent baseline has the highest scores for the recall, so Transformer has a better choice to copy from the input text. Interestingly, the Transformer model achieves the best performance in terms of ROUGE-2-Recall, even better than 3sent baseline. We also could draw the reader’s attention to the fact that Transformer models lower extraction score with an extension of input text from one first sentence to the whole text. While Seq2seq models do not follow this regularity.

Regarding the extraction score, the lowest one is demonstrated by NER baseline, but this also accompanied with the lowest ROUGE metrics. The Seq2seq approach over NER basic summary drastically improves the ROUGE results, but also gain some extraction score, showing the second best one. Interestingly, a Transformer model has a much higher extraction score with this summary as input.

Table 2: Samples of headlines generated by the studied models.

We present only unique generated headlines

Original text, truncated:	пожар, произошедший в среду в ресторане в центре москвы, ликвидирован, пострадавших нет, сообщил риа новости источник в правоохранительных органах столицы. «пожар в ресторане „эль гаучо“ на садовой-триумфальной улице в двухэтажном здании ликвидирован. по предварительным данным, горели жировые отложения в вентиляции. возгорание произошло в вентиляционной системе», — сказал собеседник агентства. в настоящее время причины пожара устанавливаются. по данным представителя мчс, сообщение о пожаре поступило на пульт дежурного «01» в 21.25 мск. он отметил, что, благодаря своевременной эвакуации, никто из посетителей и сотрудников ресторана не пострадал.
Original headline:	пожар в ресторане в центре москвы ликвидирован, никто не пострадал
Transformer+1sent:	пожар в ресторане в центре москвы потушен
Transformer+3sent:	пожар в ресторане в центре москвы ликвидирован, пострадавших нет
Transformer+full:	пожар в ресторане в центре москвы ликвидирован
Transformer+ner:	пожар в центре москвы потушен

Some samples of headlines produced by different models are presented in **Tab. 2**. As one could see, the approaches are differ with details, and quality of a headline is seemingly correlated with BLEU score, for example, the named entity recognition approach for summary constructing (the worst one by BLEU score) suffers from lack of useful words, such as «потушен» or «ликвидирован».

7. Conclusion

We have presented a comparison of summary construction strategies, where the constructed summaries are used as input for headline generation. We have studied the classic first sentence strategy and extended it to the three first sentences one. The latter strategy shows the best performance by the means of recall itself and also gives a boost for Transformer architecture model which achieves new state of the art ROUGE-2 results. This model outperforms other approaches even those which are using whole text as input. The Seq2seq models are consistently gaining lower ROUGE scores in all variants in comparable setups, although they have lower extraction score also.

As the direction for future research authors see two main ones: an application of the proposed approach to other languages, and its modification for abstractive summarization task itself, which has significantly different text structure and so states open question of applicability of the proposed approach.

7.1. Acknowledgements

The work of the first author was funded by RFBR, project number 19-37-60027.

References

1. *Cibils, A. et al.*: Diverse beam search for increased novelty in abstractive summarization. CoRR. abs/1802.01457, (2018).
2. *Gavrilov, D. et al.*: Self-attentive model for headline generation. 41st European Conference on Information Retrieval. (2019).
3. *Gusev, I.*: Importance of copying mechanism for news headline generation. In: Computational linguistics and intellectual technologies. (2019).
4. *Hayashi, Y., Yanagimoto, H.*: Headline generation with recurrent neural network. In: New trends in e-service and smart computing. pp. 81–96 Springer (2018).
5. *Hermann, K. M. et al.*: Teaching machines to read and comprehend. In: Advances in neural information processing systems. pp. 1693–1701 (2015).
6. *Hochreiter, S., Schmidhuber, J.*: Long short-term memory. Neural computation. 9, 8, 1735–1780 (1997).
7. *Lin, C.-Y.*: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out. (2004).
8. *Malykh, V., Kalaidin, P.*: Headline generation shared task on Dialogue’2019. In: Proceedings of the international conference “Dialogue 2019”. (2019).

9. *Mihalcea, R., Tarau, P.*: Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing. (2004).
10. *Nallapati, R. et al.*: Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Proceedings of the 20th SIGNLL conference on computational natural language learning. pp. 280–290 Association for Computational Linguistics, Berlin, Germany (2016).
11. *Putra, J. W. G. et al.*: Experiment on using topic sentence for neural news headline generation. In: Proceedings of 24th annual conference of japanese association for natural language processing. (2018).
12. *Rush, A. M. et al.*: A neural attention model for abstractive sentence summarization. In: Empirical methods in natural language processing. pp. 379–389 (2015).
13. *Sandhaus, E.*: The new york times annotated corpus ldc2008t19. DVD. Linguistic Data Consortium, Philadelphia (2008).
14. *Sokolov, A.*: Phrase-based attentional transformer for headline generation. In: Computational linguistics and intellectual technologies. (2019).
15. *Starostin, A. et al.*: Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. (2016).
16. *Stepanov, M.*: News headline generation using stems, lemmas and grammemes. In: Computational linguistics and intellectual technologies. (2019).
17. *Sutskever, I. et al.*: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014).
18. *Tan, J. et al.*: From neural sentence summarization to headline generation: A coarse-to-fine approach. In: Proceedings of the 26th international joint conference on artificial intelligence. pp. 4109–4115 AAAI Press (2017).
19. *Vaswani, A. et al.*: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017).

RUSSE'2020: FINDINGS OF THE FIRST TAXONOMY ENRICHMENT TASK FOR THE RUSSIAN LANGUAGE

Nikishina I. (Irina.Nikishina@skoltech.ru),
Logacheva V. (V.Logacheva@skoltech.ru),
Panchenko A. (A.Panchenko@skoltech.ru)

Skolkovo Institute of Science and Technology (Skoltech),
Moscow, Russia

Loukachevitch N. (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

This paper describes the results of the first shared task on taxonomy enrichment for the Russian language. The participants were asked to extend an existing taxonomy with previously unseen words: for each new word their systems should provide a ranked list of possible (candidate) hypernyms. In comparison to the previous tasks for other languages, our competition has a more realistic task setting: new words were provided without definitions. Instead, we provided a textual corpus where these new terms occurred. For this evaluation campaign, we developed a new evaluation dataset based on unpublished RuWordNet data. The shared task features two tracks: "nouns" and "verbs". 16 teams participated in the task demonstrating high results with more than a half of them outperforming the provided baseline.

Keywords: shared task, taxonomy, language resources, taxonomy enrichment, hypernyms, RuWordNet

DOI: 10.28995/2075-7182-2020-19-579-595

RUSSE'2020: ТЕСТИРОВАНИЕ МЕТОДОВ ПОПОЛНЕНИЯ ТАКСОНОМИИ ДЛЯ РУССКОГО ЯЗЫКА

Никишина И. (Irina.Nikishina@skoltech.ru),
Логачева В. (V.logacheva@skoltech.ru),
Панченко А. (A.Panchenko@skoltech.ru)

Сколковский институт науки и технологий (Сколтех),
Москва, Россия

Лукашевич Н. (louk_nat@mail.ru)

Московский государственный университет
им. М. В. Ломоносова, Москва, Россия

В данной работе описываются результаты первой дорожки RUSSE'2020 по пополнению таксономии терминов русского языка. Задача участников состоит в расширении существующей таксономии (RuWordNet): для новых слов необходимо предсказать их возможные гиперонимы. В отличие от соревнований, проводившихся для других языков, RUSSE'2020 имеет более реалистичную постановку: участникам не предоставляются определения для новых слов, при этом они могут использовать корпус текстов, в которых встречаются новые термины. Для оценки качества методов был подготовлен «золотой стандарт»: новые слова и их гиперонимы из неопубликованной версии RuWordNet, размеченные вручную. RUSSE'2020 включает в себя два трека по частям речи: «существительные» и «глаголы». Всего в соревновании участвовали 16 групп, более чем половине из них удалось превзойти предоставленное организаторами базовое решение.

Ключевые слова: соревнование, таксономия, обогащение таксономии, лингвистические ресурсы, гиперонимия, RuWordNet

1. Introduction

Taxonomies are tree structures that organize terms into a semantic hierarchy. Taxonomic relations (or hypernyms) are “is-a” relations: cat *is-a* animal, banana *is-a* fruit, Microsoft *is-a* company, etc. This type of relations is useful in a wide range of Natural Language Processing (NLP) tasks for performing semantic analysis.

While substantial interest is drawn to the extraction of hypernyms and taxonomic structures from text [6], [7], [9], the fully automatic taxonomy induction methods are still not widely used for routine construction of lexical resources, such as taxonomies. Nevertheless, the automatic hypernym candidate generation can facilitate and accelerate the manual taxonomy extension. Therefore, it is extremely useful to develop support tools for creation, enrichment, and maintenance of the existing semantic resources as well as their tuning to specific tasks and/or text collections.

Multiple evaluation campaigns tackling taxonomy problems have been organized for English and other Western European languages. Among them are SemEval-2018 task 9 on hypernym extraction [9], SemEval-2016 task 13 [7] and SemEval-2015 task 17 [6] on taxonomy induction, and SemEval-2016 task 14 [13] on taxonomy enrichment.

The main contribution of this paper is to report about RUSSE'2020—the first shared task on Taxonomy Enrichment for Russian, as well as for any other Slavic language. The goal of this semantic task is to extend an existing taxonomy with previously unseen words. For each new word—an *orphan*—the participants should provide a ranked list of possible hypernyms. RUSSE'2020 is similar to the SemEval-2016 task 14 [13], but has a more realistic setting. The participants are not given the definitions of the words to be added, but only a list of these words. However, the participants are allowed to use any additional resources.

We create a gold standard dataset for evaluating the participating systems. We consider the unreleased data from RuWordNet [19] as our gold standard and split it into two subsets: “nouns” and “verbs”. Moreover, we develop and release a baseline taxonomy enrichment model that uses an unsupervised approach based on word embeddings.

This paper is organized as follows. **Section 2** reviews the previous shared tasks on taxonomy creation, extension, and maintenance as well as hypernym extraction. **Section 3** introduces the task, the data, and the baseline model. The participating systems are described in **Section 4**, the overall results are provided in **Section 5**.

2. Related Work

Various methods were proposed for hypernym extraction, including pattern-based methods [12], [23], unsupervised and supervised methods based on word embeddings [27], [32], and hybrid approaches integrating several types of features [4], [29], [30].

In the majority of settings, hypernym extraction is cast as a binary classification task. Thus, the hypernym extraction algorithms are usually evaluated on purpose-built datasets containing positive and negative examples. One of such datasets is BLESS created by Baroni and Lenci [3] to test distributional models that predict several types of relationships between words.

In the semantic taxonomy enrichment task at SemEval 2016 [13], the organizers studied the possibilities of automatic addition of concepts from online glossaries and lexicographic resources into existing taxonomies such as WordNet [22]. Each new word was provided with a definition (gloss) from Wiktionary. The baseline model attached a new term to the first word from its gloss with the matching part of speech. Despite its simplicity, this approach turned out to be difficult to beat. It was outperformed by only one participating system. All participants used only Wiktionary glosses and did not try to employ any additional features from Wiktionary or text collections.

Bordea et al. [6], [7] evaluated taxonomy construction models based on the extracted hypernym relations. The evaluation was performed for several domains. Gold standard datasets were collected from WordNet and EUROVOC thesaurus¹. The authors suggested several metrics tailored for taxonomy evaluation.

Levy et al. [17] suggested that the results achieved in classification settings of hypernym extraction are mainly explained by the so-called “lexical memorization phenomenon”—a situation when models learn that in a relation “ x is-a y ” a word y is a prototypical hypernym. For example, if a classifier obtains many positive examples with the word y =animal, it may learn that anything that appears with y =animal should generate a positive answer. Camacho-Collados [8] argues that hypernym classification is not a realistic scenario. Instead, hypernym-oriented evaluation should be organized as a hypernym discovery task, i.e. given a word *dog*, the system should be able to discover its hypernyms *mammal* or *animal* among a large number of other possible candidates. He suggests evaluating models’ performance in this task with information-retrieval evaluation measures such as mean reciprocal rank (MRR) or mean average precision (MAP).

In the hypernym discovery task at SemEval 2018 [9], the organizers attempted to improve the quality of evaluation and formulated the hypernym extraction task as a ranking task. They created a list of hypernym candidates—these were all unigrams, bigrams, and trigrams that occurred more than N (for example, 5 times in the

¹ Eurovoc: <http://eurovoc.europa.eu/drupal>

corpus). For each of the new words and phrases, the participants were asked to rank the hypernym candidates by their relevance. Moreover, the participants had to find as many hypernyms as possible. The gold standard list of answers contained hypernyms of all hierarchy levels excluding only the most abstract concepts such as “entity”.

Panchenko et al. [24] describe the shared task on semantic similarity for Russian. One of the subtasks was to predict the similarity between words (synonym or hypernym relations). Each target word had the same number of related and unrelated source words. Reference answers were taken from the RuThes thesaurus [18].

Compared to the above mentioned competitions, RUSSE’2020 is closely related to the SemEval-2016 Taxonomy Enrichment Task [13] and SemEval-2018 Hypernym Discovery Task [9]. As in the mentioned SemEval tasks, in our competition the participants are asked to attach new words to the existing synsets, to create a ranked list of hypernym candidates, and the performance is evaluated using MAP and MRR metrics.

3. Shared Task Description

The goal of the task can be formulated as follows: given words that are not yet included in the taxonomy, we need to associate each word with the appropriate hypernym synset(s) from the existing taxonomy RuWordNet. For example, given an input word “утка” (duck) the participants are asked to provide a ranked list of its most probable 10 candidate hypernym synsets, e.g. “животное” (animal), “птица” (bird), and so on. We assume that an *orphan* may be a “child” of one, two, or more “ancestors” (hypernym synsets) at the same time.

The task featured two tracks: detection of hypernym synsets for nouns and verbs. We provided to participants the following resources: (i) training set based on the RuWordNet taxonomy, (ii) a collection of news texts from the year 2017 (2.2 billion tokens), (iii) a parsed Wikipedia corpus², and (iv) a hypernym database from the Russian Distributional Thesaurus³ [26], [28], which contains a set of hypernyms and a set of distributionally related terms both extracted from a huge text corpus. The participants were allowed to use any additional data and were asked to indicate the additional resources in their model descriptions.

The competition was hosted on the Codalab platform⁴. To allow the participants to evaluate their models on real data, we split the gold standard data into public and private test sets (denoted as “PRACTICE” and “EVALUATION” phases in Codalab). Thus, the participants could test their models before the deadline on the public test set by submitting the results to the “PRACTICE” leaderboard. During the “EVALUATION” phase the leaderboard was hidden, so the participants were not able to overfit the test data.

² <https://doi.org/10.5281/zenodo.3827903>

³ <https://doi.org/10.5281/zenodo.3827834>

⁴ <https://competitions.codalab.org/competitions/22168>

Table 1: Number of RuWordNet synsets in datasets used in the shared task

	Nouns	Verbs
Total in RuWordNet	29,297	7,636
Train set	12,393	2,109
Private test set	1,525	350
Public test set	763	175

3.1. Datasets and Additional Resources

We provided the gold standard dataset which contains words with manually defined hypernyms. These words were included in the extended version of RuWordNet which has not been published yet. We split this data into two parts: public (763 nouns and 175 verbs) and private (1,525 nouns and 350 verbs).

The words included in the gold standard test dataset (*orphans* in [Table 1](#)) were collected in the following way. First, we extracted words (nouns and verbs) which are present in the extended RuWordNet, but absent in the published RuWordNet. We selected only single words (not phrases) with at least 50 occurrences in the corpus of news texts from 2017. Then we filtered the obtained list excluding the following words:

- all three-symbol words and the majority of four-symbol words;
- diminutive word forms and feminine gender-specific job titles;
- words which are derived from words which are included in the published RuWordNet;
- words denoting inhabitants of cities and countries;
- geographic and personal names;
- compound words that contain their hypernym as a substring.

The gold hypernyms of the *orphan* words were assigned manually by linguists. However, it should be noted that these gold hypernyms are not necessarily the closest hypernyms. The extended RuWordNet can contain whole chains of hypernyms none of which is included in the published version. If one of the synsets selected for the test set belongs to this chain and its immediate hypernym is not presented in the published version, we set its closest published “ancestor” as a gold hypernym.

The training dataset (words paired with hypernyms) was generated from the current version of the RuWordNet taxonomy and annotated analogously to the test data. To create the training set we sampled all leaves (synsets with no hyponyms) of depth equal or more than 5. Overall, it comprises 12,393 nouns and 2,102 verbs.

The news text collection, which was provided to the participants, consists of 8 million news articles written in 2017 collected from more than 1,000 news sources. It contains a total of 2.2 billion tokens. This corpus was initially collected so that it contains at least 50 occurrences of the majority of words from the test data. However, it was further discovered that 17 words in the public test (1.8%) and 42 words in the private set (2.2%) have fewer occurrences in the corpus, due to the use of different lemmatization tools and morphological ambiguity.

Table 2: Various ground truth representations for the term “cruise”.

The connectivity component representation allows us to take into account the fact that all three direct hypernyms are related to the same word sense, as depicted in Figure 1(d), and do not wrongly penalise a system that predicted only one of them

Set of direct hypernyms	{ entertaining journey, journey, tour }
Sets of direct hypernyms and their parents	{ entertaining journey , travel, entertainment, active leisure}, { journey , travel, move}, { tour , travel, journey, active leisure}
Connectivity component	{ entertaining journey, journey, tour , travel, entertainment, active leisure, move}

3.2. Evaluation Metrics

The participants were asked to generate a ranked list of 10 most probable hypernym candidates for each word in the test set. The results were evaluated using the Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) scores. MAP score evaluates the whole range of produced hypernym candidates, whereas MRR looks at how close the first correct prediction is to the top of the list. We consider MAP as the official metric of our competition.

Both metrics are widely used in the Hypernym Discovery shared tasks, where systems also need to output ranked lists of candidate hypernyms [9]. In contrast to [9], we limited the number of possible answers to $k = 10$, because the correct answers from lower positions will have small weights and will not contribute much to the final score.

To be less restrictive during the evaluation, we consider as correct answers not only the immediate hypernyms of new words but also the hypernyms of these hypernyms. Therefore, if a system predicts a hypernym of a correct hypernym, this will also be considered a match.

One hypernym may be a “parent” of another hypernym (synset “plane” has two parents—“aircraft” and “aviation technology”, whereas “aviation technology” itself is the hypernym for “aircraft”). While computing the MAP score, it may not be clear which hypernym gains the score: “aviation technology” synset as the immediate hypernym or “aviation technology” as the second-order hypernym. Hypernyms may also have common parents: “string instrument” and “folk instrument” both have a hypernym “musical instrument”. In this case, if “musical instrument” appears in the candidate list, the MAP score will also be confused.

To avoid this hypernym ambiguity, we split all hypernyms of a word (both immediate and second-order) into groups. Each group corresponds to the connectivity component in the subgraph reconstructed from all hypernyms. The process is shown in Figure 1. We see that the first and the second subgraphs consist of only one connectivity component, whereas in the third graph the immediate hypernyms form different hypernym groups. Therefore, the list of possible candidates of a given word should contain at least one hypernym from each hypernym group. Thus, connectivity components allow us to distinguish between cases depicted in Figure 1(a) where a system must predict hypernyms for both word senses from two independent branches and (b)/(c)/(d) where only one word sense is to be predicted.

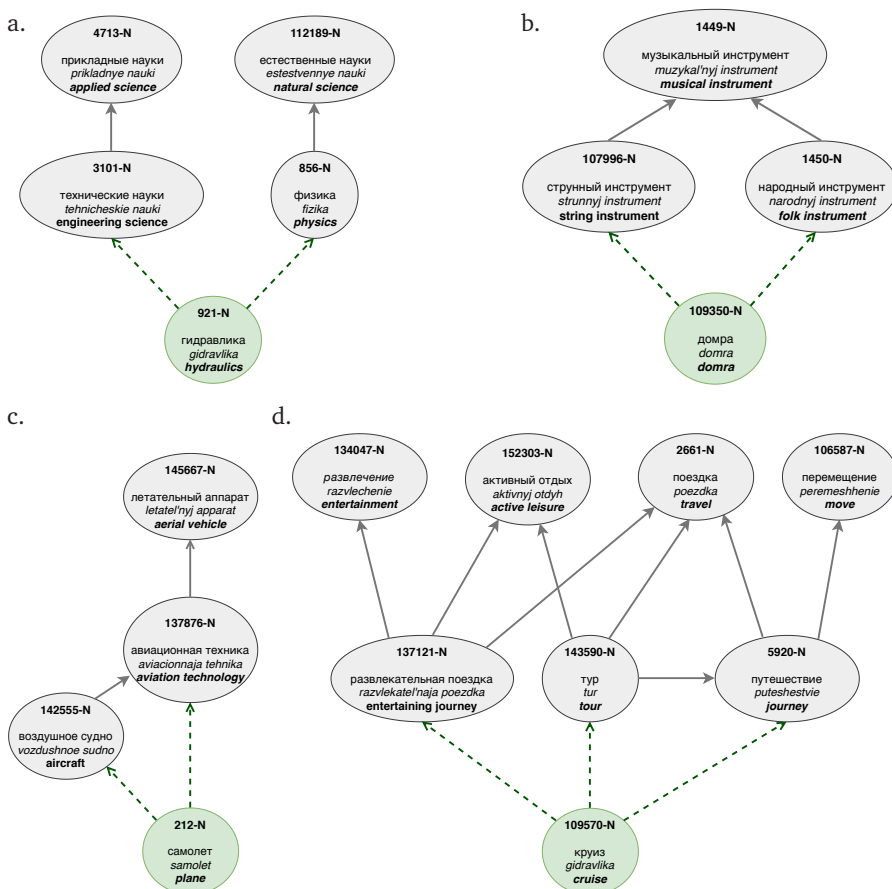


Figure 1: Examples of hypernym subgraphs from RuWordNet ground truth: direct and second-order hypernyms may be related in various ways motivating the evaluation metric based on connectivity components. While in (a) two parents lead to different senses, in (b, c, d) two parents lead to the same sense. Dashed lines indicate ground truth hypernyms.

Overall, to compute the score, we extend the standard MAP reference and group hypernyms into connectivity components (see evaluation examples in **Table 2** for the word “cruise”). **The answer is given a full score if there is at least one hypernym from each connectivity component in the list of possible candidates.** To get the highest score for the example from **Table 2**, it is enough to predict one of the synsets. Moreover, all hypernyms of all connectivity components are considered equally relevant: predictions starting with “applied science” and “physics” or with “natural science”, and “engineering science” will get the same score.

3.3. Baseline

We implemented a simple baseline that makes use of non-contextualized (standard) word embeddings. We chose fastText embeddings⁵ [5] to solve this task for two reasons: pre-trained fastText models are easy to deploy and they do not require any additional data or training for the out-of-vocabulary words, because they incorporate subword tokens.

Our baseline comprises the following steps:

1. Compute embeddings of all synsets in RuWordNet by averaging embeddings of all words from senses belonging to a synset.
2. Get embeddings for *orphans*. For multi-word *orphans* the embeddings are computed by averaging vectors for all words comprising an *orphan*.
3. For each *orphan* compute the top $k = 10$ closest synsets of the same part of speech as the orphan using the cosine similarity measure.
4. Extract hypernyms for each of these closest synsets from the previous step. Take the first $n = 10$ results (as each synset may have several hypernyms).

Our method is unsupervised and does not require any additional data. Nevertheless, it turned out to be a strong baseline as shown below.

4. Participating Systems

RUSSE'2020 shared task attracted 16 participants in the “nouns” track and 14 in the “verbs” track (excluding the baseline). We provide descriptions of the top 7 solutions which outperformed the baseline at any track. We denote 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 **Tables 3** and **4**.

4.1. Yuriy

This participant-generated candidate hypernyms and calculated features for them. Then candidates were ranked by a linear model with handcrafted weights. The list of features is provided below:

1. candidate is in top 10 similar words from RuWordNet;
2. candidate is in hypernyms of top 10 similar words from RuWordNet;
3. candidate is in hypernyms of hypernyms of top 10 similar words from RuWordNet;
4. candidate is in hypernyms on Wiktionary⁶ page about the word;
5. candidate is in hypernyms of hypernyms on Wiktionary page about the word;

⁵ <https://fasttext.cc/docs/en/crawl-vectors.html>

⁶ <https://ru.wiktionary.org>

6. candidate is in “en-ru” translation of WordNet [21] hypernyms of “ru-en” translation of the word (extracted with Yandex Machine Translation model⁷);
7. candidate is in the word definition in the Wiktionary page;
8. candidate is in the Yandex search result page;
9. candidate is in the Google search result page.

The candidates were collected using features 1–6. Features 1–3 are based on the fastText model⁸. This approach was applied for both “nouns” and “verbs” tracks.

4.2. xeno

This participant merged candidates extracted by several methods. Those methods included: Russian Wiktionary semantic graph (taxonomic relations, synonymy, antonymy); rule-based plain text definition parsing; rule-based plain text parsing with Hearst patterns on Russian Wikipedia from [26] and Russian language corpus; graph-based analysis of the nearest neighbor list obtained from word2vec. The definitions were taken from Russian Wiktionary, Russian Wikipedia, Big English-Russian polytechnic dictionary, Efremova dictionary [11]. The above-mentioned methods were used for nouns. For verbs, the team used only the Russian Wiktionary semantic graph and rule-based plain text definition parsing.

4.3. KuKuPl

[14] This team trained a classifier on the official train data provided by the organizers. They considered synsets (occurring more than n times in the training data) as classes, representing words with the embeddings (standard CBOW from word2vec) pretrained a concatenation of four corpora: Araneum Russicum Maximum, Russian Wikipedia, Russian National Corpus, and a corpus of Russian news (9.5 billion word tokens overall). The corpus was specially tailored for this task: all multi-word entities which also occurred in the RuWordNet were merged into single tokens, thus making sure that the majority of the RuWordNet entries received their respective vector representations.

A neural network classifier with one hidden layer (size 386), dropout of 0.1, ReLu activation, and a softmax output layer was trained on all the training data until convergence, using hypernym synset ids as class labels. At test time, the trained model obtains the vector representation of a query word and predicts possible classes (hypernym synsets) for this vector. 10 synsets with the highest probability are considered predictions. This approach is applied for both “nouns” and “verbs” tracks.

4.4. RefalMachine, Parkat13

[31] This team implemented the algorithm consisting of three stages. Firstly, they created a list of similar words using a combination of vector representations of words

⁷ <https://translate.yandex.ru>

⁸ <https://fasttext.cc/docs/en/crawl-vectors.html>

obtained with PPMI (positive pointwise mutual information) weighting and SVD factorization (window = 1). Secondly, they selected candidates from those similar words (depending on pattern matching), their hypernyms, and second-order hypernyms. These candidates were ranked based on the following features:

- cosine similarity;
- patterns matching co-hyponyms;
- patterns matching hypernyms (Hearst patterns). The patterns were extracted from the news corpus provided by the organizers;
- the number of synset occurrences in the candidate list;
- probabilities based on ruBERT predictions [15].

The final rank for each candidate was computed using the weighted feature combination; the weights are hand-picked during the experiments. This approach was applied for both “nouns” and “verbs” tracks.

4.5. MorphoBabushka (alvadia, maxfed, joystick)

[1] This team used the following pipeline. First, they retrieved nearest neighbors for the target word from word2vec “SkipGram with Negative Sampling” model trained on Librusec book collection [2] and search for their direct and indirect hypernyms in RuWordNet. Then they counted direct and indirect hypernyms of the nearest neighbors, combining their counts, converting (or excluding if not possible) inappropriate ones with wrong part-of-speech. They took 10 most frequent hypernyms of nearest neighbors’ synsets. Finally, they combined those hypernyms with the hypernyms extracted from Wiktionary by matching definition N-grams with the synsets. This method was applied for both “nouns” and “verbs” tracks.

4.6. cointegrated

[10] The participant used similarity scores between word embeddings to predict hypernym relations. For each RuWordNet synset, the team computed the embedding of its title, all senses, and the mean embedding of the title and all senses. Each type of the above-mentioned embeddings was computed as an L2-normalized weighted mean of its word embeddings from RusVectors[16] (weight is of 1.0 nouns, 0.1 for prepositions, and 0.5 for all other POS). For OOV words, the embedding was computed as a mean embedding of all words in the vocabulary with the longest prefix matching the target word.

For each query word (orphan), the participant found its 100 nearest neighbors from RuWordNet and all the first and second-order hypernyms of the corresponding synsets, considering them as answer candidates. The resulting list of hypernyms comprises 10 candidates with the highest scores. The score for each candidate is a sum of “neighbor scores” overall nearest neighbors from RuWordNet; if the candidate is a second-order hypernym, its “neighbor score” is multiplied by 0.5. The “neighbor score” is calculated as $\exp(-3 \cdot d) \cdot s^5$, where d is the distance between the queries and neighbor embeddings; s is their cosine similarity. The described approach was applied for both “nouns” and “verbs” tracks.

5. Results

Tables 3 and 4 present respectively the results “nouns” and “verbs” tracks. As one can observe, the absolute difference in scores of the two tracks is quite large. Apparently, the “verbs” track is more difficult, because word embeddings for verbs are not as accurate and exhaustive as for nouns: verbs are more abstract and can be seen in a context with a wider range of words than nouns [25].

All the methods applied by the participants can be divided into two classes. The first class applies supervised learning (binary or multi-class classification). The second one performs ranking based on a range of features (similarity measures, hypernyms of different orders, etc.). Surprisingly, the majority of approaches are not stable across the tasks: they can demonstrate promising results on the “nouns” track, but lag behind on “verbs” (e.g. *KuKuPl*, *RefalMachine*) or vice versa (e.g. *cointegrated*).

Another interesting point is the type of embeddings that was used by the top-7 participants. Apart from *RefalMachine*'s, no methods used contextualized embeddings. The most popular vector model is word2vec [20], pre-trained (*Yuriy*, *cointegrated*) or trained on the provided datasets (*KuKuPl*, *MorphoBabushka*).

Interestingly, all the top-7 participants resort to additional data. The most popular additional source are text corpora: *KuKuPl*, *MorphoBabushka* use corpora to train custom word embeddings, *cointegrated* and *Yuriy* apply pre-trained embeddings. The 2017 news corpus with contexts for word occurrences is used by three teams (out of the top 7 teams described in this paper): *KuKuPl*, *Parkat13* and *RefalMachine*. Another promising source of information are dictionaries: *MorphoBabushka* and *Yuriy* give their preference to Wiktionary, whereas *xeno* uses Big English-Russian polytechnic dictionary, Efremova dictionary. The most outstanding range of additional resources (from *Yuriy*) includes Wiktionary, Yandex Translate, Google, and Yandex search pages results. However, we cannot draw any conclusions about the efficiency of the use of additional data, as these sources are not the only factors that influenced the final results.

Table 3: Evaluation results for “nouns” track on the private test dataset

Rank	User	Entries	MAP	MRR
1	Yuriy	5	0.5522	0.5940
2	xeno	5	0.5054	0.5433
3	KuKuPl	2	0.4976	0.5332
4	RefalMachine	6	0.4930	0.5314
5	MorphoBabushka	5	0.4497	0.4835
6	baseline	1	0.4210	0.4518
7	cointegrated	5	0.4178	0.4503
8	adhaesitadimo	1	0.3759	0.4043
9	vvyadrincev	2	0.3095	0.3342
10	vimary	4	0.2951	0.3187

Table 4: Evaluation results for “verbs” track on the private test dataset

Rank	User	Entries	MAP	MRR
1	cointegrated	3	0.4483	0.5049
2	Yuriy	2	0.4355	0.5135
3	MorphoBabushka	5	0.3890	0.4419
4	baseline	1	0.3335	0.3817
5	xeno	2	0.3075	0.3547
6	RefalMachine	5	0.2542	0.2969
7	KuKuPl	3	0.2470	0.2897
8	vimary	2	0.1783	0.2115
9	vvyadrintsev	3	0.1474	0.1786
10	Arshehremen	2	0.0000	0.0000

In order to analyse the results obtained by the participants, we provide several examples for both verbs and nouns (Tables 5 and 6)⁹. We took 3 nouns from Yuriy’s answer and 3 verbs from *cointegrated*’s to compare with the gold standard hypernym synset subgraphs (“ground truth” part of Tables 5 and 6). For the nouns “сахарин” (saccharin), “селфи” (selfie) and the verb “тусить” (to party) candidate lists contain either all hypernyms or at least one hypernym from all subgraphs. These examples also demonstrate that the systems are capable of accurate and correct predictions. Moreover, even for verbs “прохлаждаться” (to be hanging around) and “фотошопить” (to photoshop) and for the noun “кэшбэк” (cashback) the systems predicted synsets which are very close to the correct meaning, but they either cannot predict the whole variety of synsets or predict hypernyms in the proximity to the correct ones. The task of automatic taxonomy enrichment is technically feasible, but it still requires more sophisticated approaches.

As has been noted above, the most similar competition to ours is the SemEval-2018 hypernym discovery task (task 9). However, the setting used at SemEval is still quite different from ours—in particular, there, the participants of the task had to construct a taxonomy from scratch, whereas we ask our participants to extend an existing taxonomy. If we compare the scores of SemEval participants and models submitted to our task, we can see that models participating in our task yielded significantly higher MRR scores—almost 0.6 for the best-performing models compared to 0.3 for the winners of SemEval. This suggests that our task turns out to be easier than the full taxonomy construction. Obviously, the settings are quite diverse and cannot be compared rigorously—we asked participants to output $K = 10$ hypernym candidates, while at SemEval K was set to 15, the lexis were different, so we have no information about whether one test set was easier than the other. Finally, the tasks were for different languages. However, we can still speculate that such a large difference in scores is mainly because in our task the participants were using the existing taxonomy for their predictions. If they were not using it, as in SemEval, this task would not be any easier.

⁹ English: https://competitions.codalab.org/competitions/22168#learn_the_details-results

Table 5: Predicted hypernym synsets from RuWordNet for nouns from *Yuriy's* answer (top-1 for nouns).

Green color denotes predictions of the model from the ground truth.

rank	сахарин	селфи	кэшбэк
1	подсластитель	изображение (результат)	скидка
2	заменитель	фотографическое изображение	сфера деятельности
3	пищевые добавки	фотосъемка	предоставление услуги
4	добавление (то, что добавлено)	кинофотосъемка	учетная операция
5	вещество	портрет (изображение)	вексельная операция
6	сахарозаменитель	ателье бытовых услуг	учетная ставка
7	материал для изготовления	фотоателье	понизить величину
8	сахара	движение, перемещение	льгота
9	сахар	автопортрет	действие, целенаправленное действие
10	продукты питания	постоянная сущность	банковская операция
ground truth	заменитель подсластитель сахарозаменитель пищевые добавки	автопортрет	вернуть взятое
		портрет (изображение)	возврат имущества, средств
		фотографическое изображение	премия
		фотопортрет	бонус (вознаграждение)

Table 6: Predicted hypernyms synsets from RuWordNet for verbs from *cointegrated's* answer (top-1 for verbs).

Green color denotes predictions of the model from the ground truth.

rank	тусить	прохлаждаться	фотошопить
1	собраться в одном месте	бездельничать	воспроизвести (воссоздать, повторить в копии)
2	общение, связь	недостойное поведение	исправить недостатки, ошибки
3	веселиться	бродить туда-сюда	копирование, снятие копии

rank	тусить	прохлаждаться	фотошопить
4	занятие, деятельность	находиться, пребывать	изобразить (воспроизвести)
5	отношения между людьми	лежать (находиться всем телом на поверхности)	проверить, удостовериться в правильности
6	пробыть, провести время	пробыть, провести время	обеспечить, снабдить
7	развлечься, приятно провести время	отдых	создать (сделать существующим)
8	добраться до места	идти ногами	устранить (уничтожить)
9	идти ногами	веселиться	исправить, улучшить
10	отдых	медлить	находиться, пребывать
ground truth	пробыть, провести время развлечься, приятно провести время занятие, деятельность	недостойное поведение бездельничать	преувеличить представить в виде приукрасить, выгодно представить
		медлить действие, целенаправленное действие	
	тусоваться добраться до места собраться в одном месте	освежить, восстановить силы восстановить прежнее состояние	изменить, сделать иным видеоизменить

6. Conclusion

We present the results of the first shared task on Taxonomy Enrichment for Russian. For this shared task, we created a new dataset from the unpublished data of RuWordNet. 16 teams participated in the task, and almost half of them outperformed the baseline model.

Undoubtedly, the provided gold standard may not be perfect and exhaustive. Such manual evaluation of system answers would provide a more objective result, but we did not perform it because of the time constraints. Manual inspection of system outputs by an expert could reveal valid hypernyms identified by systems but absent in the gold standard data.

Moreover, the best-performing methods presented by participants might not be optimal for some words. These methods are based on fastText and similar distributional models, such as word2vec. However, it is known that these low-variance and high-bias models tend to identify the dominant meaning of a word and populate nearest neighbor lists with words related to this dominant meaning. Therefore, some rare senses of hypernyms can be underrepresented based on such methods. Identifying them correctly requires using alternative approaches.

According to the provided results, we see that the automatic hypernym candidate generation from an existing taxonomy is a feasible task, so it can be used to assist manual taxonomy enrichment. We hope that the evaluation datasets will foster further development of taxonomy induction and enrichment methods. Besides, the obtained levels of quality will allow direct use of some of the best-performing methods in the further development of lexical resources, such as thesauri, taxonomies, and ontologies.

7. Acknowledgements

The work of Natalia Loukachevitch in the current study (preparation of RuWordNet data for the shared task) is supported by the RFBR foundation (project N 18-00-01226 (18-00-01240)). We thank Dmitry Ustalov for updating the RUSSE web site with the information about the current shared task. Finally we are grateful to RUSSIR, AIST, and AINL conference organizers, Moscow NLP Seminar organizers, and Vladislav Lialin for sharing the information about this shared task in their media resources.

References

1. *Arefyev, N. et al.*: Word2vec not dead: predicting hypernyms of co-hyponyms is better than reading definitions. In: Computational linguistics and intellectual technologies: Papers from the annual conference “Dialogue”. (2020).
2. *Arefyev, N. et al.*: Evaluating three corpus-based semantic similarity systems for russian. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue-2015”. 28, (2015).
3. *Baroni, M., Lenci, A.*: How we blessed distributional semantic evaluation. In: Proceedings of the gems 2011 workshop on geometrical models of natural language semantics. pp. 1–10 Association for Computational Linguistics (2011).
4. *Bernier-Colborne, G., Barrière, C.*: CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In: Proceedings of the 12th international workshop on semantic evaluation. pp. 725–731 Association for Computational Linguistics, New Orleans, Louisiana (2018).
5. *Bojanowski, P. et al.*: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 5, 135–146 (2017).
6. *Bordea, G. et al.*: Semeval-2015 task 17: Taxonomy extraction evaluation (texteval). In: Proceedings of the 10th international workshop on semantic evaluation (semeval-2016). pp. 1081–1091 (2016).
7. *Bordea, G. et al.*: Semeval-2016 task 13: Taxonomy extraction evaluation (texteval-2). In: Proceedings of the 10th international workshop on semantic evaluation (semeval-2016). pp. 1081–1091 (2016).
8. *Camacho-Collados, J.*: Why we have switched from building full-fledged taxonomies to simply detecting hypernymy relations. arXiv preprint arXiv:1703.04178. (2017).

9. *Camacho-Collados, J. et al.*: SemEval-2018 task 9: Hypernym discovery. In: Proceedings of the 12th international workshop on semantic evaluation (semeval-2018); 2018 jun 5–6; new orleans, la. Stroudsburg (pa): ACL; 2018. P. 712–24. ACL (Association for Computational Linguistics) (2018).
10. *Dale, D.*: A simple solution for the Taxonomy enrichment task: Discovering hypernyms using nearest neighbor search. In: Computational linguistics and intellectual technologies: Papers from the annual conference “dialogue”. (2020).
11. *Efremova, T.*: New dictionary of the russian language. Explanatory-derivational. Moscow: Russky yazyk. (2000).
12. *Hearst, M. A.*: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on computational linguistics-volume 2. pp. 539–545 Association for Computational Linguistics (1992).
13. *Jurgens, D., Pilehvar, M. T.*: SemEval-2016 task 14: Semantic taxonomy enrichment. In: Proceedings of the 10th international workshop on semantic evaluation (semeval-2016). pp. 1092–1102 (2016).
14. *Kunilovskaya, M. et al.*: Taxonomy Enrichment: Linear Hyponym-Hypernym Projection vs Synset ID Classification. In: Computational linguistics and intellectual technologies: Papers from the annual conference “dialogue”. (2020).
15. *Kurатов, Y., Arkhipov, M.*: Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213. (2019).
16. *Kutuzov, A., Kuzmenko, E.*: WebVectors: A toolkit for building web interfaces for vector semantic models. In: Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers. pp. 155–161 Springer International Publishing, Cham (2017).
17. *Levy, O. et al.*: Do supervised distributional methods really learn lexical inference relations? In: Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies. pp. 970–976 (2015).
18. *Loukachevitch, N., Dobrov, B. V.*: RuThes linguistic ontology vs. Russian wordnets. In: Proceedings of the seventh global wordnet conference. pp. 154–162 (2014).
19. *Loukachevitch, N. V. et al.*: Creating russian wordnet by conversion. In: Computational linguistics and intellectual technologies: Papers from the annual conference” dialogue. p. 22 (2016).
20. *Mikolov, T. et al.*: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013).
21. *Miller, G. A.*: Nouns in wordnet. WordNet: An electronic lexical database. 23–46 (1998).
22. *Miller, G. A.*: WordNet: A Lexical Database for English. Communications of the ACM. 38, 11, 39–41 (1995).
23. *Nakashole, N. et al.*: PATTY: A taxonomy of relational patterns with semantic types. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 1135–1145 Association for Computational Linguistics (2012).

24. *Panchenko, A. et al.*: RUSSE: The first workshop on russian semantic similarity. In: Computational linguistics and intellectual technologies: Papers from the annual international conference “Dialogue-2015”. pp. 89–106 (2015).
25. *Panchenko, A. et al.*: Noun sense induction and disambiguation using graph-based distributional semantics. In: Proceedings of the 13th conference on natural language processing (konvens 2016). pp. 192–202 (2016).
26. *Panchenko, A. et al.*: Human and machine judgements for russian semantic relatedness. In: International conference on analysis of images, social networks and texts. pp. 221–235 Springer (2016).
27. *Roller, S. et al.*: Inclusive yet selective: Supervised distributional hypernymy detection. In: Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers. pp. 1025–1036 (2014).
28. *Sabirova, K., Lukanin, A.*: Automatic extraction of hypernyms and hyponyms from russian texts. In: Ignatov, D. I. et al. (eds.) Supplementary proceedings of the 3rd international conference on analysis of images, social networks and texts (aist’2014), yekaterinburg, russia, april 10–12, 2014. pp. 35–40 CEUR-WS.org (2014).
29. *Shwartz, V. et al.*: Improving hypernymy detection with an integrated path-based and distributional method. In: Proceedings of the 54th annual meeting of the association for computational linguistics. pp. 2389–2398 (2016).
30. *Snow, R. et al.*: Semantic taxonomy induction from heterogenous evidence. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics. pp. 801–808 Association for Computational Linguistics (2006).
31. *Tikhomirov, M. et al.*: Combined Approach to Hypernym Detection for Thesaurus Enrichment. In: Computational linguistics and intellectual technologies: Papers from the annual conference “Dialogue”. (2020).
32. *Weeds, J. et al.*: Learning to distinguish hypernyms and co-hyponyms. In: Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers. pp. 2249–2259 Dublin City University; Association for Computational Linguistics (2014).

СОВМЕСТНЫЙ СИНТАКСИС В ДИАЛОГАХ С ЗАИКАЮЩИМИСЯ

Оленикова А. В. (tradaladei@gmail.com),

Федорова О. В. (olga.fedorova@msu.ru)

МГУ имени Ломоносова, Москва, Россия

Ключевые слова: заикание, совместный синтаксис, расширения, завершения, анализ бытового диалога, диалог

DOI: 10.28995/2075-7182-2020-19-596-605

CO-CONSTRUCTED SYNTACTIC UNITS IN DIALOGUES WITH INDIVIDUALS WHO STUTTER

Olenikova A. V. (tradaladei@gmail.com),

Fedorova O. V. (olga.fedorova@msu.ru)

Lomonosov Moscow State University, Moscow, Russia

Dialogue implies a high degree of coordination between the interlocutors, which makes possible the existence of co-constructed turns used by speakers for various purposes. One of the reasons for them to appear is difficulties in articulation experienced by one of the participants and prompting the other participant interested in achieving the communicative goal to increase their own contribution to the dialogue. In conversations with people who stutter, co-constructions are more common than in conversations between people who have no diagnosed speech disorders; among them completions prevail, because one of the interlocutors more often spells out uncompleted constructions. The study of stuttering from a linguistic perspective is of considerable interest, since it provides an opportunity to study dialogue as a process including cooperation between participants. During this collaborative process one interlocutor's contribution affects the contribution of another and can trigger non-standard turn-taking techniques.

Key words: stuttering, co-constructed turns, extensions, completions, conversation analysis, dialogue

1. Введение. Совместный синтаксис

Важнейшей особенностью диалога, обеспечивающей его существование, является координация реплик между участниками. Реплики разных говорящих в диалоге согласуются между собой на синтаксическом, семантическом и прагматическом уровне и следуют друг за другом с минимальным количеством наложений.

Свидетельством такой координации служат так называемые «совместные построения» (англ. *co-constructed turns*). Этот термин возник в рамках анализа бытового диалога для обозначения реплик, в которых одна и та же синтаксическая конструкция начинается одним участником диалога и продолжается другим или даже несколькими другими; существование таких построений и их значение для лингвистической теории обсуждается в [Lerner 1991], [1996]. Сама возможность разделения конструкции между двумя говорящими означает высокую степень координации между ними в процессе планирования, порождения и обработки высказываний в дискурсе [Гренобль 2008]. Граница между репликами говорящих, образующими совместное построение, может возникать в любой точке высказывания, в том числе внутри составляющей [Purver et al. 2009].

В плане соотношения между дискурсом и синтаксисом совместные построения делят на *расширения* (англ. *extensions*), при которых второй говорящий добавляет элемент к уже законченной конструкции, и *завершения* (англ. *completions*), при которых второй говорящий дополняет не законченную первым синтаксическую структуру [Ono, Thompson 1995]. *Точка перехода* (англ. *transition relevance place*) возникает в ситуации, когда текущая реплика может быть завершена и другой говорящий может начать свою реплику. В качестве сигналов точки перехода могут выступать просодические факторы (интонация, паузы) и семантическая и синтаксическая завершенность. В случае расширения в примере (1) говорящий А произносит законченную синтаксическую структуру и достигает точки перехода, после которой говорящий В решает дополнить его реплику¹, расширив ее до сочинительной конструкции. Говорящий А повторяет расширение, соглашаясь принять его.

- (1) танграмма [L] — «японка с ногами зигзагом» (см. рис. 1 ниже)²
 А: Пятый присогнул колени.
 В: Но не сел.
 А: Но не сел.

корпус TanAdult_{rus} [Федорова 2016]

В примере (2) говорящий А испытывает трудности при описании изображения и не может завершить начатую сочинительную конструкцию. Говорящий В реагирует на хезитацию собеседника, предлагая свое завершение конструкции, впоследствии принятое говорящим А.

- (2) танграмма [С] — «привидение»
 А: А последний — это человечек, который расставил руки и-и-и=
 В: и наклонился налево туловищем.
 А: Да, и наклонился налево.

корпус TanAdult_{rus} [Федорова 2016]

¹ Здесь и далее выделяется в реплике курсивом; полужирный шрифт используется в примерах для содержательного выделения.

² Здесь и далее: диалоги были расшифрованы в упрощенной дискурсивной транскрипции; в скобках дается время пауз хезитации, как заполненных, так и незаполненных; знаком равенства обозначен обрыв слова, буквами через дефис — удлинения.

И расширяющая, и завершающая реплики не могут рассматриваться как независимые высказывания, поскольку в обоих случаях происходит согласование предиката с субъектом предыдущей клаузы, произнесенной другим говорящим. Различие между ними заключается в том, что в случае завершения говорящий А на момент точки перехода не успел закончить свою речь. Преобладание того или иного типа совместных построений, как и их частотность, зависит от специфики дискурса и индивидуального стиля говорящих: в материале некоторых исследований больше представлены расширения, см. эксперимент с носителями немецкого языка в [Skuplik 1999], в других — завершения, например 75 % завершений из 200 случаев в англоязычных неформальных диалогах, см. [Szczeppek 2000].

Первое исследование совместных построений на материале русского языка предприняла Л. Гренобль, которая использовала корпус интервью радио «Эхо Москвы». В [Гренобль 2008] исследовательница отмечает, что среди встретившихся в течение 10 часов записи типов расширений самым частым оказалось добавление предложной группы. Среди завершений встретились совместно построенные ходы различных типов как внутри составляющих, так и на границах; отмечено, что в некоторых случаях первая часть может накладывать существенные морфосинтаксические ограничения на продолжение, например, определять форму глагола.

Говорящие используют совместные построения для достижения разнообразных целей, в числе которых исследователи выделяют следующие: дополнить сказанное предыдущим говорящим, получить от него дополнительную информацию, перехватить инициативу в диалоге, завершить диалог [Гренобль 2008], выразить отношение к предыдущей реплике [Ono, Thompson 1996], в том числе не согласиться и высказаться иронически; говорящий может использовать совместное построение, чтобы продемонстрировать понимание темы разговора или снять с себя часть ответственности за собственное высказывание [Szczeppek 2000].

2. Совместный синтаксис и заикание

В поисках постоянных и ситуативных факторов, вызывающих нарушения плавности речи, исследователи обратились к методам анализа бытового диалога, чтобы собрать достаточно достоверных эмпирических данных о коммуникации людей с заиканием [Acton 2004]. Многие исследования при этом были посвящены общению в диадах, состоящих из пациентов с заиканием и их родственников или друзей. В одной из таких статей рассматриваются совместные построения в ситуации диалога [Lind, Sønsterud 2014]. Авторы считают, что собеседники используют совместный синтаксис в ситуациях, когда заикающиеся испытывают значимые речевые затруднения, что в целом оказывает позитивное влияние на коммуникацию. Повторы звуков, слогов и слов, паузы и блоки, во время которых заикающийся человек пытается, но не может говорить, побуждают собеседника использовать различные средства поддержания коммуникации, в том числе совместные построения. Описанию совместных построений такого типа и посвящена эта статья.

Для обозначения явлений, связанных с заиканием, далее будут использоваться следующие дополнительные символы, адаптированные из [Ratner et al. 1996].

Символ	Явление	Пример
&	повтор звука/слога	з ^н ачит п(&Зп)ервая картинка с(&5с) лева вот верхняя
/	повтор слова/группы слов	[ну как](/2) одна нога вот поднята
^	блоки	человек с (э 0.72) ^ кастрюлей что ли
~	блоки внутри слов	путеш~ественник очень уставший

3. Корпус TanStutter_{rus}

Корпус TanStutter_{rus}, состоящий из 8 записей, был собран летом-осенью 2019 г. в Психологическом институте РАО³.

3.1. Процедура, стимульный материал и оборудование

Парам испытуемых, в каждой из которых один из участников был диагностирован как заикающийся, был предложен эксперимент по методу референциальной коммуникации, суть которого состоит в том, что два собеседника совместными вербальными усилиями решают некоторую когнитивную задачу, успешность решения которой во многом зависит от слаженности их вербального поведения. Инструктор должен вербально донести до второго собеседника, Раскладчика, информацию, которая понадобится последнему, чтобы выполнить действия, необходимые для успешного решения поставленной задачи. Эксперимент проводился по методике «беспорядок» с использованием танграмм — фигурок из китайской игры-головоломки, которые с трудом поддаются вербальному описанию (см. **рис. 1**). Данная методика была усовершенствована Г. Кларком с коллегами в целой серии экспериментов с танграммами [Clark, Wilkes-Gibbs 1986].

Работа [Clark, Wilkes-Gibbs 1986] была посвящена вопросам референции, однако впоследствии этот метод позволил перейти к построению общей модели взаимодействия собеседников в диалоге, а стимульный материал широко использовался в других психолингвистических исследованиях. На русском материале метод «Танграммы» был ранее неоднократно использован в исследованиях с взрослыми испытуемыми, детьми 10–12 лет, пациентами с афазией [Зинова и др. 2011]. Данный метод представляет собой, на наш взгляд, удачный компромисс между полностью неподготовленным общением испытуемых и строгим экспериментальным дизайном: с одной стороны, он дает возможность получить сравнимые между собой диалоги; с другой стороны, коммуникация

³ Авторы выражают благодарность Н. Л. Карповой и Ж. М. Глозман за возможность участия в этой работе.

собеседников не ограничена с точки зрения вербальных и невербальных средств. Сложность описания танграмм, особенно некоторых из них (самыми сложными для испытуемых обычно являются танграммы [В] — «мышка с бантиком» и [G] — «чеченская танцовщица») дает возможность увидеть изучаемое явление (как то: речевые сбои, оговорки, различного рода синтаксические конструкции) в более «концентрированном» виде, чем при обычном общении.

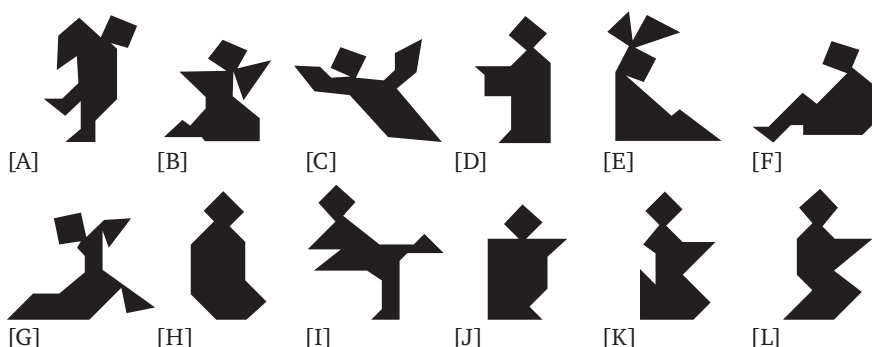


Рисунок 1. Стимульный материал из [Clark, Wilkes-Gibbs 1986].

Наиболее частотные наименования танграмм: [А] — «медведь с топором», [В] — «мышка с бантиком», [С] — «привидение», [D] — «монах с квадратным рукавом», [E] — «зайчик у стенки», [F] — «студент у стенки», [G] — «чеченская танцовщица», [H] — «кокон», [I] — «фигурист», [J] — «кавказец в бурке», [K] — «горбатая японка с хвостиком», [L] — «японка с ногами зигзагом» [Федорова 2014]

Перед обоими участниками, которые сидели за столом напротив друг друга (см. **рис. 2**), но при этом не видели карточки друг друга благодаря специальной перегородке, был одинаковый набор из 6 танграмм, однако перед Инструктором карточки были расположены в одном порядке, а перед Раскладчиком — в другом. Целью Инструктора было объяснить Раскладчику свой порядок, а целью Раскладчика — воспроизвести на своем столе порядок расположения карточек на столе Инструктора.

Эксперимент записывался на видеокамеру Panasonic 4K HC-WX970⁴, которая имеет два объектива и позволяет снимать сразу два видеоплана; результирующий видеофайл устроен по принципу «два в одном» (основной участник — Инструктор — изображен на крупном плане, второй, Раскладчик — во встроенном окне, см. **рис. 2**), что решает проблему синхронизации видеофайлов. У каждой из 8 пар было 2 попытки, в каждой из которых они раскладывали по 6 карточек; при этом первым в качестве Инструктора выступал заикающийся участник, что побуждало его к активному участию в коммуникации.

⁴ Также при записи были использованы две пары очков-айтрекеров Tobii Glasses II Eye Tracker с частотой 50 Гц, однако в настоящей статье собранные данные не приводятся.



Рисунок 2. Скриншот с видеокмеры: сбор корпуса TanStutter_{rus}

3.2. Результаты

В этом разделе мы рассмотрим количественные и качественные результаты исследования.

Общая длительность записей 8 диалогов корпуса TanStutter_{rus} составила около 40 минут, объем — более 2700 словоупотреблений и около 500 реплик. Было зафиксировано 24 случая совместных построений, 17 из которых оказались завершениями и 7 расширениями.

В корпусе TanAdult_{rus}, представляющем собой запись того же эксперимента в парах взрослых без диагностированных нарушений речевой функции, длительностью около 10 часов и объемом более 62 000 словоупотреблений и около 8500 реплик [Федорова 2016], встретилось 74 случая совместных построений. Таким образом, в корпусе TanStutter_{rus} совместные построения встречаются значительно чаще: один случай на 112 словоупотреблений по сравнению с одним случаем на 850 словоупотреблений. При этом отличается и соотношение между двумя основными типами: в корпусе TanStutter_{rus} чаще встречались совместные построения по типу завершения, тогда как в диалогах из корпуса TanAdult_{rus} количество завершений и расширений было примерно одинаковым, см. табл. 1⁵.

⁵ Небольшое количество полученных случаев совместных построений не дает возможности провести полноценный статистический анализ.

Таблица 1. Распределение совместных построений в корпусах TanStutter_{rus} и TanAdult_{rus}

TanStutter _{rus}		TanAdult _{rus}	
Всего: 24		Всего: 74	
Завершений: 17 (70,8%)	Расширений: 7 (29,2%)	Завершений: 39 (52,7%)	Расширений: 35 (47,3%)

Рассмотрим теперь качественный анализ типов совместных построений. Завершения могут быть представлены существительными, которые необходимы в качестве вершины уже начатой именной группы.

- (3) танграмма [B] — «мышка с бантиком»⁶
 А: Второе (е 0.27) ^ мышка с ^ квадратным и ^ треугольным (е 0.33)
 В: *бантиками?* Что-то вроде бантиков, да?
 А кивает

Также завершение может содержать стандарт сравнения, в случае, когда первый говорящий уже произнес сравнительный союз.

- (4) танграмма [B] — «мышка с бантиком»
 А: И сидит при этом, ну как, сидит как (1.63) ну сидит как
 В: *как ромбик там.*
 А: Ну, не-ет, нет.

В случае если первый говорящий произнес сочинительный союз и испытывает трудности, второй говорящий может завершить сочинительную конструкцию за собеседника, как в (5).

- (5) танграмма [A] — «медведь с топором»
 А: Человек ш~аг(&12г)= пыт~ается сделать ш-ш=
 В: шаг.
 А: ш-ш и-и...
 В: *и несет все.*
 А: и у него (0.85) флаг на (0.93) плече

Завершение также может включать в себя предикат для начатого первым говорящим высказывания.

- (6) танграмма [B] — «мышка с бантиком»
 А: Средневе~ковый раб стоит на кол-ленях и (2.15)
 В: *молит о пощаде.*
 А: молит о пощаде прям вот руками.

Последним типом завершений, которые встретились в рассмотренных диалогах, стали завершения с существительными, дополняющими предложную группу, когда первый говорящий уже произнес предлог.

⁶ Здесь и далее: А — участник с заиканием, В — без заикания.

- (7) танграмма [В] — «мышка с бантиком»

А: Он стоит на (1.12)

В: *коленях прямо.*

А: да, прямо.

Среди расширений встретились наречия и предложные группы, использованные для детального описания.

- (8) танграмма [В] — «мышка с бантиком»

А: Ну вот два треугольника в стороны.

В: *сверху, да, не на квадрате.*А: **У курицы** когда вот так.

- (9) танграмма [D] — «монах с квадратным рукавом»

А: ^дальше (э 0.93) человек с (э 0.72) ^кастрюлей что ли, вот так вот, впереди

В: **перед ним? Руками** держит что-то вроде кастрюли?

А: Да.

Также расширение может быть преобразованием уже законченной реплики первого говорящего в сочинительную конструкцию.

- (10) танграмма [F] — «студент у стенки»

А: путеш~ественник очень уставший.

В: ага, *и прислонившийся спиной сидит.*

А: Вот да, да, да.

В качестве последнего из встретившихся типов расширений можно выделить сравнительные конструкции, присоединяемые к уже законченным репликам первого говорящего.

- (11) танграмма [E] — «зайчик у стенки»

А: потом пятая (э 1.11) с^ ушками зайчика (э 7.69)

В: *как птичка что ли ногу который задрал?*

А: Нет^, с ^

В: А-а, сидит такой, на этой.

3.3. Обсуждение результатов

Совместный синтаксис не является часто встречающимся явлением. В [Гренобль 2008] автор пишет: «Частота совместных построений сильно различается в разных интервью. В некоторых интервью они сравнительно часто встречаются (6–8 раз в течение получаса), а в других они вообще отсутствуют». В работе [Szczepiek 2000] в материалах корпуса продолжительностью 40 часов было обнаружено 200 случаев совместных построений. Однако нарушения плавности речи у одного из собеседников побуждают другого увеличивать свой вклад в диалог, в том числе активно использовать совместные построения, что делает исследования совместных построений в парах

«здоровый — заикающийся» перспективной задачей⁷. Однако, как отмечено в [Гренобль 2008], употребление совместного синтаксиса зависит и от индивидуальных предпочтений говорящих, «характера собеседников». Действительно, в одном из восьми диалогов, рассмотренных в рамках настоящего исследования, не было зафиксировано ни одного случая совместного синтаксиса.

Обнаруженное преобладание завершений в исследуемых диалогах может быть объяснено тем, что заикание — нарушение, которое препятствует порождению речи на этапе артикуляции. На момент проявления симптома говорящий может сформулировать завершённую структуру, но из-за выраженного заикания до его собеседника доходит только часть задуманного высказывания, сопровождаемая паузами. Начатая, но незаконченная одним из участников сравнительная или сочинительная конструкция критична для структуры диалога, поэтому второй собеседник использует совместное построение, чтобы восстановить конструкцию.

4. Заключение

Диалог предполагает высокую степень координации между собеседниками, что делает возможным существование совместных построений, используемых говорящими в различных целях. Одной из причин для их использования становятся затруднения в артикуляции, испытываемые одним из участников и побуждающие другого участника, заинтересованного в достижении коммуникативной цели, увеличивать собственный вклад в диалог. В диалогах с заикающимися совместные построения встречаются чаще, чем в норме; среди них преобладают реплики по типу завершения, потому что один из собеседников чаще озвучивает незаконченные конструкции. Исследование заикания в рамках лингвистики представляет значительный интерес, так как предоставляет возможность изучения диалога с позиций кооперации между участниками и зависимости вклада и используемых средств одного участника от вклада другого.

Литература

1. Acton C. (2004), A conversation analytic perspective on stammering: Some reflections and observations, *Stammering Research*, Vol. 1(3), pp. 249–270.
2. Clark H. H., Wilkes-Gibbs D. (1986), Referring as a collaborative process, *Cognition*, Vol. 22(1), pp. 1–39.
3. Grenoble L. A. (2008), Syntax and co-constructed turns in Russian dialogue [Синтаксис I совместное построение реплики в русском диалоге], *Issues in Linguistics [Voprosy jazykoznanija]*, Vol. 1, pp. 25–36.

⁷ Среди близких родственников людей с заиканием, которые часто с ними общаются, существуют определенные сознательные тактики ведения диалога: например, не пребывать заикающегося и говорить медленно самому. Однако задание данного исследования «как можно быстрее и правильнее совместно справиться с поставленной задачей» побуждала их, на наш взгляд, более активно перехватывать роль говорящего.

4. *Fedorova O. V.* (2014), Experimental discourse analysis [Eksperimentalnyy analiz diskursa]. *Yazyki Slavyanskoy Kultury*, Moscow.
5. *Fedorova O. V.* (2016), Perspectives on using referential communication Tangram method in test subjects solving a joint cognitive task [Perspektivy ispolzovaniya metodiki referentsialnoy kommunikatsii Tangrammy v khode resheniya ispytuyemyimi sovmestnoy kognitivnoy zadachi], *Procedures and methods of experimental psychological research. Integration of Academic and University Psychology [Procedury i metody eksperimentalno-psikhologicheskikh issledovaniy. Integratsiya akademicheskoy i universitetskoy psikhologii]*, pp. 565–571.
6. *Lerner G.* (1991), On the Syntax of Sentences-in-Progress, *Language in Society*, Vol. 20, pp. 441–458.
7. *Lerner G.* (1996), On the “semi-permeable” character of grammatical units in conversation: conditional entry into the turn space of another speaker, *Interaction and Grammar*, Cambridge University Press, pp. 238–276.
8. *Lind M., Sønsterud H.* (2014), Co-construction of turns at talk: Active listening or disruptions in conversation with persons who stutter? *Journal of Interactional Research in Communication Disorders*. Vol. 5(2), pp. 141–165.
9. *Ono T., Thompson S. A.* (1995), What can conversation tell us about syntax? Alternative linguistics: Descriptive and theoretical modes, *J. Benjamins*, Amsterdam.
10. *Purver M., Howes C., Gregoromichelaki E., Healey P.* (2009), Split Utterances in Dialogue: a Corpus Study, *Proceedings of the SIGDIAL 2009 Conference: 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, London, pp. 262–271.
11. *Ratner N., Rooney B., Macwhinney B.* (2009), Analysis of stuttering using CHILDES and CLAN, *Clinical Linguistics & Phonetics*, Vol. 10, pp. 169–187.
12. *Skuplik K.* (1999), Co-constructed sentences, definition and an empiric study. [Satzkooperationen. Definition und empirische Untersuchung], *Technical Report Series*, Bielefeld University, Vol. 3.
13. *Szczepek B.* (2000), Formal Aspects of Collaborative Productions in English Conversation, *Interaction and Linguistic Structures (InLiSt)*, available at: <http://www.uni-potsdam.de/u/inlist/issues/17/>.
14. *Zinova Ju. A., Dragoy O. V., Fedorova O. V.* (2011), Experimental study of verbal interaction: language pathology data [Eksperimentalnoe issledovanie rechevogo vzaimodeystviya: dannye yazykovoy patologii], *Vestnik MGU, Ser. 9. Philology*, 4, pp. 167–175.

DOC2VEC OR BETTER INTERPRETABILITY? A METHOD STUDY FOR AUTHORSHIP ATTRIBUTION

Pimonova E. (hpimonova@gmail.com),
Durandin O. (oleg.durandin@gmail.com),
Malafeev A. (amalafeev@yandex.ru)

National Research University Higher School of Economics,
Nizhny Novgorod, Russia

In this work, we perform a method study for the problem of authorship attribution in Russian and English. The datasets used consist of 324 works written in Russian and 207 works in English. We propose a set of text representation models that reflect various linguistic phenomena, in particular, morphological and syntactic ones. One distinctive feature of the proposed models is that they are interpretable. These models are used individually and in combination against a Doc2Vec baseline. For Russian, some of our models outperform Doc2Vec, but this does not happen in the case of English, for various reasons. However, the proposed models can also be used together with Doc2Vec, dramatically improving its performance: by 16.79% in the case of Russian and by 7.2% for English. Additionally, we experiment with two different methods for separating texts into blocks of K sentences (contiguous and bootstrapped) and performed parameter tuning of K . Finally, we conduct a feature importance analysis and show which linguistic markers of author style are the most pertinent for Russian, English and for both these languages. All code used in this work is made freely available to the community¹.

Key words: authorship attribution, author style, text classification, text representation, feature engineering, natural language processing

DOI: 10.28995/2075-7182-2020-19-606-614

DOC2VEC ИЛИ ИНТЕРПРЕТИРУЕМОСТЬ? СРАВНЕНИЕ МЕТОДОВ ОПРЕДЕЛЕНИЯ АВТОРСТВА

Пимонова Е. (hpimonova@gmail.com),
Дурандин О. (oleg.durandin@gmail.com),
Малафеев А. (amalafeev@yandex.ru)

¹ <https://github.com/OlegDurandin/AuthorStyle>

1. Introduction

Authorship attribution is the task of determining the author of a written text based on a set of texts by candidate authors. Automatic algorithms for authorship attribution greatly simplify the solution of these problems and provide reliable and replicable results, which is especially important in criminal law and security matters. Most modern author attribution algorithms are based on formal and statistical models. Despite showing high accuracy in the classification problem, the algorithm results are difficult to interpret. In order to solve this problem, we propose more linguistically-grounded models for solving the attribution problem. We believe that our approach helps identify stylistic markers that can be used as guidelines when attributing a text to a particular author.

This paper is structured as follows. In [Section 2](#), a literature review is given on the problem of authorship attribution. [Section 3](#) describes the proposed four text representation models. We report on the experiments conducted, discuss the results obtained and illustrate the findings of feature importance analysis in [Section 4](#). Finally, [Section 5](#) gives some conclusive remarks.

2. Related Work

Approaches to authorship attribution can be formal or linguistics-based. One of the most common formal models is the n-gram model. Some of the first published algorithms for authorship attribution in Russian used character bigrams [[Khmelev, 2000](#)] and trigrams [[Borisov, et al. 2013](#)]. N-grams are successfully used for author profiling as well; in the work by Litvinova et al. [[2018](#)] on age identification using the first age-annotated corpus for the Russian language, the authors supplemented word n-grams with part-of-speech n-grams.

In the PAN competitions [[Kestemont et al. 2019](#)], the baseline character trigram model was improved on by using variable-length character and word n-grams [[Custodio and Paraboni 2018](#)], as well as by extracting n-grams after text distortion [[Muttenthaler, et al. 2019](#)]. Other n-gram-based models [[Muraier et al., 2018](#)]; [[Bacciu et al., 2019](#)] also showed high accuracy in the PAN competitions.

Other formal approaches to solving the attribution problem are text compression [[Halvani and Graner 2019](#)] and frequency analysis of various text features: word frequencies [[Poddubny, et al. 2010](#)], the number of sentences, text length, character and punctuation frequencies [[Safin and Ogaltsov 2018](#)].

When linguistic (and usually language-dependent) features are used, morphology and syntax are most commonly modeled [[Baayen et al. 1996](#)], [[Rogov et al. 2007](#)], [[Hosseinia and Mukherjee 2018](#)]. Other linguistic methods involve modeling semantics [[Panicheva et al. 2016](#)].

The linguistic approach to authorship attribution is not as widespread as formal language-independent models, but it performs on par with them. While linguistic models are language-dependent, they are often more interpretable. In this work, we propose morphology and syntax models for the Russian and English languages. We believe that these models can help identify reliable stylistic markers that are useful both for computational analysis of author style and for text analysis performed by human experts.

3. Text Representation Models

3.1. Doc2Vec

We used five text representations. The Doc2Vec [Le, Mikolov 2014] model was chosen as the baseline. It is an embedding model for representing sentences, paragraphs or entire documents as vectors. Doc2Vec is known to perform well on various text classification tasks. To improve the quality of the baseline model, we developed two morphological and two syntactic models that differ in representation complexity.

3.2. Simple Morphology and Syntax Models

The so-called ‘simple’ morphological and syntactic models include relative frequencies of parts of speech and syntactic relations present in the text. The number of the morphological features (17, including punctuation, special characters, and foreign words) is the same for Russian and English since we used the language-agnostic UDPipe tool [Straka et al. 2016]. In the simple syntax model, we identified 38 types of syntactic relations such as *nsubj* (subject) and *fixed* (non-free phrase) for Russian and 45 for English.

3.3. Complex Morphology Model

To increase interpretability, we developed the so-called ‘complex’ morphology and syntax models that encompass higher-level language phenomena. The complex morphological model relies on semantic features of words (e.g. the noun “running” denotes a process, etc.). In this model, we used the OpenCorpora markup for the Russian language, since it distinguishes between a larger number of morphological types than Universal Dependencies. The English model still used the UD markup, which resulted in a loss of some features available for Russian (16 versus 10 features).

The semantic attributes used are closely tied to morphological characteristics of words, hence the name of the model. The most ambiguous part of speech in terms of determining the semantic attribute was the noun. We used the “Russian semantic dictionary” by N. Yu. Shvedova for grouping nouns based on their semantic features. An example of a feature under this model is “dynamism”, or the ratio of words with the semantic attribute *process* to all content words. This criterion allows one to determine how much the author is inclined to narration and active change of action.

There were also criteria in our model that took into account the morphological characteristics of the entire text. Some examples are the proportion of verbs in the passive voice or verbs in the past tense to all verbs.

3.4. Complex Syntax Model

Similarly, a complex syntax model was developed, with distinguishing features at the phrase and sentence level. Phrases are categorized according to communication type (coordination, agreement, verb government, or contiguity), structural type (simple and complex phrase), the degree of phrase component unity (syntactically free and non-free phrases) and lexical-grammatical type (nominal, verbal and adverbial).

Each criterion is represented by several types of relations, normalized by the total number of relations. For example, the proportion of syntactically non-free phrases in the text was calculated by the formula: $(flat + fixed + compound) / N$, where *flat* is the number of named entities, *fixed* is non-free phrases, *compound* is compound and composite numerals, *N* is the total number of syntactic relations.

At the sentence level, we considered contracted sentences, vocatives, genitives, various types of one-member sentences and semi-complex sentences. These parameters were calculated taking into account not only the syntactic relations representing the class, but also the morphological characteristics of the words associated with these syntactic relations. For example, indefinite-personal sentences include those that do not have the relations *nsubj* and *csubj* (the connection between the subject and the predicate) coming from the root word. In this case, the root word must also be a verb either in 3rd person plural form, present or future tense, e.g. *govoryat* (\approx people say) or in the form of the plural past tense, like *pogovarivali* (\approx there were rumors). Adapting the originally Russian-based complex syntax model to the English language, we omitted genitive sentences and one-member sentences (except nominative ones), since in English most well-formed sentences have a subject. Thus, we got 28 features for Russian and 22 for English.

4. Experiments

4.1. NLP Framework and Dataset

As mentioned above, we relied on the UDPipe library as the natural language processing framework. The following language models were used: English-EWT and Russian-SyntagRus.

For Russian, we used a corpus that contains 324 works of Russian literature, created by 30 authors spanning XVIII–XXI centuries. For English, we selected 207 works by 34 classical English authors from the Gutenberg Project (gutenberg.org). We divided the entire set of works into training and test sets in such a way that all authors were present, but different works by these authors were used for training and testing, like in PAN competitions. For Russian, the training set included 186 texts, \sim 5M words, while the test set had 138 texts, \sim 2.5M words. For English, there were 137 texts, \sim 15M words in training and 70 texts, \sim 7.4M words in the test dataset.

4.2. Evaluation Method

Following competitions such as PAN [Kestemont et al. 2019], we used classification accuracy as an evaluation metric, that is, the proportion of works whose authors were correctly attributed by the system. Since many literary works in the dataset are quite large, we divide them into blocks. Each of the blocks is classified by the system, then a prediction is made by majority vote as to who authored the entire text. Only final, post-vote predictions are evaluated.

4.3. Experiment Setting

In the experiments, the representations proposed in [Section 3](#) were evaluated against the Doc2Vec baseline, independently and in various combinations. Classic machine learning algorithms were used, namely logistic regression with L1 regularization, random forest, and a linear SVC.

Apart from text representations, we also experimented with some methods for separating texts into blocks of K sentences. The value K is a hyperparameter that affects classification accuracy, so we performed some parameter tuning. We tested two alternative approaches to extracting blocks of text: contiguous (non-overlapping blocks) and bootstrapped (blocks can overlap and are randomly sampled from each text).

4.4. Results and Discussion

We will only list results for logistic regression because it significantly outperformed random forest and linear SVC. Due to size constraints, we will not show the results obtained with each configuration (language, text representation or a combination of text representations, machine learning algorithm, the value of K and text sampling strategy) that we tested, of which there were over 500. Only the best results for each text representation model will be discussed in this section (see [Table 1](#) and [Table 2](#)). Optimized Doc2Vec parameter values were as follows: window = 10, min_count = 3, negative = 5, vector size = 100.

Table 1. Authorship Attribution Classification Accuracy on the Russian-language test set of 138 texts
(Legend: SM—simple morphology, CM—complex morphology, SS—simple syntax, CS—complex syntax, SMS—simple morphology and syntax, CMS—complex morphology and syntax, SCMS—simple + complex morphology and syntax, K—number of sentences per block of text, CB—contiguous blocks, BB—bootstrapped blocks)

Configuration	Classification accuracy
SM, K=350, BB	0.511
CM, K=500, BB	0.430
SS, K=500, BB	0.693
CS, K=500, CB	0.526
SMS, K=500, BB	0.737
CMS, K=350, BB	0.693
SCMS, K=400, BB	0.774
Doc2Vec, K=300, CB	0.613
Doc2Vec + SMS, K=450, CB	0.766
Doc2Vec + SCMS, K=400, CB	0.781

As can be observed for both Russian and English, the complex morphology and syntax models, when used individually, performed much worse than the simple morphology and syntax models, respectively. For Russian, the simple syntax model outperformed the baseline Doc2Vec method, while for English none of the proposed models (or their combinations) surpassed the baseline. This is partly due to the fact

that the complex morphology and syntax models were originally developed for Russian, so they had to be somewhat simplified to accommodate English. Another factor in the higher accuracy of morphosyntactic features for Russian is that, unlike English, Russian is a morphologically-rich language and thus authors have more tools for expression at this level. For English, however, lexical features (as captured by Doc2Vec in our approach) are much more powerful.

Table 2. Authorship Attribution Classification Accuracy on the English-language test set of 70 texts

Configuration	Classification accuracy
SM, K=300, BB	0.6
CM, K=300, BB	0.371
SS, K=400, BB	0.773
CS, K=300, BB	0.586
SMS, K=400, CB	0.787
CMS, K=300, BB	0.671
SCMS, K=400, BB	0.792
Doc2Vec, K=400, CB	0.886
Doc2Vec + SMS, K=300, CB	0.929
Doc2Vec + SCMS, K=400, CB	0.957

Importantly, the proposed text representation models succeed in improving Doc2Vec results. In particular, the combination of all four proposed models (SCMS) resulted in an improvement of 16.79% over Doc2Vec for Russian. For English, the improvement was 7.2%, still very considerable.

5. Feature Importance Analysis

We conducted an analysis of important features in each of the four proposed text representation models to determine which linguistic markers help distinguish one author from another. **Tables 3–5** list such style markers for both Russian and English, as well as language-specific ones.

Table 3. Style markers for both Russian and English

	Simple Morphology	Complex Morphology	Simple Syntax	Complex Syntax
Universal	Function words (conjunctions and particles)	—	conj—relationship between homogeneous members, cc—connection with a means of communication	Homogeneous members
	Noun	—	nsubj—connection between subject and predicate	—
	Punctuation	—	—	Complex structures (participle, adjective and verb-adverb constructions)

Table 4. Russian-specific style markers

	Simple Morphology	Complex Morphology	Simple Syntax	Complex Syntax
Russian	Adverb	Action feature, action descriptiveness (used in the text to describe an action)	advmod, advcl—connection with adjunct	Contiguity linkage
	Noun	Abstractness (used in the text to state abstract notions), objectivity (used in the text to state facts)	nsubj—connection between subject and predicate	Coordination and agreement linkage
	Pronoun	Pronominal replacement (used to replace a noun or noun phrase)	—	—

Table 5. English-specific style markers

	Simple Morphology	Complex Morphology	Simple Syntax	Complex Syntax
English	—	—	flat—relationship between named entities,	Syntactically non-free combinations
	Auxiliary words	—	aux—connection with an auxiliary word	—
	—	Real modality, passive voice	—	—

6. Conclusion and Future Work

Thus, we have proposed and tested a novel approach to authorship attribution that consists in supplementing Doc2Vec with frequencies of parts of speech and syntactic relations, as well as with manually-designed features that reflect larger-scale morphological and syntactic phenomena relevant to author style. This approach is suitable for Russian and English, although we found that lexis has a lot more impact on authorship attribution accuracy in English, while the proposed features at the morphology and syntax levels perform much better on Russian texts.

It is also worth noting that our approach is only suitable for larger chunks of text. The performance with $K < 250$ drops significantly.

For future work, it is possible to test similar approaches on other languages and perform a comparative study of feature importance for a larger set of languages.

References

1. *Baayen R., Halteren H. van, Tweedie F.* (1996), Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, *Literary and Linguistic Computing*, Vol. 11(3), pp. 121–132.
2. *Bacciu A., Morgia M. La, Mei A., Nemmi E. N., Neri V., Stefa J.* (2019), Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features: Notebook for PAN at CLEF 2019, available at: http://ceur-ws.org/Vol-2380/paper_220.pdf.
3. *Borisov L. A., Orlov Y. N., Osminin K. P.* (2013), Authorship attribution by the distribution of letter combination frequencies [Identifikatsiya avtora teksta po raspredeleniyu chastot bukvosochetaniy], *Keldysh Institute Preprints [Preprinty IPM im. M. V. Keldysha]*, Vol. 27, pp. 3–26.
4. *Custódio J. E., Paraboni I.* (2018), EACH-USP Ensemble Cross-domain Authorship Attribution: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_76.pdf.
5. *Halvani O., Graner L.* (2018), Cross-Domain Authorship Attribution Based on Compression: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_90.pdf.
6. *Hosseinia M., Mukherjee A.* (2018), A Parallel Hierarchical Attention Network for Style Change Detection: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_91.pdf.
7. *Kestemont M., Stamatatos E., Manjavacas E., Daelemans W., Potthast M., Stein B.* (2019), Overview of the cross-domain authorship attribution task at {PAN} 2019, *CEUR Workshop Proceedings*, Vol. 2380, pp. 1–15.
8. *Khmelev D. V.* (2000), Recognition of the text author using the Markov chains [Raspoznavaniye avtora teksta s ispolzovaniyem tsepey A.A. Markova], *MSU Bulletin [Vestnik MGU]*, Vol. 9 (2), pp. 115–126.
9. *Le Q., Mikolov T.* (2014), Distributed representations of sentences and documents, in *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pp. 1188–1196.

10. *Litvinova T. A., Sboev A. G., Panicheva P. V.* (2018), Profiling the Age of Russian Bloggers, Proceedings of the 7th International Conference, AINL 2018, St. Petersburg, pp. 167–177.
11. *Murauer B., Tschuggnall M., Specht G.* (2018), Dynamic Parameter Search for Cross-Domain Authorship Attribution: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_84.pdf.
12. *Muttenthaler L., Lucas G., Amann J.* (2019), Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams: Notebook for PAN at CLEF 2019, available at: http://ceur-ws.org/Vol-2380/paper_49.pdf.
13. *Panicheva P. V., Ledovaya Y. A., Bogolyubova O. N.* (2016), Lexical, Morphological and Semantic Correlates of the Dark Triad Personality Traits in Russian Facebook Texts, Proceedings of the AINL FRUCT 2016 Conference, St. Petersburg, pp. 72–79.
14. *Poddubny V. V., Shevelev O. G., Kravtsova A. S., Fatykhov A. A.* (2010), Vocabulary and analytical block of the Style Analyzer [Slovarno-analiticheskiy blok sistemy “Stileanalizator”], 14th Russian Scientific and Practical Conference [Nauchnoye tvorchestvo molodezhi: Materialy XIV Vserossiyskoy N76 nauchno-prakticheskoy konferentsii], Tomsk, pp. 138–140.
15. *Rogov A. A., Sidorov U. V., Solopova A. I., Surovtsova T. G.* (2007), The information-analytical system “SMALT” [Informatsionno-analiticheskaya sistema “SMALT”], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2007” [Kompyuternaya lingvistika i intellektualnyye tekhnologii: Trudy Mezhdunarodnoy konferentsii “Dialog 2007”], Bekasovo, pp. 470–474.
16. *Safin K., Ogaltsov A.* (2018), Detecting a Change of Style Using Text Statistics: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_104.pdf.
17. *Shvedova N. Yu.* (2003), Russian semantic explanatory dictionary. Explanatory dictionary, systematized by classes of words and meanings [Russkiy semanticheskii slovar. Tolkovyy slovar. sistematizirovannyi po klassam slov i znacheniy]. Azbukovnik, Moscow.
18. *Straka M., Hajic J., Straková J.* (2016), UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pp. 4290–4297.

РУССКИЙ ЯЗЫК И КОРПУСНОЕ РАЗНООБРАЗИЕ¹

Пиперски А. Ч. (apiperski@gmail.com)

РГГУ / НИУ ВШЭ, Москва, Россия

В статье даётся обзор применений наиболее известных корпусных ресурсов исследования для русского языка. На примере лингвистических публикаций 2019 года демонстрируется, что русистика недостаточно активно использует возможности, которые открываются перед исследователями благодаря наличию широкого разнообразия корпусов. В качестве примеров демонстрируется, какую пользу различные «не-классические» корпуса могут принести в исследованиях, посвящённых анализу явлений на различных уровнях языка: в морфологии и синтаксисе, в словообразовании и лексике, в частности в исследовании субстандартных языковых явлений, а также в сфере конструкций. Обсуждаются достоинства и недостатки отдельных корпусов с точки зрения интерфейса и удобства для использования в различных аспектах.

Ключевые слова: русский язык, корпусная лингвистика, интернет-корпуса, методология корпусных исследований

DOI: 10.28995/2075-7182-2020-19-615-627

RUSSIAN LANGUAGE AND CORPUS DIVERSITY

Piperski A. Ch. (apiperski@gmail.com)

Russian State University for the Humanities / National Research University Higher School of Economics, Moscow, Russia

This paper discusses the use of most widely-known Russian corpora, namely Russian National Corpus, ruTenTen, General Internet Corpus of Russian, and Araneum Russicum Maximum, for the theoretical study of Russian language. Based on a sample of papers from 2019, I demonstrate that scholars, especially theoretical linguists, tend to ignore the opportunities provided by a wide range of Web corpora, even though these resources are well-known to the NLP community. I present a selection of case studies to show that data from “non-classical” corpora can be used for studying various linguistic phenomena, such as: 1) variation in morphology and syntax; 2) word formation and lexical change; 3) construction grammar. I also claim that the underuse of non-classical corpora is partly due to the fact that they are (perceived as) not quite user-friendly.

Key words: Russian, corpus linguistics, Web as corpus, methodology of corpus-based studies

¹ Исследование выполнено за счет гранта Российского научного фонда (проект № 19-78-10081). Автор сердечно благодарит анонимных рецензентов за высказанные ими замечания.

1. Введение

Современный русский язык обеспечен большим количеством корпусных ресурсов. Разумеется, по их количеству он не может сравниться с мировым лидером — английским языком, но всё же их диапазон чрезвычайно широк. Это приводит к тому, что в разных областях науки используются принципиально различные корпуса, которые порой оказываются незнакомы специалистам из других областей. Так, исследователи, которые занимаются грамматикой русского языка, могут и не знать о существовании корпуса Taiga объёмом 6 млрд словоформ [Shavrina, 2018] или корпуса Omnia Russica объёмом 33 млрд словоформ [Shavrina & Benko, 2019]. В свою очередь, разработчики языковых моделей, которые опираются на большие объёмы текстовых данных, наверняка используют Taiga и Omnia Russica, но едва ли заинтересуются 100-тысячным Хельсинкским аннотированным корпусом (ХАНКО) с его глубокой разметкой [Мустайоки & Копотев, 2003], тем более что этот корпус недоступен для скачивания, что особенно важно для разработки моделей.

Не существует полного перечня русских корпусов, предназначенных для всех разнообразных целей в этом континууме от традиционной русистики до NLP, хотя обзоры корпусных ресурсов делаются достаточно регулярно. Некоторое приближение к такому перечню есть на сайте NLPub², однако этот список явно неполон; из первых работ такого рода можно упомянуть [Резникова & Копотев, 2005], а из более новых — [Zakharov, 2013]; [Хохлова, 2016] и список в учебнике [Копотев, 2014].

В статье будут рассмотрены лингвистические корпуса, которые находят применение в современной теоретической русистике; особое внимание уделяется тому, какие плоды может принести более широкое использование некоторых из них. Разумеется, автор отдаёт себе отчёт, что многим читателям этот обзор покажется бесполезным, поскольку они и так знают, как устроены описанные здесь ресурсы, и пользуются ими — однако кажется, что такие читатели всё же не составляют большинства среди русистов.

2. Использование корпусов в русистике: опыт наукометрического анализа

В англоязычном мире польза корпусов для лингвистов стала очевидной уже в начале 1990-х годов — в частности, благодаря работам [Fillmore, 1992]; [Sinclair, 1991]. В русистике обращение к корпусам стало массовым лишь полтора десятилетия спустя. Так, в лекции В. А. Плунгяна «Почему современная лингвистика должна быть лингвистикой корпусов» [Плунгян, 2009] говорится:

«Все знают, что есть две вещи, нужные, чтобы овладеть языком, это словарь и грамматика, они нужны и теоретику, и практику. Так вот, если в результате моей лекции вы ничего не поймете и не запомните, это

² https://nlpub.ru/Ресурсы#.D0.9A.D0.BE.D1.80.D0.BF.D1.83.D1.81_.D1.82.D0.B5.D0.BA.D1.81.D1.82.D0.BE.D0.B2

совершенно не страшно. Запомните единственное: теперь для овладения языком человеку нужны не две, а три вещи: словарь, грамматика и корпус текстов данного языка».

В 2009 году эта мысль звучала в популярной лекции, но ещё за пять лет до того она была новой для научного сообщества: [Перцов 2006] сообщает, что В. А. Плуныгин говорил это в 2004 году на презентации Национального корпуса русского языка³ (НКРЯ). Впрочем, довольно скоро лингвистам стало ясно, что теоретику нужен корпус не в единственном, а во множественном числе: НКРЯ со всем разнообразием представленных в нём подкорпусов всё же не может служить единственным авторитетным ресурсом [Беликов и др., 2013]. Во втором десятилетии XX века появилось сразу несколько новых значимых ресурсов для русского языка, созданных в парадигме Web as Corpus [Schäfer & Bildhauer, 2013]: ruTenTen⁴, Генеральный интернет-корпус русского языка [ГИКРЯ; Беликов и др., 2013], семейство корпусов Araneum Russicum [Benko, 2014]. Представляется важным оценить, какую роль эти новые, «неклассические» ресурсы играют в современной русистике на фоне единственного широко используемого «классического» корпуса — НКРЯ. Стоит добавить, что ещё одним ресурсом, который обрёл новую жизнь несколько лет назад, стал корпус СинТагРус: он был конвертирован в формат Universal Dependencies [Droganova & Zeman, 2016] и благодаря этому стал широко использоваться в типологических исследованиях.

В качестве инструмента для корпусометрии — или, точнее, для наукометрии, — можно использовать Google Scholar и посмотреть, насколько часто исследователи опираются в своих работах на данные, полученные с помощью тех или иных русских корпусов. Для этого в Google Scholar задавались запросы с названиями и/или характерными фрагментами URL-адресов различных корпусов; поиск вёлся только по публикациям 2019 года. Затем был произведён ручной подсчёт количества найденных публикаций. Полученные результаты представлены в **таблице 1**:

Таблица 1. Число публикаций в Google Scholar за 2019 год с упоминанием различных корпусов русского языка

№	Корпус	Запрос(ы)	Число публикаций
1	НКРЯ	ruscorpora	872
2	ГИКРЯ	ГИКРЯ, GICR corpus	30
3	ruTenTen	ruTenTen, ruTenTen11	25
4	Araneum	“Araneum Russicum”	7

Как можно видеть, доля НКРЯ огромна: именно на этот корпус приходится 872 / 934 ≈ 93% упоминаний. Это и определяет структуру статьи: несколько

³ <http://www.ruscorpora.ru>

⁴ <http://sketchengine.eu>

не умаляя достоинств НКРЯ, я бы хотел поговорить не о нём, а о других, «неклассических» корпусах — о том, какую пользу они могут принести лингвистам в их исследованиях, и указать на то, почему эти ресурсы всё-таки находят недостаточно широкое применение.

3. Примеры (возможных) исследований на «неклассических» корпусах

3.1. Морфология и синтаксис

«Неклассические» корпуса, такие как ГИКРЯ, Araneum и ruTenTen, предоставляют широкий диапазон возможностей для изучения некодифицированных грамматических явлений, в частности — вариативности. Из недавних исследований именно в таком русле выполнена работа [Nesset, 2019], автор которой строит статистические модели, описывающие устройство русских количественных словосочетаний типа *две серьёзные аварии / две серьёзных аварии*. Применение больших корпусов, отражающих устройство современной живой речи, а в случае ГИКРЯ и позволяющих проследить её социальную вариативность, показывает, что во многих случаях мы имеем дело не просто с вариантами, а с вариантами, частотность которых зависит от взаимодействия множества факторов. Образцом такого исследования может служить доклад [Беликов, 2019], в котором автор демонстрирует, что употребление вокализованных и невокализованных вариантов предлогов, в первую очередь *в(о)*, обусловлено региональными и возрастными характеристиками говорящих: к примеру, в Псковской области доля варианта *во Пскове* составляет 46,8%, а по мере удаления от Пскова она снижается: в Санкт-Петербурге — 32,5%, в Москве — 26,0%, а в Сибири — всего 19,3%; у молодёжи доля невокализованных форм выше, чем у более старшего поколения.

Рассмотрим в качестве примера употребление предлога *с / со* перед словами, начинающимися на *щ*. Категориальный подход к этому явлению позволяет принять два возможных решения: либо назвать один из двух вариантов правильным в этом контексте, а другой — неправильным, либо признать их равно допустимыми. Так, первый из этих подходов цитируется со ссылкой на «Орфоэпический словарь русского языка» [Аванесов, 1988], а второй принимается самим автором в работе [Иткин, 2007]:

«Согласно [ОСРЯ: 683], употребление варианта *со* является нормой также перед словами, начинающимися на *щ*. По нашим наблюдениям, это не вполне верно. В спонтанной речи соответствующие конструкции избегаются или вызывают у говорящего неуверенность. Опрос носителей языка выявляет некоторое предпочтение морфа *с* на фоне значительных идиолектных расхождений; нередко информанты признают не вполне удачными обе возможных альтернативы. На наш взгляд, варианты *с* и *со* перед *щ* — *с/со щукой*, *с/со щавелем*, *с/со щек*, *с/со щедрыми дарами* и т. д. — должны трактоваться как равноправные» [Иткин, 2007, с. 81–82]

Однако просмотр частотных списков, извлечённых из большого корпуса — в данном случае использовался *Araneum Russicum Maximum*, — показывает, что мы имеем дело с весьма нетривиальным распределением. В **таблице 2** приведены 82 словоформы, которые суммарно встречаются с предлогами *с / со* более 100 раз в корпусе; они отсортированы по убыванию доли употреблений с вариантом *с* (буквы *е* и *ё* считаются различными; регистр символов не учитывается).

Таблица 2. Частотность вариантов «*с / со X*», где *X* начинается на букву *щ*, по корпусу *Araneum Russicum Maximum*

Словоформа	<i>с / со X</i>	Доля « <i>с X</i> » (%)	Словоформа	<i>с / со X</i>	Доля « <i>с X</i> » (%)
<i>щедрым</i>	959	81	<i>щелочными</i>	1130	45
<i>Щербаковым</i>	136	76	<i>щебенкой</i>	251	41
<i>щебеночным</i>	258	71	<i>щепками</i>	100	39
<i>щедростью</i>	499	70	<i>щупом</i>	374	39
<i>щедрыми</i>	504	70	<i>щитками</i>	199	39
<i>щелевым</i>	265	68	<i>щёткой</i>	346	38
<i>щедрой</i>	539	68	<i>щеткой</i>	2060	38
<i>щечной</i>	153	67	<i>щелятами</i>	170	38
<i>щемящим</i>	200	64	<i>щелчком</i>	1168	38
<i>Щелковским</i>	107	64	<i>щепнем</i>	1703	38
<i>Щелковского</i>	306	63	<i>щётками</i>	171	37
<i>щелкунчиком</i>	115	61	<i>щитовкой</i>	292	37
<i>щадящей</i>	264	61	<i>щетками</i>	973	37
<i>щадящим</i>	492	60	<i>щелями</i>	1196	36
<i>щитовой</i>	110	58	<i>щёлочью</i>	150	35
<i>щупальцами</i>	1071	58	<i>щитка</i>	203	34
<i>Щербинкой</i>	121	57	<i>щитовками</i>	110	34
<i>щитовидкой</i>	2186	57	<i>щелоком</i>	110	33
<i>щадящими</i>	232	57	<i>щуками</i>	117	32
<i>щемящей</i>	262	56	<i>щелячьего</i>	1066	32
<i>щучьей</i>	101	55	<i>щукой</i>	967	32
<i>щавелевой</i>	156	55	<i>щелочами</i>	1022	30
<i>щипцами</i>	422	55	<i>щелочью</i>	862	30
<i>щечками</i>	136	54	<i>щупами</i>	146	30
<i>щетинками</i>	423	54	<i>щитком</i>	585	30
<i>щелевыми</i>	214	54	<i>щавелем</i>	1777	29
<i>щеточкой</i>	432	54	<i>щеками</i>	317	29
<i>щелячим</i>	178	54	<i>щепками</i>	2100	29
<i>щитовым</i>	109	53	<i>щёк</i>	166	28
<i>щелочным</i>	559	53	<i>щепком</i>	4148	27
<i>щелевой</i>	433	51	<i>щепой</i>	189	26
<i>щеточным</i>	148	50	<i>щеки</i>	974	26
<i>щадящих</i>	100	50	<i>щетки</i>	175	25
<i>щепоткой</i>	2846	50	<i>щек</i>	788	24

Словоформа	с / со X	Доля «с X» (%)	Словоформа	с / со X	Доля «с X» (%)
щитовидной	4541	50	щитами	1753	23
щетиной	1602	48	щитом	3626	19
щелчками	174	47	щечкой	281	19
щелью	592	47	щетка	206	17
щелевидными	144	47	щитов	269	13
щелочной	919	46	щита	541	13
щеньчества	104	45	щами	450	7

Можно отметить по крайней мере три фактора, которые явно влияют на доли с и со: частеречная принадлежность, падеж и длина словоформы:



Если построить линейную регрессионную модель, предсказывающую долю «с X» в зависимости от свойств словоформы, мы получим следующую зависимость:

$$f(\text{«с X»}) = 13,8 + 16,7 \cdot \textit{Adjective} + 10,3 \cdot \textit{Instrumental} + 5,3 \cdot \textit{Syllables}$$

Все эти три переменные: *Adjective* (прилагательное: 1, существительное: 0), *Instrumental* (творительный падеж: 1, родительный падеж: 0) и *Syllables* (количество слогов в словоформе) — вносят значимый вклад в определение доли «с X» ($p = 9 \cdot 10^{-6}$, $p = 0,0128$ и $p = 0,0096$ соответственно). Другие протестированные переменные: место ударения при счёте слогов от предлога и частотность словоформы — значимого влияния не оказывают.

Этот пример демонстрирует, что большой неклассический корпус позволяет практически за любой вариативностью увидеть интересные статистические закономерности. На материале НКРЯ такой анализ сделать было бы невозможно, поскольку там насчитывается в общей сложности лишь 2030 примеров «с / со X», где X начинается на букву щ.

3.2. Словообразование и лексика

Использование неклассических корпусов даёт широкие возможности для анализа таких явлений в сфере словообразования и лексики, которые не попадают в корпуса типа НКРЯ. Речь идёт в первую очередь про неологизмы, недавние заимствования и жаргонизмы.

В качестве примера работы, которая для исследования таких явлений опирается на различные типы корпусов, можно привести статью [Пиперски, 2019], посвящённую словам типа *толерасты*, *либерасты*, *флудерасты* и т. п. В ней изложение начинается с примеров из НКРЯ, но там удаётся найти лишь

10 лексем, иллюстрирующих эту словообразовательную модель, поэтому дальнейшее исследование основано на корпусе *Araneum Russicum Maximum*, где обнаруживается 47 таких единиц, встретившихся хотя бы 5 раз. Собранный материал позволяет показать, что такие образования обычно возникают путём наложения *-раст-* на основы, содержащие *р* после второго гласного (точнее: в конце группы согласных после второго гласного, причём длина этой группы может быть равна 1), однако есть и примеры, где *-раст-* уже не требует такого наложения: ср. *единорасты* ‘члены партии «Единая Россия»’, где *р* в производящей основе присутствует, но не в нужной позиции, и *сталирасты* ‘сторонники Сталина’, где *р* в производящей основе нет вовсе. Из этого делается вывод, что *-раст-* постепенно морфологизуется как продуктивный пейоративный суффикс, а не только как элемент языковой игры.

Однако вывод о такой морфологизации на материале *Araneum Russicum Maximum* может быть лишь гипотетическим; чтобы его подтвердить или опровергнуть, следует привлечь материал ГИКРЯ, где тексты снабжены хронологической разметкой. Морфологизация суффикса будет подтверждена, если продемонстрировать, что слова с наложением *-раст-* на *р* появляются раньше, чем слова, в которых *-раст-* выделяется без такого наложения. И действительно, поиск [word="*расты"] в подкорпусе «Живого журнала» в ГИКРЯ демонстрирует, что в более ранние годы слова на *-раст-* образуются от слов с *р* после второго гласного основы, и лишь несколько лет спустя появляются другие типы. Из частотных слов, которые встретились не менее 10 раз, с 2002 года фиксируются *либерасты* и *гитарасты*, с 2003 года — *байдарасты* и *сладострасты*, с 2004 года — *федерасты* ‘сторонники федерализации’, *флудерасты* и *пиарасты*, с 2005 года — *толерасты*, *модерасты* и *питерасты* ‘жители Санкт-Петербурга’, с 2006 года — *поттерасты* и т. д.; все эти слова соответствуют описанной модели. В то же время только в 2005 году в ГИКРЯ появляется слово *восьмидерасты*, в 2006 году — *едрасты* и в 2007 году — *едирасты* и *единорасты*. Разумеется, оценки веб-корпусов могут и должны проверяться другими источниками: например, первые примеры слова *восьмидерасты* в Google Books встречаются уже в 1993 году, но, тем не менее, грамматикализация *-раст-* в суффикс в целом подтверждается данными ГИКРЯ.

Этот пример показывает, что неклассические корпуса являются ценнейшим источником сведений о лексике, в том числе о неологизмах. Это возможно как благодаря объёму этих корпусов, так и благодаря тому, что они нередко включают субстандартные разновидности языка.

3.3. Грамматика конструкций

Неклассические корпуса дают широкий простор для изучения конструкций, в том числе новых и разговорных, которые подчас остаются вне поля зрения исследователей, опирающихся только на тщательно подготовленные к публикации и вычитанные тексты.

Для примера рассмотрим два недавних доклада на очень близкие между собой темы: [Endresen, 2019] и [Урысон, 2020]. В обоих докладах анализируются

конструкции типа с повтором «X и X». Авторы приходят к схожим выводам о возможных значениях этих конструкций, однако оперируют при этом достаточно небольшим числом примеров, либо порождённых ими самими, либо извлечённых из НКРЯ по запросам типа **работа и работа** с конкретными словоформами. В то же время использование неклассических корпусов позволило бы заметно убыстрить извлечение примеров и получить намного более представительную выборку: так, при поиске конструкции «ну(,) X и X» в корпусе *Aganeum Russicum Maximum* находится 5045 примеров, которые можно подвергнуть статистическому анализу (например, установить частотность различных типов конструкций), а наиболее интересные из них проанализировать с подробной интерпретацией.

4. Почему неклассические корпуса не становятся классическими?

Примеры, приведённые в предыдущем разделе, демонстрируют множество полезных применений неклассических корпусов. В этой связи возникает резонный вопрос: если неклассические корпуса так хороши, почему же они настолько редко — в 15 раз реже, чем НКРЯ, как показывает поиск в Google Scholar — используются в практике исследователей русского языка?

Ответ на этот вопрос отчасти заключается в силе привычки: из всех перечисленных ресурсов НКРЯ появился в публичном доступе первым, и пользователи привыкли именно к нему. Ещё одна причина, разумеется, кроется в более широком хронологическом охвате текстов, присущем НКРЯ, а также в том, что содержащиеся в нём тексты воспринимаются как более нормативные и «качественные», хотя это суждение едва ли доказуемо. Схожие соображения высказывает [Нессет 2019] в своей работе:

First, as opposed to the Russian National Corpus, the RuTenTen corpus is not balanced, and it is therefore not clear to what extent it is representative of the Russian language. Second, since the RuTenTen corpus is based on data from the internet, we cannot be sure about the quality of the language in the examples, which may for instance involve machine translated text of poor quality and other “noise”. Third, while the Russian National Corpus provides important metadata such as the time the example was created, the genre of the text, the name and gender of the author, etc., no such metadata are found in the RuTenTen corpus.

Первая претензия — к репрезентативности и сбалансированности корпуса ruTenTen — представляется весьма показательной. В статье [Беликов и др., 2013] после обстоятельного обсуждения понятий репрезентативности и сбалансированности авторы приходят к заключению:

[К]орпус считается представительным и сбалансированным тогда, когда на этот счет имеется негласный договор между его создателями и пользователями.

Хотя со времени публикации этой статьи прошло семь лет, более формального определения репрезентативности и сбалансированности не появилось — но, как видим, создателям веб-корпусов так и не удалось вступить в такой негласный договор с лингвистами. В то же время пользователей не смущает, что, например, 899 из 1808 вхождений слова *удлинённый* в НКРЯ приходится на текст Л. Жильцовой «Веснянки (plecoptera). Группа euholognatha» (2003).

Сверх этого следует отметить ещё одну важную проблему, которая приводит к недостаточному использованию неклассических корпусов: на многих пользователей они производят пугающее впечатление своей технической сложностью (иногда мнимой). Так, для работы с *ruTenTen* или *Araneum Russicum* желательно освоить язык запросов CQL и язык регулярных выражений и понимать, какие возможности они предоставляют. В частности, примеры на *c / co X*, приведённые в [разделе 3.1](#), можно найти по запросу `[lc="co?"] [lc="щ.*"]`, а примеры конструкции «*ну(,) X и X*», упомянутые в [разделе 3.3](#), по запросу `1:[lc="ну"] 2:[lc=","]? 3:[lc!=","] 4:[lc="и"] 5:[] & 3.lc = 5.lc` (1-й токен — *ну*; 2-й токен — запятая, которая может и отсутствовать; 4-й токен — *и*; 3-й токен совпадает с 5-м, но при этом они не являются запятыми). Такие выражения могут показаться громоздкими и непонятными и отбить желание пользоваться неклассическими корпусами. Здесь некоторую помощь может оказать визуальный интерфейс по образцу НКРЯ, который реализован в ГИКРЯ (но не в *ruTenTen* и *Araneum*): он позволяет выбирать необходимые характеристики слов и текстов без составления длинных запросов.

Ещё одна сложность неклассических корпусов для пользователя состоит в том, что большие массивы данных предполагают выгрузку и последующую техническую обработку результатов, что тоже не всегда тривиально и требует умения уверенно конвертировать между собой различные форматы данных: например, выгружать данные в формате *.csv, открывать их в Excel и т. п. Естественно, это отнюдь не непреодолимые преграды, однако исследователям порой даже не приходится в голову, что они могут и должны проводить такого рода обработку материала.

5. Заключение

Рассмотрев, насколько часто используются различные корпусные ресурсы в недавних исследованиях по русистике, мы можем прийти к выводу, что «классический» Национальный корпус русского языка применяется на порядок чаще, чем все прочие «неклассические» корпуса, важнейшими из которых являются Генеральный интернет-корпус русского языка, *ruTenTen* и *Araneum Russicum Maximum*. Недостаточно широкое использование неклассических корпусов вызывает некоторое огорчение, поскольку эти корпуса могут пролить свет на явления, которые трудно или невозможно изучать с помощью традиционных источников. Примеры таких явлений и возможные пути их исследования приведены в разделе 3. Однако приходится признать, что у недостаточного внимания к неклассическим корпусам есть свои причины, в частности их неудобство для технически неподготовленного пользователя. Хочется надеяться, что в будущем параллельно

будут происходить два процесса: адаптация неклассических корпусов к широким кругам пользователей и адаптация пользователей к тому, что работа с корпусом требует определённого уровня технических навыков. Всё вместе это позволит расширить и углубить наши знания о различных явлениях русского языка.

Литература

1. *Benko, V.* (2014), Yet another family of (comparable) Web corpora. In P. Sojka, A. Horák, I. Kopeček & K. Pala (eds.), Text, Speech and Dialogue. 17th International Conference, Brno, Czech Republic, September 8–12, 2014, Springer International Publishing Switzerland, Cham, pp. 35–60.
2. *Droganova, K. & Zeman, D.* (2016). Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies (TR-2016-60; ÚFAL Technical Report). Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.
3. *Endresen, A.* (2019). Ну работа и работа: Russian constructions with reduplication. Novemberseminaret i russisk, UiT Norges Arktiske Universitet. November 29, 2019.
4. *Fillmore, C. J.* (1992). “Corpus linguistics” or “Computer-aided armchair linguistics.” In J. Svartvik (ed.), Directions in Corpus Linguistics, Mouton de Gruyter, Berlin, pp. 35–60.
5. *Neset, T.* (2019), Big data in Russian linguistics? Zeitschrift Für Slawistik, vol. 64(2), pp. 157–174. <https://doi.org/10.1515/slav-2019-0012>.
6. *Schäfer, R., & Bildhauer, F.* (2013), Web corpus construction. Morgan & Claypool. <http://dx.doi.org/10.2200/S00508ED1V01Y201305HLT022>.
7. *Shavrina, T.* (2018). Differential approach to Web corpus construction. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. <http://www.dialog-21.ru/media/4555/shavrinato.pdf>.
8. *Shavrina, T., & Benko, V.* (2019), Omnia Russica: Even larger Russian corpus. In В. П. Захаров (ed.), Труды международной конференции «Корпусная лингвистика—2019», Издательство Санкт-Петербургского университета, Санкт-Петербург, pp. 94–102.
9. *Sinclair, J.* (1991), Corpus, concordance, collocation, Oxford University Press, Oxford.
10. *Zakharov, V.* (2013), Corpora of the Russian language. In I. Habernal & V. Matoušek (Eds.), Text, Speech, and Dialogue. Springer, Berlin; Heidelberg, pp. 1–13. https://doi.org/10.1007/978-3-642-40585-3_1.
11. *Аванесов, Р. И.* (ред.) (1988), Орфоэпический словарь русского языка. Произношение, ударение, грамматические формы, Русский язык, Москва.
12. *Беликов, В. И.* (2019), Статистический взгляд на чередование предлогов. VI Международная конференция «Культура русской речи», Москва.
13. *Беликов, В. И., Копылов, Н. Ю., Пиперски, А. Ч., Селегей, В. П., & Шаров, С. А.* (2013), Корпус как язык: От масштабируемости к дифференциальной полноте. Материалы ежегодной международной конференции «Диалог-2013», с. 84–96.
14. *Иткин, И. Б.* (2007), Русская морфонология, Гнозис, Москва.

15. *Коптев, М. В.* (2014), Введение в корпусную лингвистику, Animedia Company, Prague.
16. *Мустайоки, А., & Коптев, М. В.* (2003), Принципы создания Хельсинкского аннотированного корпуса русских текстов ХАНКО в сети Интернет. Научно-техническая информация, 6, с. 33–37.
17. *Перцов, Н. В.* (2006), К суждениям о фактах русского языка в свете корпусных данных, Русский язык в научном освещении, 1(11), с. 227–245.
18. *Пиперски, А. Ч.* (2019), Экспрессивные неологизмы на *-раст* в русском языке // В. Н. Степанов (ред.), Русская грамматика: Активные процессы в языке и речи. Сборник научных трудов Международного научного симпозиума, ЯГПУ им. К. Д. Ушинского, Ярославль, с. 190–195.
19. *Плунгян, В. А.* (2009), Почему современная лингвистика должна быть лингвистикой корпусов. <https://polit.ru/article/2009/10/23/corpus/>.
20. *Резникова, Т. И., & Коптев, М. В.* (2005), Лингвистически аннотированные корпуса русского языка (Обзор общедоступных ресурсов) // Национальный корпус русского языка: 2003–2005, Индрик, Москва, с. 31–61.
21. *Урысон, Е. В.* (2020), Об одном типе русских предложений тождества (*Платье и платье, ничего особенного*). Конференция в честь 90-летия Ю. Д. Апресяна, Институт русского языка им. В. В. Виноградова РАН, Москва. 04.02.2020.
22. *Хохлова, М. В.* (2016), Обзор больших русскоязычных корпусов текстов. In Компьютерная лингвистика и вычислительные онтологии: Сборник научных статей. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2016), Санкт-Петербург, 22–24 июня 2016 г., Университет ИТМО, Санкт-Петербург, с. 74–77.

References

1. *Avanesov, R. I.* (ed.) (1988), A pronouncing dictionary of Russian. Pronunciation, stress, grammatical forms. [Orfoëpičeskij slovar' russkogo jazyka. Proiznošenie, udarenie, grammatičeskie formy], Russkij jazyk, Moscow.
2. *Belikov, V. I.* (2019), A statistical study of alternations in prepositions [Statističeskij vzgljad na čeredovanie predlogov], 6th International Conference “Culture of Russian Speech”, Moscow. February 23, 2019.
3. *Belikov, V. I., Kopylov, N. Yu., Piperski, A. Ch., Selegey, V. P., & Sharoff, S. A.* (2013), Corpus as language: From scalability to differential completeness [Korpus kak jazyk: Ot masštabiruemosti k differencial'noj polnote]. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2013”, 1, pp. 84–96.
4. *Benko, V.* (2014), Yet another family of (comparable) Web corpora. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (eds.), Text, Speech and Dialogue. 17th International Conference, Brno, Czech Republic, September 8–12, 2014, Springer International Publishing Switzerland, Cham, pp. 35–60.
5. *Droganova, K., & Zeman, D.* (2016). Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies (TR-2016-60; ÚFAL Technical Report). Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.

6. *Endresen, A.* (2019). Ну работа и работа: Russian constructions with reduplication. Novemberseminaret i russisk, UiT Norges Arktiske Universitet. November 29, 2019.
7. *Fillmore, C. J.* (1992). "Corpus linguistics" or "Computer-aided armchair linguistics." In J. Svartvik (ed.), *Directions in Corpus Linguistics*, Mouton de Gruyter, Berlin, pp. 35–60.
8. *Itkin, I. B.* (2007), Russian morphophonology [Russkaja morfonologija], Gnozis, Moscow.
9. *Khokhlova, M. V.* (2016), A survey of big corpora of Russian [Obzor bol'six russkojazyčnyx korpusov tekstov], In *Computational Linguistics and Ontologies [Kompjuternaja lingvistika I vyčislitel'nye ontologii]*, Proceedings of the 19th International Conference "Internet and Modern Society" (IMS-2016), Saint Petersburg, June 22–24, 2016, ITMO University, Saint Petersburg, pp. 74–77.
10. *Kopotev, M. V.* (2014), Introduction to corpus linguistics [Vvedenie v korpusnuju lingvistiku], Animedia Company, Prague.
11. *Mustajoki, A., & Kopotev M. V.* (2003), Principles of construction of the HANCO Corpus on the Web [Principy sozdanija Xelsinksskogo annotirovannogo korpusa russix tekstov HANCO v seti Internet], Scientific and Technical Information [Naučno-texničeskaja informacija], 6, pp. 33–37.
12. *Neset, T.* (2019), Big data in Russian linguistics? *Zeitschrift Für Slavistik*, vol. 64(2), pp. 157–174. <https://doi.org/10.1515/slav-2019-0012>.
13. *Pertsov, N. V.* (2006), Judgments on the facts of Russian language in the light of corpus data [K suždenijam o faktax russkogo jazyka v svete korpusnyx dan-nyx], *Russian Language and Linguistic Theory [Russkij jazyk v naučnom osveščeenii]*, issue 1(11), c. 227–245.
14. *Piperski, A. Ch.* (2019), Expressive neologisms in -rast in Russian [Èkspressivnye neologizmy na -rast v russkom jazyke], In V. N. Stepanov (ed.), *Russian grammar: Active processes in language and speech [Russkaja grammatika: Aktivnye process v jazyke i reči]*, Yaroslavl State Pedagogical University named after K. D. Ushinsky, Yaroslavl, pp. 190–195.
15. *Plungian, V. A.* (2009), Why modern linguistics must be corpus-based [Počemu sovremennaja lingvistika dolžna byt' lingvistikoj korpusov], <https://polit.ru/article/2009/10/23/corpus/>.
16. *Reznikova, T. I., & Kopotev, M. V.* (2005), Linguistically annotated corpora of Russian: A survey of publicly available resources [Lingvističeski annotirovannye korpusa russkogo jazyka (Obzor obščedostupnyx resursov)], In *Russian National Corpus: 2003–2005 [Nacional'nyj korpus russkogo jazyka: 2003–2005]*, Indrik, Moscow, pp. 31–61.
17. *Schäfer, R., & Bildhauer, F.* (2013), Web corpus construction. Morgan & Claypool. <http://dx.doi.org/10.2200/S00508ED1V01Y201305HLT022>.
18. *Shavrina, T.* (2018). Differential approach to Web corpus construction. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"*. <http://www.dialog-21.ru/media/4555/shavrinato.pdf>.

19. *Shavrina, T., & Benko, V. (2019), Omnia Russica: Even larger Russian corpus.* In Б. П. Захаров (ed.), In Proceedings of the International Conference “Corpus Linguistics’2019”, Saint Petersburg State University, Saint Petersburg, pp. 94–102.
20. *Sinclair, J. (1991), Corpus, concordance, collocation,* Oxford University Press, Oxford.
21. *Uryson, E. V. (2020), On a certain type of Russian equative constructions (Plat’je i plat’je, ničego osobennogo ‘Just a dress, nothing more’) [Ob odnom tipe russkix predloženíj toždestva (Plat’je i plat’je, ničego osobennogo)].* Conference on the occasion of the 90th birthday of Yu. D. Apresyan. V. V. Vinogradov Institute of Russian Language, Moscow, February 04, 2020.
22. *Zakharov, V. (2013), Corpora of the Russian language.* In I. Habernal & V. Matoušek (Eds.), *Text, Speech, and Dialogue.* Springer, Berlin; Heidelberg, pp. 1–13. https://doi.org/10.1007/978-3-642-40585-3_1.

«А ТОТ ПЕРОВСКОЙ НЕ ДАЛ ВСЛАСТЬ ПОСПАТЬ»: ПРОСОДИЯ И ГРАММАТИКА АНАФОРИЧЕСКОГО ТОТ В ЗЕРКАЛЕ КОРПУСНЫХ ДАННЫХ

Подлесская В. И. (vi_podlesskaya@il-rggu.ru)

РГГУ, Москва, Россия

DOI: 10.28995/2075-7182-2020-19-628-643

“A TOT PEROVSKOJ NE DAL VSLAST’ POSPAT’”: PROSODY AND GRAMMAR OF ANAPHORIC TOT THROUGH THE LENS OF CORPUS DATA

Podlesskaya V. I. (vi_podlesskaya@il-rggu.ru)

Russian State University for the Humanities, Moscow, Russia

Based on data from the Russian National Corpus and the General Internet-Corpus of Russian, the paper addresses syntactic, semantic and prosodic features of constructions with the demonstrative TOT used as an anaphor. These constructions have gained some attention in earlier studies [Paducheva 2016], [Berger, Weiss 1987], [Kibrik 2011], [Podlesskaya 2001], but their analysis (a) covered primarily their prototypical uses; and (b) was based on written data. The data from informal, esp. from spoken discourse show however that the actual use of these constructions may deviate considerably from the known prototype. The paper aims at bridging this gap. I claim (i) that the function of TOT is to temporarily promote a referent from a less privileged discourse status to a more privileged one; and (ii) that TOT can be analyzed on a par with switch reference devices in the languages where the latter are grammatically marked (e.g. on verb forms). The following parameters of TOT-constructions are discussed: syntactic and semantic roles of TOT and of its antecedent in their respective clauses, linear and structural distances between TOT and its antecedent, animacy of the maintained referent. Special attention is paid to the information structure of the TOT construction: I give structural and prosodic evidence that TOT never has a rhematic status. The revealed actual distribution of TOT (a) adds to our understanding of cross-linguistic variation of anaphoric functions of demonstratives; and, hopefully, (b) may contribute to further developing computational approaches to coreference and anaphora resolution for Russian, e.g. by improving datasets necessary for this task.

Key words: anaphora, switch reference, Russian, prosody

1. Постановка задачи

Предметом данной работы являются конструкции, в которых средством анафорической отсылки выступает указательное местоимение ТОТ в субстантивном употреблении¹. Канонический пример такой конструкции представлен в (1), где антецедентом местоимения ТОТ является именная группа (*своему осмотрительному другу*) *Фиме Королёву*:

- (1) *Тем временем Коля рассказывает обо всём своему осмотрительному другу Фиме Королёву. И тот убеждает Колю затаиться.*

Указательным местоимениям в функции анафоров в русском и ряде других языков посвящена значительная литература, см. [Кибрик 1987], [Крейдлин, Чехов 1988], [Падучева 1985], [2016], [Подлеская 2001], [Селезнев 1987], [Berger, Weiss 1987], [Kibrik 2011], [Patel-Grosz, Grosz 2017] inter alia. Однако в описании этого феномена сохраняется ряд белых пятен.

Во-первых, чаще всего имеющиеся работы опираются в анализе на канонические случаи таких конструкций, хотя большинство исследователей признает, что отклонения от прототипа, не нарушающие грамматичности, встречаются (по оценке [Kibrik 2011:329] канонические случаи составляют около 90% всех употреблений анафорического ТОТ). Проиллюстрирую один класс отклонений от канона. Прототипически, антецедентом местоимения ТОТ является лицо. Однако в ряде случаев приемлемой оказывается и отсылка к предметному референту (*гора* в (2), *камфара* в (3)):

- (2) *Ей показали гору / на которую она собиралась идти / та была вся в облаках.*
(3) *Если бы он мимоходом глянул на обложку нашего журнала, то та показалась бы ему скоплением расцвеченных точек.*

Во-вторых, имеющиеся работы рассматривают почти исключительно структурные свойства конструкций ТОТ, оставляя в стороне их просодические свойства. Между тем, понимание того, как устроена коммуникативно-просодическая структура таких конструкций, позволило бы пролить свет на ряд дистрибутивных и функциональных ограничений на их использование.

Учитывая сказанное, я ставлю перед собой две взаимосвязанные задачи, которые попытаюсь решить на корпусном материале:

- А. Систематизировать прототипические структурные ограничения на употребление анафора ТОТ и установить, какие из этих ограничений могут нарушаться в реальном узусе. Этому будет посвящен **раздел 2**.
Б. Описать коммуникативно-просодические свойства конструкций с ТОТ и выявить возможную связь этих свойств со структурными ограничениями на их употребление. Эта задача решается в **разделе 3**.

¹ Работа выполнена при поддержке РФФ, грант 17-18-01184. Я очень признательна О. В. Федоровой и Я. Г. Тестельцу, которые дали себе труд прочесть первоначальную версию статьи, высказали ценнейшие соображения и дополнили мой библиографический список очень важными работами.

Для анализа структурных свойств использовался основной корпус НКРЯ (ruscorpora.ru) и, частично, данные ГИКРЯ (webcorpora.ru), для анализа просодии использовался мультимедийный подкорпус НКРЯ (МУРКО). Для инструментального изучения просодии из видеофайлов выдачи МУРКО извлекались аудиофайлы и подавались на вход анализатора PRAAT [Boersma, Weenink 2014]. Для вынесения суждений о приемлемости примеров я пользуюсь оценками носителей языка, в том числе, и своими собственными оценками. Примеры, полученные методом интроспекции, помечены аббревиатурой «ВИП» после номера, примеры из ГИКРЯ также имеют соответствующую помету. Все примеры без пометы взяты из НКРЯ, они приводятся в той графической форме, в которой они даны в корпусе, за исключением отдельных случаев, когда я вношу собственную просодическую разметку — они ниже специально оговариваются.

2. Прототипические структурные свойства конструкций с анафором ТОТ и возможные отклонения от прототипа

В прототипических случаях на конструкции с анафором ТОТ накладываются следующие ограничения (наиболее подробно эти ограничения исследованы в [Кибрик 1987] и уточнены в [Подлесская 2001]):

- 1) ТОТ и его антецедент расположены в соседних клаузах;
- 2) Клауза, предшествующая ТО, обозначает ситуацию с двумя участниками-лицами: первый участник является актором (наиболее активным участником) и кодируется подлежащим, второй участник кодируется прямым или косвенным дополнением, расположен (в соответствии с базовым порядком слов) правее первого и является антецедентом анафора ТОТ.
- 3) ТОТ в своей клаузе является актором и кодируется подлежащим.

Указанные ограничения не являются абсолютными и так или иначе могут нарушаться в реальном узусе. Рассмотрим их подробнее.

2.1. В прототипическом случае анафор ТОТ и его антецедент расположены в соседних клаузах, причем эти клаузы могут быть:

- (а) синтаксически подчинены одна другой таким образом, что вторая подчинена первой, (4а);
 - (б) синтаксически подчинены одна другой таким образом, что первая подчинена второй, (4б);
 - (в) сочинены с использованием союза, (4в)
 - (г) соположены асиндетически, (4г):
- (4) (а) *и пошёл он к царю / Ивану Грозному / российскому самодержцу / чтобы тот разрешил ему сочетаться законным браком с дворовой девкой / Парашей*

- (б) *Когда настойчивые средневековые власти прижали брадобрея, тот раскололся.*
- (в) *Валентин Михайлович обратился с просьбой к бармену, но тот сказал, что шашлыки у них готовят только после девяти вечера.*
- (г) *Илья Иосифович укорял Павла Алексеевича в приспособленчестве, тот пытался оправдываться*

Требование непосредственного соседства двух клауз в реальном узусе может не выполняться. Например, в диалоге ТОТ регулярно используется для переключения говорящих в чередующихся репликах. При этом реплика, предшествующая ТОТ и отделяющая его от своего антецедента, может включать протяженную последовательность клауз, как в примере ниже, где имеется референциальная цепочка *редакторша > она > та*, каждый из элементов цепочки удален от соседнего более, чем на одну клаузу

- (5) *И вот однажды Дурново приходит / а там была какая-то редакторша / он её называл / её фамилию / но я забыла / какая-то известная / вы можете тоже знать. Она была такая коммунистка / очень такая м... выдержанная / целеустремлённая и / вообще / знающая и ориентирующаяся во всём. Ну вот значит / приходит Дурново и говорит / «На улице траурные флаги / умер какой-то Сэн Катаяма / по суффиксу / должно быть / японец». «Так / а та / значит / возмутилась / как это по суффиксу / вообще / Сэн Катаяма / понимаете ли».*

Сходным образом устроен и пример (6), но тут дополнительно в референциальную цепочку *поэтесса > своей > её > Вашей > Вы > Вашего > та* включено местоимение второго лица, обращенное к собеседнику, в составе цитаты, реализованной в форме прямой речи:

- (6) *А вот однажды к ней пришла поэтесса и прочитала большую поэму о своей любви к убитому на войне мужу. Выслушав ее / Анна Андреевна сказала: «Главный недостаток Вашей поэмы / что по существу / Вы сейчас любите другого человека / о нем Вы пишете в этой поэме / и только прикрываетесь фигурой Вашего убитого мужа». И та сказала: «Это правда».*

Большое расстояние между ТОТ и его антецедентом может наблюдаться и в нарративе. Так в (7) антецедентом местоимения ТА является *Ольга*, между антецедентом и анафором — шесть предложений, два из которых мультиклаузальны! Разумеется, семантически расстояние не так велико, поскольку референт, к которому отсылает ТА (*Ольга*), имплицитно присутствует в упоминаниях ее места проживания — *там и туда*.

- (7) *И потом это второе сообщение — о том, что Сонька ходила к Ольге ... Мне самому казалось странным, что этот факт поразил меня едва ли не так же, как новость про отца и Соньку. Я сам не мог понять, что со мной творится. Конечно, это действительно было странно. Что она там забыла? Зачем было туда ходить? Выяснить отношения? Просить, чтобы та уступила? Ни за что не поверю! Это совсем, совсем не похоже на Соньку.*

Таким образом, приходится признать, что минимальное линейное расстояние между клаузой с ТОТ и клаузой с его антецедентом не является самоцелью: главное требование — чтобы антецедент адекватно восстанавливался адресатом. Разумеется, чем ближе друг к другу расположены анафор и антецедент, тем легче происходит идентификация антецедента, но и при значительном расстоянии между ними лингвистический и\или экстралингвистический контекст могут помочь успешной идентификации.

2.2. В прототипическом случае в клаузе, предшествующей ТОТ, выведены два участника, ср. в (4а) — он и царь, в (4б) — власти и брандмейстер, в (4в) Валентин Михайлович и бармен, в (4с) — Илья Иосифович и Павел Алексеевич. Первый из пары является актором (наиболее активным участником) и кодируется подлежащим, второй кодируется прямым или косвенным дополнением, первый из пары расположен в клаузе левее, как и положено подлежащему при нейтральном порядке слов. Местоимение ТОТ отсылает к тому участнику, который расположен правее и имеет более низкий ролевой и грамматический статус в клаузе. Главное отличие анафора ТОТ от анафора ОН состоит в том, что у последнего нет этого ограничения, поэтому в предложении (8а) антецедентом может быть только Коля, а в предложении (8б) — и Петя, и Коля²:

- (8) (а — ВИП) *Петя_i встретил Колю_j. Тот^{*i/j} был не в духе и не поздоровался*
 (б — ВИП) *Петя_i встретил Колю_j. Он_{i/j} был не в духе и не поздоровался*

Указанное ограничение на выбор антецедента приводит к тому, что ТОТ становится обязательным в случае референциального конфликта [Кибрик 1987], [Кибрик 2011]. Однако ТОТ вполне допустимо и вне конфликта, например, при несовпадении двух потенциальных антецедентов по роду и\или числу, как (8в); согласно [Кибрик 2011:345], доля случаев с референциальным конфликтом составляет 28% обследованной им выборки предложений с анафорическим ТОТ:

- (8) (в — ВИП) *Петя_i встретил Машу_j. Та^{*i/j} была не в духе и не поздоровалась*

Требование, чтобы ИГ — антецедент ТОТ — располагалась правее и имела более низкий ролевой и грамматический статус, чем конкурирующая ИГ, в реальном узусе может не выполняться и в своей линейной, и в своей структурной части. Во-первых, антецедент ТОТ может оказаться левее подлежащего:

- (9) *Колю Гнидюка, как тот ни стремился попасть на торжество по случаю «помолвки», Шевчук не пригласил.*
 (10) *Когда консула Украины по Уралу и Западной Сибири останавливают тюменские гаишники на предмет красных номеров, тот шутит, что работает в пожарной охране*

² По свидетельству А. П. Майорова, линейное ограничение на антецедент местоимения ТОТ существовало в русском деловом дискурсе уже в XVIII в.: «Если в контексте речь идет о двух объектах, то среди них первый антецедент в дальнейшем имеет отсылку посредством местоимения *оной*, а второй — с помощью местоимения *тотъ*» [Майоров 2004: 236].

Во-вторых, антецедентом ТОТ (в примере ниже — ТА) может оказаться подлежащее-актор, ср. *соседка* в (11):

- (11) *соседка, нагнувшись под парту, больно ущипнула её за ногу. Тогда Таня поняла, что и под локоть та толкнула её нарочно, и заплакала.*

Как видим, в (11) рассматриваемое ограничение нарушено одновременно и в линейной, и в структурной части: антецедент местоимения ТОТ (*соседка*) имеет ролевой статус актора, кодируется подлежащим и расположено левее ИГ, кодирующей конкурирующего референта (*её=Таня*). Более того, вполне приемлемыми оказываются случаи, когда антецедент местоимения ТОТ содержится в первой клаузе и является в ней актором-подлежащим, за ней (непосредственно или на некотором расстоянии) следует клауза, содержащая конкурирующую ИГ, и лишь затем — клауза, содержащая местоимение ТОТ. Так, в (12) антецедентом ТОТ является *сосед*, конкурирующей ИГ — *Лохматый*. В (13) антецедентом ТОТ является *кот*, референт конкурирующей ИГ — *говорящий*, но сама конкурирующая ИГ опущена, так как местоимение 1-ого лица легко восстанавливается из контекста:

- (12) (ГИКРЯ) *Опять приходил этот сосед. Лохматый (это такой позывной у Домовенка) как обычно, что-то ляпнул и тот убежал, держась за задницу*

- (13) (ГИКРЯ) *Кот сказал, что до смерти боится закрытых дверей. Открыл. Тот сказал, что обожает сидеть на пороге. Закрыл.*

В-третьих, даже если конкурирующая ИГ и антецедент местоимения ТОТ оказываются внутри одной клаузы, они не обязательно разведены как два разных актанта одного глагола. Допускается, например, чтобы ИГ — антецедент ТОТ была вложена в ИГ — подлежащее. Так, в (14) антецедентом ТОТ является *Бертенев*, а подлежащим — ИГ *кто из партнёров Бертеньева по преферансу*:

- (14) *Не догадался ли кто из партнёров Бертеньева по преферансу, что тот ведёт не честную игру.*

Конкурирующая ИГ может и вовсе физически отсутствовать. Мы уже видели выше в примере (13), что конкурирующая ИГ может эллиптироваться. Потенциальный конкурирующий референт может быть «зашит» в обобщенно-личной конструкции:

- (15) *Стоит только побрызгать немножко керосина на хвост Наполеону / как тот сдаётся в плен.*

Наконец, не только эксплицирующая референт ИГ, но и сам конкурирующий референт может вообще отсутствовать:

- (16) (ВИП) *На диване осталась лежать шляпка Марии Петровны, которую та, убегая, забыла надеть*

Таковы, очень кратко, возможные отклонения от прототипа, связанные со статусом антецедента местоимения ТОТ. Рассмотрим теперь отклонения, связанные со статусом самого местоимения.

2.3. В прототипическом случае местоимение ТОТ в своей клаузе имеет максимально высокий ролевой и грамматический статус: обычно это подлежащее и актор. Соответственно, будучи подлежащим и актором, ТОТ тяготеет к крайней левой позиции в клаузе. Однако в реальном узусе и ролевое, и грамматическое, и линейное ограничения могут нарушаться, причем одновременно, ср. косвенные дополнения с *той* в (17), с *той* в (18). В (17) нарушено еще и рассмотренное выше ограничение на антецедент, так как антецеденты местоимения ТОТ (*женщина с инфарктом* и *она*) — подлежащие в своих клаузах:

(17) *И п... после этого она ничего не говоря / несмотря на то / что двадцать лет они дружили и вообще / ну / были как сёстры / она просто / ничего не объявляя / может перестать с той разговаривать.*

(18) *есть у нее приятельница какая-то или кормилица что ль ее дочерей / женщина с инфарктом / вот с такими ногами / со слоновой болезнью. И вот когда она совсем слегла / Нина Григорьевна в своем состоянии ездила к той мыть посуду.*

В примере (19) ниже нарушение менее глобальное — линейно местоимение ТУ (аккузатив от ТА) занимает, как в прототипических случаях, крайнюю левую позицию, но это — в отличие от прототипа — не подлежащее и не актор, а прямое дополнение и пациенс:

(19) (ГИКРЯ) *Я девушка, мне 16, а моей лучшей подруге уже около 20. Она ведет богемный образ жизни, и мне раньше хотелось стать похожей на нее. Но недавно она рассказала про свою знакомую. Ту лишили родительских прав за наркотики, а моя подруга стала ее выгораживать.*

Таким образом, все рассмотренные нами прототипические грамматические ограничения на употребление анафора ТОТ преодолеваются в узусе без нарушения правильности\приемлемости. Это приводит к предположению, что сами эти ограничения должны формулироваться не исключительно в грамматических терминах, а многофакторно — с подключением критериев, связанных с прагматикой, семантикой и актуальным членением.

2.4. Наиболее полно многофакторный подход к описанию анафорических конструкций с ТОТ разработан А. С. Чеховым и Г. Е. Крейдлиным [Чехов, Крейдлин 1988]. В качестве центрального для правильности таких конструкций они выдвигают следующее требование (в изложении несколько менее формализованном, чем у авторов): антецедент местоимения ТОТ должен быть в составе «своей клаузы» ремой (или частью ремы), а само местоимение в составе «своей» клаузы — темой (или частью темы). Такое требование выглядит очень продуктивным и, как кажется, позволяет объяснить довольно много случаев,

когда приемлемыми и уместными оказываются высказывания, нарушающие прототипические грамматические ограничения. Вместе с тем, подключение требования к актуальному членению конструкции без систематического подключения просодической аргументации приходится принимать с некоторой осторожностью. Во-первых, работая только с письменным текстом, мы можем лишь гипотетически реконструировать его наиболее вероятную линейно-акцентную структуру и часто не имеем достаточно оснований, чтобы решить, выполняется сформулированное выше требование или нет. Во-вторых, даже если допустить такой реконструктивный подход, имеются случаи, когда данное требование не выполняется, а текст оказывается правильным и уместным. Например, антецедент может не иметь статуса ремы — это особенно касается случаев, где антецедент местоимения ТОТ располагается не в непосредственно предшествующей клаузе, а далеко в предтексте и встроен в протяженную референциальную цепочку. Так, в примере (6), по-видимому, ни одно из упоминаний антецедента местоимения ТА в цепочке *поэтесса > своей > её > Вашей > Вы > Вы > Вашего > та* не формирует ремы — оно либо входит в состав нерасчлененного сегмента, либо входит в тему. Трудности возникают и в примерах типа (14) или (16), где вложенные генитивные группы (*Бертеньева, Марии Петровны*) вряд ли могут претендовать на статус ремы. Тематический статус самого местоимения ТО тоже не всегда оказывается «безупречен». Так, в примере (20) ниже доказать, что местоимение ТА формирует тему, весьма проблематично:

(20) (ВИП) *Ольге очень нужна была помощь Ивана. Но он не приехал, как та ни упрасивала.*

По-видимому, проще доказать более мягкое утверждение: местоимение ТОТ в «своей» клаузе не может быть ремой. Косвенным структурным доказательством этого «мягкого» ограничения является тот факт, что ТОТ не может быть помещен в позицию, характеризующуюся признаками тематичности. Во-первых, местоимение ТОТ в анафорическом употреблении, не будучи формально клиткой, обладает следующим характерным для клитик свойством — оно не сочетается с большинством фокусирующих частиц (см. подробнее [Подлеская 2001]):

(21) (ВИП) *Петя попросил Машу приехать, потому что помочь могла только она/*та.*

Во-вторых, по наблюдениям Я. Г. Тестельца (р.с.) местоимение ТОТ в анафорическом употреблении обладает еще рядом типичных для клитик свойств — оно не сочетается с отрицанием, не входит в ни сочинительную, ни в комитативную связь:

(22) (ВИП) *Я просил Машу, но в результате не она (*не та), а Саша откликнулась. Я сообщил Маше, и она (*и та) и ее мама пришли. Я сообщил Маше, и она (*и та) с Леной пришла //пришли*

В-третьих, местоимение ТОТ нетерпимо к заключительной позиции:

(23) (ВИП) *Сначала Ольга долго и безуспешно добивалась внимания Ивана. Потом, наоборот, Иван сделал предложение, но на этот раз отказала она/*та.*

Единственным заметным исключением являются конструкции с прямой речью с постпозитивной авторской ремаркой:

(23) (а) *Черт с тобой, бери!* — сказал тот.

Но это особая конструкция, в которой инверсия подлежащего и сказуемого в постпозитивной авторской ремарке обязательна для любого подлежащего — как выраженного местоимением, так и полной именной группой, ср.:

(23) (б — ВИП) *Черт с тобой, бери!* — сказал Иван.

Обычно такое подлежащее реализуется атонически или со слабым нисходящим тоном. Т. Е. Янко (р.с.) предположила, что ремой в таких высказываниях является собственно прямая речь, а авторская ремарка является вынесенной вправо атонической темой со статусом близким к парентезе. Таким образом, возможный нисходящий тон на подлежащем маркирует не рему, а финальность, т.е. завершенность высказывания. Косвенным подтверждением того, что инвертированное подлежащее в таких конструкциях с прямой речью не является ремой, может служить тот факт, что оно не может присоединять фокусные частицы:

(23) (в — ВИП) *Черт с тобой, бери!* — сказал *именно тот / *именно Иван.

Таковы структурные свидетельства в пользу «мягкого» утверждения: просодическое поведение анафорического ТОТ несовместимо с рематическим статусом. Ниже, в разделе 3 я постараюсь показать, что просодия также свидетельствует в пользу «мягкого» утверждения.

3. Коммуникативно-просодические свойства анафора ТОТ

Для изучения просодических свойств анафорического ТОТ в корпусе МУРКО были отобраны все случаи, удовлетворяющие запросу «*тот* (m | f) на расстоянии 1 от V», т.е. форма местоимения ТОТ мужского или женского рода, за которой следует глагольная словоформа. Форма среднего рода была изначально отсечена, чтобы уменьшить «шум» в выдаче, связанный с соотносительным, а не анафорическим ТО. Кроме того, также для уменьшения «шума», выборка была ограничена случаями, где анафорическое ТО соседствует с глаголом непосредственно. Всего было обнаружено 56 примеров интересующей нас конструкции. Перцептивно и с помощью анализатора PRAAT я проверила, имеется ли в этих примерах акцент на местоимении, и если акцент имеется, то какой вариант движения тона реализуется. Выяснилось, что из отобранных 56 примеров в 50 случаях ТО было употреблено атонически, во всех этих случаях ТО не формирует отдельной коммуникативно-просодической составляющей. Рассмотрим пример (24а). В нем (и двух следующих «просодических» примерах) движение тона показано иконически с помощью косых черт перед акцентированным словом, а границы просодических составляющих — с помощью квадратных скобок; во избежание путаницы я удалила в «просодических» примерах косые черты, которые в собственной разметке НКРЯ используются для обозначения

пауз, а также знаки словесных ударений. Как мы видим, оба вхождения анафорического местоимения — ТА и ТОТ — произносятся без акцента; напротив, их антецеденты (*Лидии* и *Василькову*), как и в большинстве прототипических случаев, являются в своих клаузах носителями главного фразового (рематического) акцента, который реализуется с падающим тоном:

- (24) (а) [*Входит \горничная*] [*и подаёт счёт \Лидии*] [*та показывает на \мужа*].
 [*Горничная подаёт счёт \Василькову*] [*тот /пробежав его*] [*кивает головой на \жену*].

В принципе, этот же сегментный материал мог бы быть реализован и таким образом, что местоимения ТА и ТОТ были бы выделены в отдельную коммуникативно-просодическую составляющую со статусом темы или контрастной темы и реализовались бы с одним из стандартных вариантов подъема тона — например, по типу ИК3 или ИК4 в номенклатуре интонационных конструкций по [Брызгунова 1982]:

- (24) (б — ВИП) ...[/*та*] [*показывает на \мужа*] [/*тот*], [/*пробежав его*], [*кивает головой на \жену*].

Однако такая тактика оказывается значительно более редкой, чем безакцентная: местоимение было акцентировано в шести случаях из пятидесяти шести. Все эти шесть случаев можно квалифицировать коммуникативно как случаи темы или контрастной темы. В пяти из них акцентированное местоимение реализуется с подъемом тона. Так, в (25) акцентированное местоимение ТА образует отдельную просодическую составляющую и реализуется с подъемом тона по типу ИК3, а следующая за ним рематическая составляющая *затуманилась* реализуется с падением тона по типу ИК2 — в узком временном интервале и широком частотном диапазоне. Интонограмма клаузы с ТА и предшествующей клаузы с антецедентом местоимения (*Мурашкину*) показана на **рисунке 1**:

- (25) (Павел Васильевич сделал усилие, чтобы разомкнуть напряжённые слипающиеся веки, зевнул, не раскрывая рта,) [*и поглядел на \Мурашкину*]. [/*Та*] [*\затуманилась*] [*закачалась в его /глазах*] [*стала /трехголовой*] [*и упёрлась головой в \потолок*].

В одном случае из шести акцентированное местоимение, формирующее отдельную просодическую составляющую, реализовалось с ровным тоном, см. (26) и **рисунок 2**:

- (26) (Уважаемый человек спросил «в пять?»). Серёжа как переводчик перевёл своему начальнику) [*уважаемому \человеку*]. [–*Тот*] [*спрашивает «в /пять?»*]

На ТОТ акцент сильный, ровный, в низком регистре, он может рассматриваться как тематический, адаптированный, по общему для русского языка правилу, к последующему подъему, маркирующему рему вопроса.

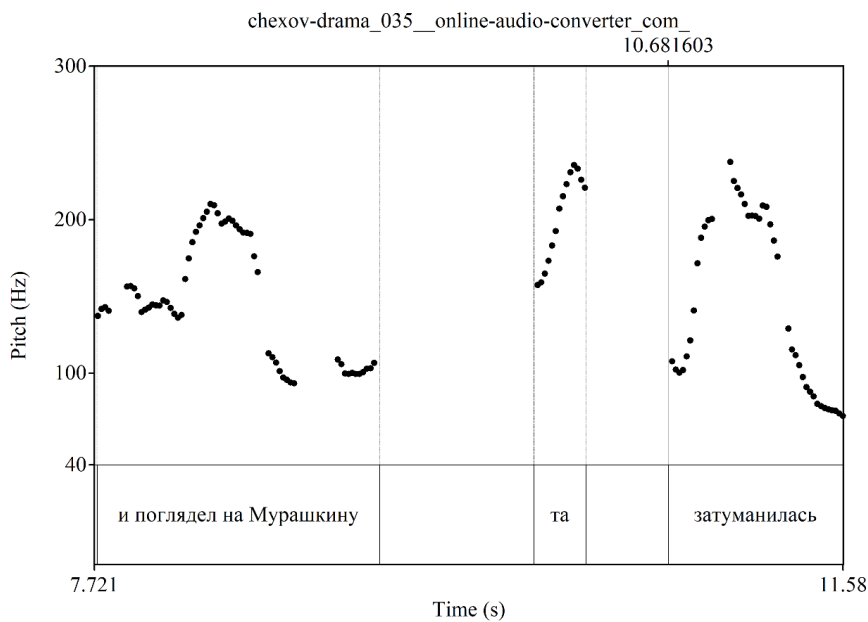


Рисунок 1. Интонограмма к примеру (25)

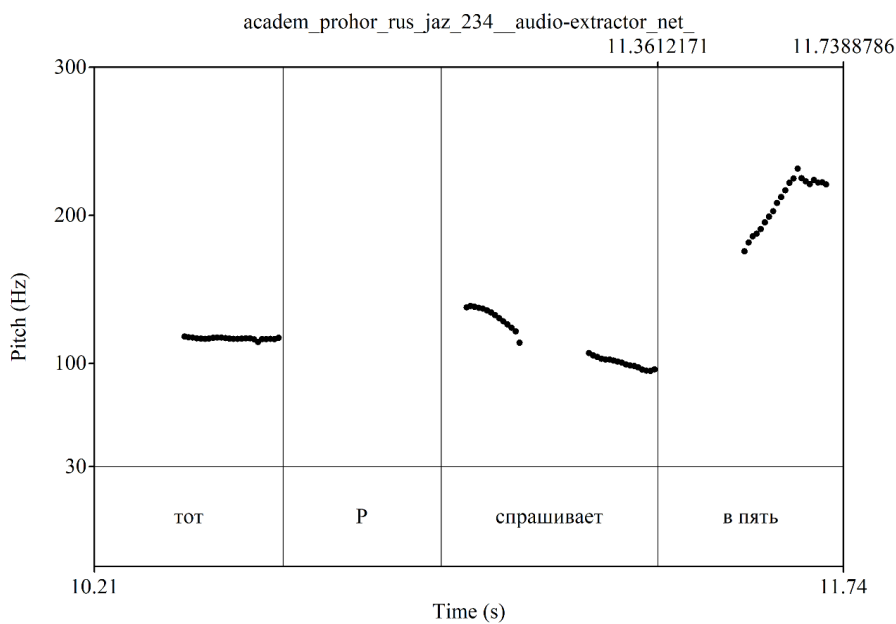


Рисунок 2. Интонограмма к примеру (26)

Таким образом, во всех примерах обследованной выборки ТОТ имеет статус, отличный от рематического.

4. Обсуждение и выводы

Мы обсудили структурные ограничения на употребление конструкций с анафором ТОТ и ограничения, связанные с актуальным членением. Однако, как справедливо сказано в работе [Крейдлин, Чехов 1988], для исчерпывающего описания этих конструкций необходимо описать также прагматические ограничения или, я бы сказала шире — дискурсивные ограничения.

Местоимение ТОТ, будучи в своей исходной дейктической функции дистальным, при анафорическом употреблении отвечает за противопоставление более периферийного референта более центральному, коммуникативно более близкому\значимому для говорящего на данном отрезке дискурса. По терминологии Е. В. Падучевой [1985: 125–127, 2016], антецедент анафорического ТОТ не должен быть «смысловой темой текста», или с уточнением Г. Е. Крейдлина и А. С. Чехова [1988: 50–51] — «смысловой темой релевантного фрагмента текста». Даже если центральный референт отсутствует в текущем дискурсивном «окне» (как в примерах (15), (16) выше), он незримо участвует в противопоставлении как точка отсчета. Само по себе повторное упоминание периферийного референта с помощью ТОТ уже продвигает его дискурсивный статус, но в прототипических случаях это продвижение сопровождается еще и повышением грамматического и коммуникативного статуса: переход от не-актора и не-подлежащего к актору и подлежащему-теме. Фактически, перечисленные нами выше грамматические феномены — это совокупность средств, грамматикализирующих дискурсивную функцию «присвоение статуса смысловой темы». О том, что функция анафорического ТОТ состоит именно в однократной смене, а не в поддержании дискурсивного статуса, говорит тот факт, что для анафора ТОТ — в отличие от ОН — затруднено повторное употребление внутри референциальной цепочки; ср. (27), где в третьей клаузе повторное упоминание Маши возможно с помощью ОНА и невозможно с помощью ТА:

(27) (ВИП) *Петя_i встретил Машу_j; Та_j была не в духе. Она_j / *Та даже не поздоровалась.*

«Родимые пятна» дейктического употребления проявляются в том, что анафорическому ТОТ требуется доступный конкретный антецедент. Поэтому, в частности, в конструкциях с ТОТ затруднена катафора, ср. (28a) и (28b):

(28) (a — ВИП) *Иван уже сообщил Ольге, что она/та отстранена от конкурса*
(b — ВИП) *Что она/*?та отстранена от конкурса, Иван уже сообщил Ольге.*

По этой же причине ТОТ не употребляется с предикатными антецедентами и в качестве «ленивого» местоимения (местоимения повтора):

(29) *Ивану всё же пришлось пойти к зубному врачу, хотя он всегда их/*тех боялся* (Цитируется по [Крейдли́н, Чехов 1988: 41])

Исходная дистальность дейктического ТОТ проявляется не только в том, что его антецедент всегда рассматривается на фоне другого, альтернативного референта, но и в том, что его антецедент, может получать негативную оценку, с метафорическим сдвигом от более далекого к менее значимому и, далее, к менее положительному. По меткому наблюдению А. Е. Крейдлина и А. С. Чехова [Крейдли́н, Чехов 1988:48], ТОТ может оказаться неуместным анафором, если у говорящего имеется подчеркнуто вежливое отношение к антецеденту ТОТ:

(30) *Автор благодарит X за то, что он/*тот прочитал статью в рукописи и сделал ряд ценных замечаний.* (Цитируется по [Крейдли́н, Чехов 1988: 48])

Интересно, что все перечисленные симптомы контрастивности — соотнесенность антецедента с другим, альтернативным референтом, предпочтительность ТОТ при референциальном конфликте, неспособность выступать в катафорической функции и в функции местоимения повтора, проявление негативной коннотации — анафорическое ТОТ разделяет с указательными местоимениями в ряде других языков, где демонстративы, как и в русском, могут выступать в анафорической функции, в частности, в немецком, португальском, французском, иврите, см. [Patel-Grosz, Grosz 2017], [Sichel, Wiltschko 2018]. Особенно вдохновляющими выглядят работы Петера Боша и его коллег [Bosch 2013], [Hinterwimmer; Bosch 2018], [Bosch; Hinterwimmer 2016], *inter alia*, в которых немецкое указательное *der* в качестве анафора исследуется корпусными, экспериментальными психолингвистическими методами и методами, использующими регистрацию движений глаз. В этих работах убедительно показано, что для *der* в функции анафора предпочтительны антецеденты, которые не являются темами (топиками), но исключения из этого правила возможны, если вмешиваются факторы контекстной (прагматической) уместности или если морфо-синтаксис конструкции позволяет избежать референциального конфликта без опоры на местоимение.

Анафорическое ТОТ можно поставить в один ряд с показателями совпадения/смены подлежащего в смежных предикациях в тех языках, где в составе глагола или в составе коннектора выражается грамматическая категория «переключения референции» (*switch-reference*). За такого рода средствами устойчиво закреплен кластер значений, связанных с дискурсивной «преемственностью» (когерентностью) текста при одновременном указании на частичное изменение того или иного дискурсивного параметра. Добавлю еще, что анафорическое ТОТ повышает дискурсивный ранг своего антецедента однократно и временно. Этот антецедент не превращается в протагониста текста, его высокий ранг сохраняется только в пределах текущего дискурсивного окна. Поэтому анафор ТОТ часто используется в контексте «многоходовой комбинации» (см. подробнее [Подлесская 2001]): когда требуется подключить участника к списку действующих лиц, которые поочередно временно выдвигаются на первый план, причем обычно последовательность выдвижения диктуется структурой общей ситуации с множеством участников — беседа, игра и т. д.

Мой заключительный пример — из стихотворения Н. Коржавина «Памяти Герцена (баллада об историческом недосыпе)» — демонстрирует употребление анафора ТОТ именно в составе «многоходовой комбинации»³:

(31) *...И, ошалев от их поступка дерзкого,
Он поднял страшный на весь мир трезвон.
Чем разбудил случайно Чернышевского,
Не зная сам, что этим сделал он.*

*А тот со сна, имея нервы слабые,
Стал к топору Россию призывать, —
Чем потревожил крепкий сон Желябова,
А тот Перовской не дал всласть поспать.*

*...Был царь убит, но мир не зажил заново.
Желябов пал, уснул несладким сном.
Но перед этим побудил Плеханова,
Чтоб тот пошёл совсем другим путём...*

Полученные результаты могут иметь двоякое значение. Во-первых — теоретическое и типологическое: они вовлекают новые данные в оборот исследований по кросс-языковой вариативности анафорических средств. Во-вторых — прикладное: они могут послужить усовершенствованию систем автоматического анализа референциальных цепочек и, в частности, использоваться для подготовки релевантных датасетов для этой задачи.

Литература

1. Брызгунова Е. А. (1982) Интонация // Русская грамматика, том 1, М.: Наука, 103–118.
2. Живов В. М. (1999) Автономность письменного узуса и проблема преемственности в восточнославянской средневековой письменности // Славянское языкознание. XII Международный съезд славистов: Доклады российской делегации. М., 212–247.

³ По-видимому, построения типу «многоходовой комбинации» с местоимением ТОТ — явление достаточно старое. Так, А. П. Майоров указывает, со ссылкой на [Живов 1999], что такие построения были типичны для деловой письменности XVIII в.: «наиболее выразительной особенностью приказного языка является его синтаксис, а именно так называемое „нанизывание“ предикативных конструкций, при котором рема... первой предикативной конструкции повторяется как тема во второй, рема... второй предикативной конструкции повторяется как тема в третьей и т. д., создавая особый тип подчинительной связи» [Живов 1999: 234]. При таком типе словесной организации текста указательное местоимение *тотъ* в составе темы каждой последующей предикативной конструкции призвано актуализировать указание на тот объект, который выступает как рема в предыдущем предложении» [Майоров 2004: 231].

3. Кибрик А. А. (1987) Механизмы устранения референциального конфликта // Кибрик А. Е., Нариньяни А. С. (ред.), Моделирование языковой деятельности в интеллектуальных системах. М.: «Наука», 128–145.
4. Крейдлин Г. Е., Чехов А. С. (1988) Соотношение семантики, актуального членения и прагматики в лексикографическом описании анафорических местоимений (на материале местоимений группы ТОТ) // Институт русского языка АН СССР. Проблемная группа по экспериментальной и прикладной лингвистике. Предварительные публикации. Вып. 178.
5. Майоров А. П. (2004) Указательные местоимения сей, тот, оной, этот в деловом языке XVII—XVIII вв. // Русский язык в научном освещении, №2 (8), 224–239.
6. Падучева Е. В. (2016) Указательные местоимения. Материалы для проекта корпусного описания русской грамматики (<http://rusgram.ru>). На правах рукописи. М.
7. Падучева Е. В. (1985) Высказывание и его соотнесенность с действительностью. М.: Наука. (изд. 6-е, испр. — М.: Изд-во ЛКИ. 2010).
8. Подлесская В. И. (2001) Переключение референции: дискурсивные функции грамматической категории // Плунгян В. А. (ред.) Исследования по теории грамматики. 1. Глагольные категории. Москва: Русские словари, 198–208.
9. Селезнев М. Г. (1987) Референция и номинация // Кибрик А. Е., Нариньяни А. С. (ред.), Моделирование языковой деятельности в интеллектуальных системах. М.: «Наука», 64–78.

References

1. Bryzgunova E. A. (1982) Intonation [Intonacija] // Russkaja grammatica. V. 1. M: Nauka, 103–118.
2. Zhivov V. M. (1999) The autonomy of the written modus of discourse and the problem of succession in Eastern Slavic medieval writing [Avtonomnost' pis'mennogo uzusa i problema preemstvennosti v vostochnoslavjanskoj srednevekovoj pis'mennosti] // Slavjanskoe jazykoznanie. XII Mezhdunarodnyj sjezd slavistov: Doklady rossijskoj delegacii. M., 212–247.
3. Kibrik A. A. (1987) Mechanisms of solving referential conflicts [Mexanizmy ustraneniya referencial'nogo konflikta] // Kibrik A. E., Narin'jani (eds.) Modelirovanie jazykovoj dejatel'nosti v intellektual'nyx sistemax. M: Nauka, 128–145.
4. Krejdlin G. E., Chexov A. S. (1988) Semantics, information structure and pragmatics in the lexicographic description of anaphoric pronouns (TOT and similar anaphoric pronouns) [Sootnoshenie semantiki, aktual'nogo chleneniya i pragmatiki v leksikograficheskom opisanii anaforicheskix mestoimenij (na materiale mestoimenij gruppy TOT)] // Institut russkogo jazyka AN SSSR. Problemnaja gruppy po eksperimental'noj i prikladnoj lingvistike. Predvaritel'nye publikacii. Vyp. 178.
5. Majorov A. P. (2004) Demonstrative pronouns *sej, tot, onoj, etot* in the business language of the XVII–XVIII centuries [Ukazatel'nye mestoimenija *sej, tot, onoj, etot* v delovom jazyke XVII–XVIII vv.] // Russkij jazyk v nauchnom osveschenii, №2 (8), 224–239.

6. *Paducheva E. V.* (2016) Demonstrative pronouns. Materials for the Russian corpus-based grammar project [Ukazatel'nye mestoimenija. Materialy dlja proekta korpusnogo opisanija russskoj grammatiki] (<http://rusgram.ru>). A manuscript. M.
7. *Paducheva E. V.* (2016) Utterance and its relation to actuality [Vyskazyvanie i ego sootnesennost's dejstvitel'nostju]. M: Nauka (6th edition, revised — M.: LKI. 2010)
8. *Podlesskaya V. I.* (2001) Switch-reference: discourse functions of a grammatical category [Pereključenje referencii: diskursivnye funkcii grammatičeskoj kategorii // Plungjan V. A. (ed.) Issledovanija po teorii grammatiki. I. Glagol'nye kategorii. M.: Russkie slovari, 198–208.
9. *Seleznev M. G.* (1987) Reference and nomination [Referencija i nominacija] // Kibrik A. E., Narin'jani (eds.) Modelirovanie jazykovoje dejatel'nosti v intellektual'nyx sistemax. M: Nauka, 64–78.
10. *Berger, T.; Weiss, D.* (1987) Die Gebrauchsbedingungen des Anaphorikums "tot" in substantivischer Verwendung // Freidhof, G.; Kosta, P. (eds.), Slavistische Linguistik, München, 9–93.
11. *Bosch, Peter.* (2013) Anaphoric reference by demonstrative pronouns in German // Presented at the Workshop on the Impact of Pronominal Form on Interpretation, Universität Tübingen, Nov. 2013.
12. *Bosch, Peter; Stefan Hinterwimmer.* (2016) Anaphoric reference by demonstrative pronouns in German // Anke Holler, & Katja Suckow (eds.), Empirical perspectives on anaphora resolution. Berlin/New York: De Gruyter, 193–212.
13. *Boersma, P., & Weenink, D.* (2012) Praat: Doing phonetics by computer (Version 5.3.04) [Computer software]. Retrieved from <http://www.praat.org/>.
14. *Hinterwimmer, Stefan; Peter Bosch.* (2018) Demonstrative pronouns and propositional attitudes // Pritty Patel-Grosz, Patrick Georg Grosz & Sarah Zobel (eds.), Pronouns in embedded contexts at the syntax-semantics interface (Studies in Linguistics and Philosophy). Springer, 105–144.
15. *Kibrik, Andrej A.* (2011) Reference in discourse. Oxford: Oxford University Press.
16. *Patel-Grosz, Pritty; Grosz, Patrick G.* (2017) Revisiting Pronominal Typology // Linguistic Inquiry, Volume 48, Number 2, 259–297.
17. *Sichel, Ivy; Wiltschko, Martina* (2018) Demonstrative Pronouns and the Linguistic Encoding of Appraisal // Wm. G. Bennett et al. (eds.), Proceedings of the 35th West Coast Conference on Formal Linguistics, 365–373. Somerville, MA: Cascadia Proceedings Project. www.lingref.com, document #3409.

RUSSIAN NATURAL LANGUAGE GENERATION: CREATION OF A LANGUAGE MODELING DATASET AND EVALUATION WITH MODERN NEURAL ARCHITECTURES

Shaheen Z. (shaheen@itmo.ru)^{1,2},
Wohlgenannt G. (gwohlg@itmo.ru)¹,
Zaity B. (bassel.zaity@gmail.com)²,
Mouromtsev D. I. (mouromtsev@itmo.ru)¹,
Pak V. G. (vadim.pak@cit.icc.spbstu.ru)²

¹Faculty of Software Engineering and Computer Systems, ITMO University;

²Institute of Computer Science and Technology Peter the Great
St. Petersburg Polytechnic University (SPbPU), St. Petersburg, Russia

Generating coherent, grammatically correct, and meaningful text is very challenging, however, it is crucial to many modern NLP systems. So far, research has mostly focused on English language, for other languages both standardized datasets, as well as experiments with state-of-the-art models, are rare. In this work, we i) provide a novel reference dataset for Russian language modeling, ii) experiment with popular modern methods for text generation, namely variational autoencoders, and generative adversarial networks, which we trained on the new dataset. We evaluate the generated text regarding metrics such as perplexity, grammatical correctness and lexical diversity.

Key words: natural language generation, variational autoencoder, dataset construction, seqGAN

DOI: 10.28995/2075-7182-2020-19-644-657

1. Introduction

Text generation is a key component in many NLP systems that produce text such as translation systems, dialogue systems, or text summarization. The quality of the generated text is critical in these systems, it should be coherent and well-formed, without grammatical mistakes, and semantically meaningful [5]. Generating human-like text is challenging, it includes modeling high-level syntactic properties and features like sentiment and topic [1].

Natural Language Generation (NLG) produces human-understandable NL text in a systematic way—based on non-textual data (eg. a knowledge base) or from meaning representations (eg. a given state of a dialogue system) [16]. Modern NLG systems often make use of (neural) language models [18]. A language model (LM) is a probability distribution over a sequence of words, and can be used to predict the next word given an input sequence.

In recent years, various types of neural network architectures have been successfully applied in NLG, such as variational autoencoders (VAE) [1], [21], generative

adversarial networks (GAN) [3], [5], [22], and recurrent neural networks (RNN) [11]. Here, we experiment with those architectures on Russian language.

The *goals* of this paper are (i) to create a reference dataset for language modeling for the Russian language, comparable to the popular Penn Tree Bank (PTB) dataset for English language, and (ii) to adapt and to train several state-of-the-art language models and to evaluate them on the task of Russian language text generation. We create a dataset of 236K sentences by sampling from the Lenta News dataset, preprocess the text, and filter sentences that do not match certain quality criteria. Then we train six models (four VAE models with different scheduling methods, seqGAN, and LSTM RNNLM) on the new corpus, and evaluate them regarding the perplexity metric, and manually validate 100 sentences for each model regarding grammatical correctness. We achieve best results with the VAE models, the *zero* variant performs well regarding perplexity, but overall the *cyclical* VAE model shows the highest performance, as it generates the largest fraction of grammatically correct sentences, which have similar characteristics (sentence length, etc.) as the training data.

2. Related Work

Our Russian language dataset is inspired by the plain-text/language-modeling part of the PTB dataset¹. PTB contains about 1M words from 1989 Wall Street Journal material, with various annotations such as POS-tags. This dataset is very popular among NLP researchers for language modeling and other NLP tasks. Many recent language models are trained and evaluated also on larger corpora, such as WikiText-103 [12], or WebText [17] (created for the GPT transformer models). For languages other than English high-quality reference datasets are rare.

In language modeling, Recurrent Neural Network Language Models (RNNLM) [14], and extensions such as long short-term memory (LSTM) [19], are frequently used architectures. RNNLMs generate text word-by-word depending on a hidden state that summarizes the previous history. These models are able to capture long-range dependencies, however, they do not expose interpretable states that represent global features like topic or sentiment. Regarding recent RNNLMs, for example Merity et al. [11] investigate different strategies for regularizing and optimizing LSTM-based models.

Variational Autoencoders (VAEs) [6] have been applied to many domains, including language modeling [1], [21]. They showed impressive results in producing interpretable representations of global features like the topic or of high-level syntactic properties. For example, Yang et al. [21] use the method for unsupervised clustering of the text. VAEs are trained using regularization to avoid overfitting and produce a regular latent space that has properties enabling the generative process. Recent research on VAEs focuses on improving the quality of the hidden representation, on exploring the properties of the latent space, and experiments with different architectures to improve VAEs.

Generative adversarial networks GANs [4] train a *generator* that tries to produce realistic samples from a data distribution. The generator is guided by a *discriminator* on how to modify its parameters. GANs have been applied successfully to computer

¹ <https://catalog.ldc.upenn.edu/LDC99T42>

vision tasks, however, adapting GANs to generate texts is challenging due to the discrete nature of natural language. Many attempts to adopt GANs to text rely on using reinforcement learning [3], [22], or on Gumbel-Softmax approximation [9] (a continuous approximation of the softmax function). Zhang et al. [23] use a feature matching scheme for training GANs to generate realistic-looking text.

Little work exists on NLG for the Russian language. Nesterenko [15] uses a simple template-based system to generate stock market news in Russian. Kipyatkova and Karpov [7] study the use of RNNLM models in Russian speech recognition. Kuratov and Arkhipov [8] train a BERT (transformer) language model on Russian text (RuBERT) and evaluate it on tasks such as paraphrase and sentiment detection. Finally, Shimorina et al. [20] present an English-Russian parallel corpus for generating natural language text from the triples of a knowledge base (data-to-text NLG). Their corpus was created with neural machine translation. However, to the best of our knowledge, for general Russian NLG no research work has been published about general-domain NLG datasets and about the evaluation of NLG models based on modern neural architectures.

3. Variational Autoencoder

In this section, we introduce the VAE variants (zero, constant, linear, cyclical) which are applied in the experiments. An autoencoder (AE) consists of an encoder that encodes an input sequence into a hidden state and a decoder that uses this hidden state to reconstruct the original sequence. In a standard AE for language modeling, an RNN is used for both the encoder and the decoder. The decoder is then used for text generation, where each output token is conditioned on the previous output tokens. A Variational Autoencoder (VAE) encodes the input sequence x into a region in the latent space rather than a single point, this region is defined using a multi-variate Gaussian prior $p(z)$, where the last hidden state of the encoder (z) is projected on two separate vectors. These vectors represent the mean and the diagonal co-variance matrix of the prior. To restore the original sequence, the initial state of the decoder is sampled from the prior, and then used to decode the output sequence. This way, the model is forced to be able to decode plausible sentences from every point in the latent space, that has a reasonable probability under the prior [1]. A standard recurrent neural network language model is based on a series of next-step predictions, thus a standard AE does not provide an interpretable representation of global features such as the topic or of high-level syntactic properties.

The VAE modifies the AE architecture by replacing the deterministic encoder with a learned posterior recognition model $q(z|x)$. If the VAE were trained with standard AE reconstruction objective, it would learn to encode x deterministically by making $q(z|x)$ vanishingly small. However, we want the posterior to be close to the prior (most often standard Gaussian), therefore we have two objectives and the goal is to optimize the following lower-bound:

$$L(\theta; x) = -KL(q\theta(z|x)||p(z)) + E_{z \sim q\theta(z|x)}[\log p\theta(x|z)] \leq \log p(x) \quad (1)$$

The first term is the KL-divergence of the posterior from the prior, and the second is the reconstruction loss, where θ stands for the parameters of the neural network. Straightforward training of the network using this objective will bring $q(z|x)$

to be exactly the same as the prior $p(z)$, and KL-divergence term in the cost function to zero. As a result, the model will behave like a standard RNNLM. Bowman et al. [1] use KL-cost annealing to solve this problem, by multiplying a variable weight β with the KL term at training time. In the beginning, β will be set to zero, and then it gets gradually increased, forcing the model to smooth out its encodings and pack them into the prior. Later research by Fu et al. [10] investigates KL annealing further, they experiment with three scheduling approaches:

- Constant Schedule: the standard approach is to keep $\beta = 1$ fixed during training, which causes the vanishing of the KL-term, the model will behave as a standard RNNLM.
- Monotonic (linear) Annealing Schedule: this is the previously described approach for VAEs by Bowman et al. [1]. It starts with $\beta = 0$ and gradually increases during training to $\beta = 1$.
- Cyclical Annealing Schedule: split the training process into M cycles, each cycle consists of two stages:
 1. Annealing, where β is annealed from 0 to 1 in the first $R[T/M]$ training steps over the cycle; R : proportion used to increase β , T : the total number of global steps, M : the number of cycles.
 2. Fixing: fix $\beta = 1$ for the rest of the cycle.

According to Fu et al. [10] cyclical KL-annealing results in better latent codes by leveraging the informative representations of previous cycles as warm re-starts.

4. SeqGAN

A Generative Adversarial Network (GAN) contains two main models: the discriminator D is trained to distinguish between real and fake data and the generator G , which is trained to fool D and tries to generate samples from the real data distribution. D and G are trained simultaneously, the improvement in one model will cause further improvements in the other model. When G is able to produce samples from the same data distribution as the real data, D will be no longer able to distinguish between real and sampled data.

GANs were applied successfully to tasks in computer vision. However, adapting GANs to text generation is not straightforward, because the original GAN works with continuous data, meanwhile, text is discrete and the generator uses a non-differential function to produce each token, therefore it is difficult to pass gradient updates from D to G . Another difficulty is that D can evaluate only the completed text; for a partially generated sequence the evaluation score depends on the score of the full sentence.

To address these problems, Yu et al. [22] use Reinforcement learning to train GANs for text generation. The generator is treated as an agent of reinforcement learning, and the state corresponds to the generated tokens up to this moment and the action is the next token to be generated. A discriminator will give the feedback reward that guides the learning process.

We can consider G as a stochastic parametric policy where Monte Carlo (MC) search is used to approximate the state-action value using the policy gradient. D is trained

by providing positive examples from the real data and negative examples from the generated samples by G . G is updated using the policy gradient and MC search by the reward signal received from D . The reward represents how likely the discriminator is fooled by the generator. Yu et al. [22] use a rollout policy with MC search, they set the rollout policy the same as G during the experiments. For G they choose a recurrent neural network with LSTM cells, and D uses a convolutional neural network with highway architecture.

The training strategy is as follows: first G is pre-trained using maximum likelihood estimation, then G and D are trained alternatively. To train D , at each training step we sample negative examples using G and use examples from the true data as the positive examples.

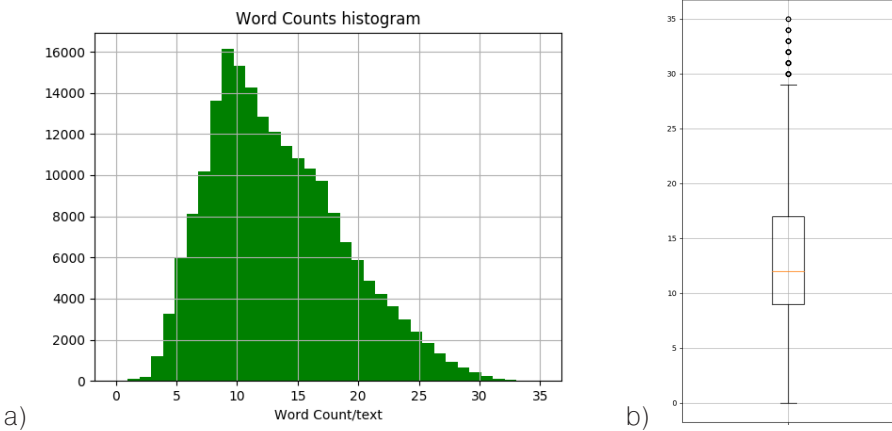


Figure 1: a) A histogram visualizing the number of words per sentence which shows the probability distribution of the data, b) A box plot for the number of words per sentence

5. Dataset

In this section we discuss the Russian language dataset for language modeling, its creation, and characteristics.

Penn Tree Bank (PTB) [13] is a very popular dataset for experimenting with language models, and many researchers use it in experiments with the same settings (dataset splits, etc.)—which allows to compare different language modeling approaches. PTB is heavily preprocessed, and the dataset vocabulary is limited to 10,000 tokens with no numbers, punctuation or capital letters.

Our goal was to create a similar reference dataset for the Russian language. Our dataset is based on the Lenta news dataset², a corpus of over 800K Russian news articles collected from Lenta.Ru between 1999–2019. To create our dataset we randomly sample sentences from the Lenta dataset after we apply preprocessing in a similar way as in PTB.

² <https://github.com/yutkin/Lenta.Ru-News-Dataset>

The *preprocessing pipeline* includes: a) lower-casing the text, b) replacing all URLs with a special token <url>, c) separating punctuation symbols from neighboring tokens with spaces, and finally d) replacing digits with a special character *D*, and for any 5 or more consecutive digits we use the special character *N*. Therefore, numbers in the dataset will take the following forms: *D*, *DD*, *DDD*, *DDDD*, or *N*. Step d) helps to keep meaningful numeric tokens, such as in these examples:

- часть *D* статьи *DDD* ук
- встреча так и завершилась со счетом *D* : *D*
- в *DDDD* году, госкорпорация обеспечивает *DD* процента электроэнергии .

Subsequently, we apply the NLTK³ PunktSentenceTokenizer⁴ to tokenize the text into sentences. We create a dictionary using the 15,000 most frequent tokens, and replace all other tokens in the text with <UNK>. In contrast to PTB, which uses a dictionary size of 10,000, we decided to keep 15,000 tokens because of the rich morphology of the Russian language where nouns, adjectives, and verbs change forms according to their role in the sentence.

Finally, we create the Russian language corpus by sampling 200,000 examples for the training set, 16,000 for the development set and 16,000 for the test set. In the sampling process, we only accepted sentences which fulfill the following conditions:

- The sentence contains less than 40 tokens.
- The sentence does not include any English words.
- The number of single (') and double (") quotation marks is even (balanced).
- Every opening bracket is followed by a closing bracket (balancing condition).
- Less than 10% of the tokens in the sampled sentence are the <UNK> token.

The dataset and code are available on GitHub⁵. Finally, we provide an overview of dataset statistics, including the mean number and standard deviation of the number of tokens per sentence in **Table 1**. The histogram and the box plot in **Fig. 1** show the distribution of sentence length (number of tokens) in the dataset. Those statistics will be helpful to compare the sentences generated with various methods (see next section) with the original dataset.

Table 1: An overview of the Russian language modeling corpus (statistics of training, development and test set)

	Training set	Development set	Test set
#examples	200,000	16,000	16,000
Mean #tokens per example	13.26	13.17	13.14
Stddev for #tokens per example	5.63	5.66	5.62
#unique tokens	14,511	13,398	13,401

³ <https://nltk.org>

⁴ https://github.com/Mottl/ru_punkt

⁵ https://github.com/zeinsh/lenta_short_sentences

6. Experiments

6.1. Evaluation Setup

We experiment with the following popular text generation methods: VAE, GAN, and RNN.

For the VAE method, we build on the implementation by Baumgärtner⁶ and add the implementation of additional scheduling methods, namely cyclical, constant and zero scheduling. Constant and cyclical scheduling are explained in [Section 3](#), and in the zero schedule we set $\beta = 0$, which excludes the KL-divergence term from the lower-bound computation in [equation \(1\)](#). For the VAE encoder, we use 300-dim input embeddings, a single layer of 256 LSTM cells, and a latent vector (bottleneck) size of 16, and in the decoder again 256 LSTM cells and a 300-dim output. We train the VAEs for 10 iterations with an embedding dropout probability of 0.5.

The experiments with **seqGAN** are based on the PyTorch implementation on GitHub⁷, We modified the original implementation to adapt it to our text dataset instead of discrete numbers created by the oracle generator in the original implementation. For seqGAN, we also use 300-dim input embeddings, and a single layer of 256 LSTM cells. We pretrain G for 10 iterations with an embedding dropout probability of 0.5, and then pretrain D (10 iterations). Then follow 10 epochs of adversarial training, each of which trains G for one iteration and D for five iterations. Finally, to measure the effect of the adversarial training on the RNNLM, we performed an evaluation on the pre-trained LSTM generator separately as **RNNLM** (see [Section 6.2](#)).

For evaluation, we use a dual strategy. First, in line with most research on language modeling, we calculate the *perplexity* on the test set for each model. Perplexity is a measure of how well a probability distribution predicts a sample. Perplexity does not always correlate with human correlation, in fact there is sometimes a negative correlation [2], for this reason we also include an human expert evaluation (see below).

Furthermore, we generate 10,000 sample texts with each method by greedily sampling word by word. Examples of the results are given in GitHub repository⁸ and in [Appendix 1](#). For the generated samples we use expert evaluation, where a Russian native speaker checks the generated text for grammatical correctness, and assigns a score of either 1 or 0. The value 1 signifies that no grammatical mistakes were found in the text. As we manually evaluate 100 unique sentences for each model, the maximum score per model is 100.

6.2. Evaluation Result

Here we analyze the models with regards to the following aspects: perplexity, token statistics of generated text, and manual evaluation of grammatical correctness. Furthermore, for VAE models we discuss the spatial distribution of latent representations.

⁶ <https://github.com/timbmg/Sentence-VAE>

⁷ <https://github.com/suragnair/seqGAN>

⁸ https://github.com/zeinsh/lenta_short_sentences/blob/master/samples.md

In **Table 2** we report on the perplexity metric for the test set. The *zero* VAE model clearly shows the best results, followed by the cyclical model. As we will see in **Tables 3** and **4**, despite good results on perplexity, the *zero* model does not excel in grammatical correctness of the generated text.

Table 2: Perplexity calculated on test set for each model

x	zero	constant	linear	cyclical	RNNLM	seqGAN
perplexity	7.19	16.27	14.36	14.11	27.88	27.93

Table 3 gives an overview of some statistics of the generated text samples. We can see that LSTM RNNLM, seqGAN, and especially the *zero* variant of VAE produce a large number of unique sentences—with very little overlap with the training sentences. On the other hand, the *constant* VAE model fails to generate a large variety of sentences. Liu et al. [10] argue that the constant schedule ignores z and treats it as noise. Regarding sentence length, most models are similar, except *constant*, which creates shorter sentences. More interestingly, the *zero* and *constant* model show little variance in sentence length, while the other models much better capture the variance in sentence length of the training dataset. Finally, only RNNLM and seqGAN are able to have a diversity in vocabulary similar to the training data.

As mentioned, a Russian native speaker manually verified 100 generated sentences for each model regarding grammar. The results (number of grammatically correct sentences) are presented in **Table 4**. The data clearly shows that the *cyclical* VAE model performs very well (91% correct sentences), while the *zero* VAE model, although providing low perplexity, produces many grammatically wrong sentences. This corresponds with Chang et al. [2], ie. that perplexity does not always correlate with human evaluation. In our experiments, RNNLM and seqGAN fail to generate a high ratio of grammatically correct sentences.

Table 3: Comparison between models regarding uniqueness and sentence length of the generated data

	zero	constant	linear	cyclical	RNNLM	seqGAN
# unique sent. (out of 10000)	10,000	231	8,810	8,868	9,972	9,979
# unique sent. not in train-set	10,000	225	8,694	8,752	9,972	9,979
Mean #tokens per sample	10.52	8.60	11.89	11.25	13.53	13.89
Std-dev of # tokens per sample	2.25	2.27	5.08	4.78	5.99	6.08
# unique words	5,324	353	5,028	4,550	11,430	11,505

Table 4: Number of grammatically correct sentences (out of 100) checked by a native speaker

	zero	constant	linear	cyclical	RNNLM	seqGAN
Score	77	79	86	91	43	51

Fig. 2 shows a projection of the latent representations in the development set into 2D space (using tSNE⁹). The figure compares the resulting distributions of the VAE methods with a zero, constant, linear and cyclical schedule. The figure shows that the zero schedule, which corresponds to a standard autoencoder, produces a rather irregular distribution of latent codes. The KL-divergence term in VAEs causes the algorithm to fill the latent space.

As discussed before, although zero gives the best perplexity on the test set, samples from the zero model contain many grammatically incorrect sentences, as it sometimes samples z from regions in the latent space with low density. That explains why samples from linear and cyclical VAEs are better in terms of grammar, where latent codes produced by these models fill the latent space.

Finally, in the appendix and in the GitHub repository¹⁰, we give examples on how VAE models can interpolate between two sentences. Following Bowman et al. [1], we sample two random points in the latent space and then decode those into two sentences. Then, starting from the first sentence, we gradually move through the latent space on a line to the second sentence, and pick points on the way, which are decoded into sentences. This process makes the ability of interpolation in the latent space explicit.

In summary, we found that the cyclical VAE produces best result with regards to grammatical correctness, followed by the VAE with linear schedule. Both also generate sentences with similar characteristics as in the training set (regarding sentence length), however concerning sentence length and diversity of the used vocabulary the plain RNNLM and the seqGAN produce better results. But RNNLM and seqGAN generate a high number of grammatically wrong sentences in the experiments.

⁹ <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>

¹⁰ https://github.com/zeinsh/lenta_short_sentences/blob/master/interpolation.md

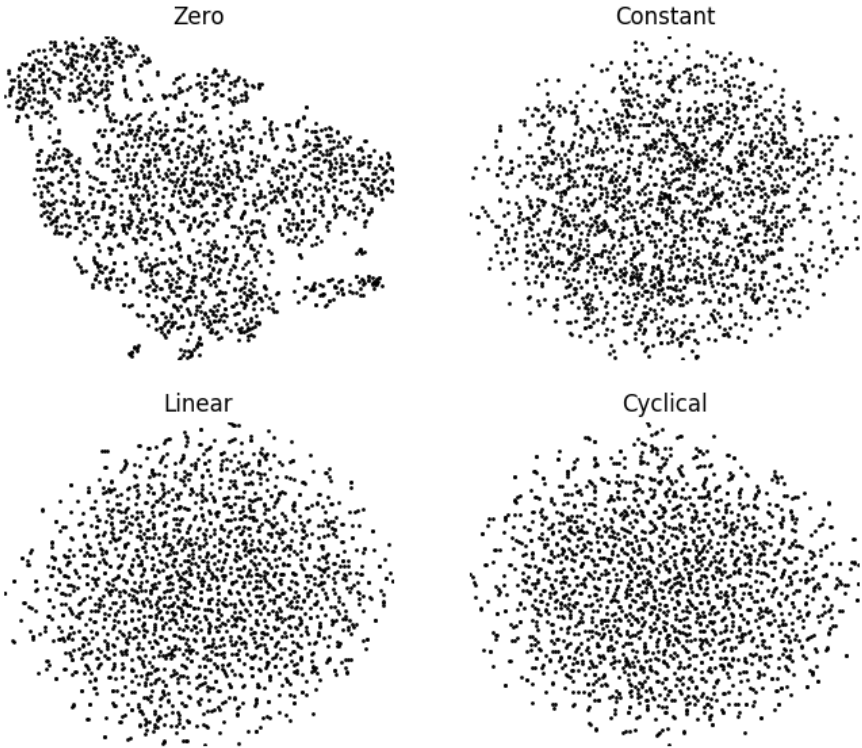


Figure 2: Latent representation of texts in the development set. The KL-divergence term in the constant, linear, and cyclical schedule forces the encoder to fill the latent space.

7. Conclusion

In this work, we present a new dataset for Russian language modeling (based on the Lenta News dataset) and perform a comparative study between two modern methods in text generation, namely VAE and seqGAN. Our results show the effect of the scheduling method on the quality of the generated text in VAEs, where linear and cyclical schedules produced the best models grammatically, however, the zero method showed the best perplexity, but an irregular distribution of the latent codes. LSTM and SeqGAN were able to replicate the mean and variance of the length of sentences in the original dataset as well as the number of unique words. The contributions of this work are: i) the provision (on GitHub) of a reference dataset for Russian language modeling with 236K sentences in total, ii) the adaption of various VAE variants and seqGAN to Russian text, iii) and extensive experiments and evaluations with the chosen deep learning methods which indicate that the cyclical VAE approach performs best overall. Future work will include a deeper investigation of the latent

representations produced by VAEs (and why VAEs produce less diverse sentences), apply state-of-the-art models like LeakGAN and studying the generation of Russian language text conditioned on topic, sentiment, etc.

8. Acknowledgments

This work was supported by the Government of the Russian Federation (Grant 074-U01) through the ITMO Fellowship and Professorship Program.

Appendix 1

In the appendix, we show the interpolation using the four VAE models trained on the reference dataset. As described in [Section 6.2](#) we sample two random points in the latent space and then decode those into two sentences. Starting from the first sentence, we gradually move through the latent space on a line to the second sentence, and pick points on the way, which are decoded into sentences. This process makes the ability of interpolation in the latent space explicit.

VAE/zero schedule

- пожары с ними возник конфликт между двумя группами и на юго—востоке <unk> . <eos>
- пожары с ними возник конфликт между двумя группами и на юго—востоке <unk> . <eos>
- пожары с рельсов сошел с одной из самых опасных технологий , вызванных <unk> на работу . <eos>
- пожары с созданием по атомной энергии (магатэ) , на <unk> островах . <eos>
- пожары с созданием по спасению ракет у ворот у берегов острова , <unk> за рубеж . <eos>
- соперник по многим показателям проходит у дома от продажи билетов , <unk> за рубеж . <eos>
- соперник по многим показателям (нсн) создает у нее , <unk> за рубеж . <eos>
- соперник (по словам представителя рфс) при этом <unk> , сообщает тасс . <eos>
- соперник (по словам) у него возникли трудности , <unk> за собой . <eos>
- соперник—гран—при россии по легкой атлетике , <unk> за рубеж . <eos>

VAE/constant schedule

Explanation: This model decoded both points from the latent space into the *same* sentence.

- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>
- в dddd году он был объявлен в международный розыск . <eos>

VAE/linear schedule

- точная дата выхода фильма пока неизвестна , пока неясно , выйдет на экраны . <eos>
- точная дата выхода фильма пока неизвестна , пока неясно , не уточняется . <eos>
- такое заявление сделал на заседании совета федерации хоккея россии , передает риа новости . <eos>
- соответствующее заявление сделал на заседании совета федерации хоккея россии по футболу , передает риа новости . <eos>
- соответствующее заявление сделал на заседании совета федерации по правам человека , передает риа новости . <eos>
- соответствующее заявление сделал на заседании совета федерации по правам человека , передает риа новости . <eos>
- соответствующее заявление сделал в четверг , dd мая , на сайте следственного комитета рф . <eos>
- соответствующее заявление сделал в четверг , dd мая , в ходе совещания . <eos>
- соответствующее заявление в четверг , dd мая , приводит риа новости . <eos>
- соответствующее постановление опубликовано на сайте ведомства в четверг , dd мая , в <unk> . <eos>

VAE/cyclical schedule

- в то же время , по словам <unk> , он посетит россию , а также в последние годы . <eos>
- в то же время , по словам <unk> , он посетит россию , а также в последние годы . <eos>
- в то же время , по словам <unk> , он был вынужден уйти в отставку . <eos>
- в <unk> , где он жил в нью—йорке , не уточняется . <eos>
- в <unk> , где он жил в нью—йорке , не уточняется . <eos>

- в результате <unk> погибли dd человек , большинство из которых были убиты . <eos>
- по его словам , <unk> был задержан в ходе проверки , проведенной полицией . <eos>
- по его словам , <unk> был задержан в ходе перестрелки с полицией . <eos>
- по предварительным данным , <unk> был ранен , а также ранен . <eos>
- причины катастрофы устанавливаются , <unk> в результате инцидента никто не пострадал . <eos>

Appendix 2

Here we present a few examples of grammatically correct and incorrect sentences generated by the VAE/Cyclical model.

Grammatically correct

- если вина будет доказана , <unk> грозит до dd лет лишения свободы . <eos>
- но это не первый случай , когда он будет <unk> в прямом эфире . <eos>
- в последние годы он жил в нью—Йорке и Вашингтоне . <eos>
- стоимость контракта оценивается в dd миллионов долларов . <eos>

Grammatically incorrect

- если бы не удастся , то <unk> , как , будет <unk> , то , как он , не будет делать какие—то проблемы , касающиеся <unk> изменений в закон " о <unk> " . <eos>
- поединок состоялся в ночь на D февраля , однако известно , что <unk> в нем принял участие около dd тысяч человек . <eos>

References

1. *Bowman, S. R. et al.*: Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349. (2015).
2. *Chang, J. et al.*: Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems. pp. 288–296 (2009).
3. *Fedus, W. et al.*: Maskgan: Better text generation via filling in the . arXiv preprint arXiv:1801.07736. (2018).
4. *Goodfellow, I. et al.*: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014).
5. *Guo, J. et al.*: Long text generation via adversarial training with leaked information. In: Thirty-second aaai conference on artificial intelligence. (2018).
6. *Kingma, D. P., Welling, M.*: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. (2013).

7. *Kipyatkova, I. S., Karpov, A. A.*: A study of neural network russian language models for automatic continuous speech recognition systems. *Automation and Remote Control*. 78, 5, 858–867 (2017).
8. *Kuratov, Y., Arkhipov, M.*: Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213. (2019).
9. *Kusner, M. J., Hernández-Lobato, J. M.*: Gans for sequences of discrete elements with the gumbel-softmax distribution. arXiv preprint arXiv:1611.04051. (2016).
10. *Liu, X. et al.*: Cyclical annealing schedule: A simple approach to mitigating kl vanishing. arXiv preprint arXiv:1903.10145. (2019).
11. *Merity, S. et al.*: Regularizing and optimizing LSTM language models. arXiv preprint arXiv:1708.02182. (2017).
12. *Merity, S. et al.*: Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843. (2016).
13. *Mikolov, T. et al.*: Empirical evaluation and combination of advanced language modeling techniques. In: 12th annual conf. of the international speech communication association. (2011).
14. *Mikolov, T. et al.*: Extensions of recurrent neural network language model. In: 2011 IEEE Intern. Conf. on acoustics, speech and signal proc. (ICASSP). pp. 5528–5531 IEEE (2011).
15. *Nesterenko, L.*: Building a system for stock news generation in russian. In: Proc. Of 2nd int. Workshop on nlg and the semantic web (webnlg 2016). pp. 37–40 (2016).
16. *Perera, R., Nand, P.*: Recent advances in natural language generation: A survey and classification of the empirical literature. *Computing and Informatics*. 36, 1, 1–32 (2017).
17. *Radford, A. et al.*: Language models are u/n-supervised multitask learners. *OpenAI Blog*. 1, 8, 9 (2019).
18. *Ruder, S. et al.*: Transfer learning in natural language processing. In: Proc. Of 2019 conf. Of the north american chapter of the association for computational linguistics: Tutorials. pp. 15–18 (2019).
19. *Sak, H. et al.*: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *research.google*. (2014).
20. *Shimorina, A. et al.*: Creating a corpus for Russian data-to-text generation using neural machine translation and post-editing. In: Proceedings of the 7th workshop on balto-slavic natural language processing. pp. 44–49 Association for Computational Linguistics, Florence, Italy (2019).
21. *Yang, Z. et al.*: Improved variational autoencoders for text modeling using dilated convolutions. In: Proc. 34th int. Conf. On machine learning-volume 70. pp. 3881–3890 JMLR. org (2017).
22. *Yu, L. et al.*: Seqgan: Sequence generative adversarial nets with policy gradient. In: 31st aai conf. On art. Intelligence. (2017).
23. *Zhang, Y. et al.*: Adversarial feature matching for text generation. In: Proc. 34th int. Conference on machine learning-volume 70. pp. 4006–4015 JMLR.org (2017).

ЛИНГВОСПЕЦИФИЧНЫЕ СЛОВА В ЗЕРКАЛЕ ПЕРЕВОДА: *ТОСКА*

Шмелев А. Д. (shmelev.alexei@gmail.com)

Институт русского языка им. В. В. Виноградова РАН;
Московский педагогический государственный университет,
Москва, Россия

DOI: 10.28995/2075-7182-2020-19-658-669

LANGUAGE-SPECIFIC WORDS IN THE LIGHT OF TRANSLATION: THE RUSSIAN *TOSKA*

Shmelev A. D. (shmelev.alexei@gmail.com)

Vinogradov Russian Language Institute of the Russian Academy
of Sciences; Moscow Pedagogical State University,
Moscow, Russia

This paper presents a semantic analysis of the most language-specific Russian word for 'sadness', namely, *toska*. The analysis is based on the hypothesis that one may regard translation equivalents and paraphrases of a linguistic unit extracted from real translated texts as a source of information about its semantics. The appearance of language-specific words in translated texts may be even more useful for studying their semantics.

It turns out that *toska* is not all that rare in Russian translated texts. The study of the incentives that lead Russian translators to use the word *toska* often reveals important aspects of the semantics of this word. Stimuli for the appearance of *toska* in translations into Russian vary greatly. In general, when the original describes some bad feelings, the word *toska* appears if the original speaks of a subject's unsatisfied desire, which desire may be vague and not well understood and it usually cannot be satisfied. In addition, the subject often feels lonely.

Key words: semantics, language-specific words, translation, parallel corpora

Традиционно в российской лингвистике в качестве языковых данных было принято использовать материал оригинальных текстов на русском языке, а также искусственные примеры, специально сконструированные носителями языка (чаще всего — самим исследователем). К примерам из переводных текстов принято было относиться с некоторым пренебрежением. Мне же кажется, что именно переводные тексты представляют собою ценнейший материал для изучения особенностей языка перевода.

С развитием корпусной лингвистики и появлением электронных параллельных корпусов русского языка возникла возможность использовать их для изучения семантики русских лингвоспецифичных слов и выражений более или менее систематически¹. В частности, выявление в оригинальных текстах стимулов, побуждающих переводчиков использовать лингвоспецифичные выражения языка перевода, позволяет не только верифицировать описание этих слов, полученное без использования электронных корпусов, но в определенных случаях по-новому взглянуть на те или иные лингвоспецифичные слова. Само появление таких слов в переводных текстах может показаться парадоксальным. В самом деле, если слово или выражение выражает конфигурацию смыслов, специфичную для языка перевода, то встает вопрос, какие же свойства оригинального текста побудили переводчика употребить это выражение. Часто ответ на этот вопрос позволяет уточнить описание семантики лингвоспецифичного слова или выражения. В указанном отношении именно переводы на русский язык показательны в большей степени, нежели переводы с русского.

Разумеется, для изучения особенностей лингвоспецифичных слов на основе их использования в переводных текстах существенно, чтобы язык этих текстов был близок языку авторских произведений на русском языке. Этот случай более других эффективен для изучения русских языковых конвенций, поскольку мы имеем дело с соответствующим языковым стандарту русским текстом, для которого уже готово его семантическое представление в виде иноязычного оригинала [Добровольский, Шмелев 2018: 46]. Именно таково большинство переводов из англо-русского подкорпуса «Национального корпуса русского языка» (НКРЯ), который и был положен в основу анализа². Об этом следует упомянуть, поскольку не все параллельные подкорпусы НКРЯ удовлетворяют данному требованию³.

Уже появился целый ряд работ, в которых русские лингвоспецифичные слова исследуются на основе их употребления в переводах на русский язык при помощи параллельных подкорпусов НКРЯ. Впрочем, почти все эти работы посвящены дискурсивным единицам русского языка. Но не менее показательны использование данных, извлеченных из переводов на русский язык в сопоставлении с оригинальными иноязычными текстами, для исследования семантики

¹ Идея об использовании данных параллельных корпусов для изучения лингвоспецифичных слов получила распространение в самое последнее время. Она высказывалась в том числе на конференциях «Диалог» [Шмелев 2015]; [Зализняк 2015]. Важно, что, в отличие от контрастивных корпусных исследований, при таком анализе в центре внимания оказывается один язык, и поэтому этот подход мы называли «унидирекциональным» [Шмелев, Зализняк 2017: 370]. Сейчас мы чаще используем слово «монофокусный».

² Исключение составляют включенные в НКРЯ переводы на русский язык англоязычных произведений Набокова, которые на этом основании были исключены из рассмотрения.

³ Так, по замечанию одного из рецензентов данной статьи, переводы, включенные в украинско-русский подкорпус НКРЯ, этому критерию не удовлетворяют, в отличие от переводов из англо-русского подкорпуса.

самостоятельных русских слов (в данной статье речь пойдет об одном из самых лингвоспецифичных слов русского языка, а именно — о слове *тоска*).

Однако изучение способов перевода лингвоспецифичных слов оригинальных русских текстов (и, в частности, слова *тоска*) тоже может оказаться до какой-то степени показательным, и с него и следует начать.

1. *Тоска* в русских оригинальных текстах и ее соответствие в переводах на английский язык

Как мы знаем, конкретные переводческие решения в отношении лингвоспецифичных слов могут опираться на разные соображения в соответствии с целями, которые ставит перед собою переводчик [Шмелев 2012]. Упрощенно говоря, можно различить две установки и, соответственно, две стратегии перевода. Одна из них ориентирована в первую очередь на автора в том смысле, что она направлена на то, чтобы передать все, что «имел в виду» («хотел сказать») автор. Для этой цели могут использоваться примечания, комментарии и т. п.; в результате адресат получает более или менее полное знание о семантическом содержании оригинального текста, но непосредственное впечатление разрушается. Другая установка в большей степени ориентирована на адресата и направлена на то, чтобы произвести на иноязычного адресата то же впечатление, которое оригинальный текст производит на носителя языка оригинала.

При этом выбор установки определяется целями перевода; не может быть универсальной стратегии, пригодной во всех случаях. Особенно это касается художественных переводов. Сталкиваясь с лингвоспецифичным языковым выражением, переводчик может ориентироваться на общий смысл высказывания, подбирая соответствие, в максимальной степени передающее смысл выражения в данном контексте (стратегия смысла), или же выбрать некоторый единый эквивалент данного выражения и стараться во всех случаях использовать именно его (стратегия конкорданса, или стратегия формы⁴). При выборе стратегии смысла часто возникает разброс переводов одного и того же лингвоспецифичного слова, поскольку единого эквивалента в языке перевода у него нет, и переводчик всякий раз вынужден выбирать соответствие, которое, по его мнению, ярче всего выражает смысловые компоненты, наиболее существенные в данном контексте⁵. Соответственно, обилие возможных переводов можно считать одним из признаков, свидетельствующих о высокой степени лингвоспецифичности слова, и, более того, разброс переводов может использоваться для количественной оценки лингвоспецифичности: чем более

⁴ «Стратегия формы», соответствующая так называемому *formal equivalence approach* (термин, предложенный Юджином Найдой), стремится к тому, чтобы подчеркнуть верность лексическим деталям и грамматической структуре оригинального текста [Nida and Taber 1969: 200].

⁵ Противопоставление «стратегии смысла» и «стратегии формы» при переводе художественного текста разбирается в статье [Добровольский, Шмелев 2018] и указанной там литературе.

разнообразны переводы лексической единицы в реально существующих переводных текстах, тем выше ее лингвоспецифичность [Шмелев 2015: 563].

Русское лингвоспецифичное слово *тоска*, как и следовало ожидать, обнаруживает разброс переводов на основные европейские языки, в том числе в пределах одного перевода одного и того же произведения. Это часто свидетельствует о том, что переводчик следует стратегии смысла, но, кроме того, может считаться дополнительным доказательством высокой степени лингвоспецифичности слова *тоска*. Обратимся к переводам на английский язык в НКРЯ и возьмем для примера переводы произведений Достоевского, выполненные Констанс Гарнетт. В переводе «Преступления и наказания» (романа, для которого мотив *тоски* является ключевым⁶) К. Гарнетт использует разнообразные соответствия. Самое частое соответствие в этом переводе (5 раз из 22 употреблений слова *тоска* в оригинальном тексте) — слово *misery*. Но в этом же переводе используются и другие соответствия: *wretchedness*, *anguish*, *dismay*, *depression*, *dejection*, *distress*, *agony*, *restlessness*, *misgivings*. В переводах других произведений Достоевского, выполненных К. Гарнетт, мы можем видеть не меньший разброс переводов. В качестве соответствия используются слова *yearning*, *distress*, *despondency*, *anguish*, *misery*, *trouble*, *depression*. При этом иногда Достоевский использует слово *тоска* в не вполне стандартном значении 'горе, страдание' (как во фразе из «Бесов» в эту несчастную неделю я вынес много тоски — в переводе К. Гарнетт *suffered a great deal during that unhappy week*), и отсюда такие переводы, как *grief* или *woe*.

Здесь уместно обратиться к толкованию слова *тоска*, данному А. Вежбицкой [Wierzbicka 1992: 172]. Упрощенно говоря, по Вежбицкой *тоска* — это когда человек чего-то хочет, сам точно не знает, чего именно, знает, что это не может произойти, и из-за этого нечто чувствует. В разных случаях на первый план может выходить любой из компонентов значения этого слова: желание чего-то (*yearning*), отсутствие точного понимания того, в чем состоит желаемое (этот компонент, как правило, не представлен в переводах К. Гарнетт), недостижимость желаемого и связанное с нею чувство, причем важно, что субъект чувствует нечто плохое (именно на этот компонент указывает большинство переводных эквивалентов, используемых К. Гарнетт)⁷.

«Братья Карамазовы» в НКРЯ тоже представлены переводом Констанс Гарнетт, и рассмотрение этого перевода на первый взгляд мало что добавляет к сделанным выводам. Однако сам по себе текст «Братьев Карамазовых» позволяет увидеть некоторые важные характеристики слова *тоска*, из которых часть можно считать общезыковыми, а часть — принадлежностью идиостиля Достоевского. Эти характеристики иногда находят отражение в переводе.

⁶ О роли *тоски* в поэтике «Преступления и наказания» см. [Топоров 1970: 100].

⁷ Многие из английских слов, которые К. Гарнетт использовала, как переводческие эквиваленты *тоски* и содержащих компонент 'произошло нечто плохое' (а именно, *sadness*, *unhappiness*, *distress*, *sorrow*, *grief*, *despair*) также были проанализированы А. Вежбицкой [Wierzbicka 1999: 60–70].

Можно обратить внимание на то, что, как и в «Бесах», в слове *тоска* в «Братьях Карамазовых» иногда на первый план выходит идея горя или страдания, и в этом случае переводчица иногда использует такие соответствия, как *grief* или *suffering* (используются и другие соответствия, уже знакомые нам по переводам «Преступления и наказания» и «Бесов», напр. *anguish*), но самым частым соответствием для *тоски* в этом переводе оказалось слово *depression*. Кажется, что в некоторых случаях она злоупотребляет им в качестве переводного эквивалента. Так, в следующем примере ярко проявляется компонент 'субъект чего-то очень хочет, но не знает, чего именно' (на него недвусмысленно указывает контекст):

- (1) *Тоска до тошноты*, а определить не в силах, чего хочу.
I feel sick with depression and yet I can't tell what I want.

Использование в переводе слова *depression* обнаруживает несовпадение его значения со значением слова *тоска*: последнее указывает на сильное желание, тогда как *depression* предполагает, что все желания уже позади⁸. В примере, когда Иван Карамазов хочет вступить на «новый, совершенно неведомый путь», «много надеясь, но не зная на что», перевод слова *тоска* как *depression* приводит к тому, что когда дело доходит до сочетания *тоска нового и неведомого* слово *тоска* переводчица передает как *apprehension*, в результате чего смысл изменяется почти на противоположный: в оригинале говорится о сильном желании Ивана, чтобы наступило нечто новое и неведомое, тогда как слово *apprehension* скорее указывает на боязнь, предчувствие чего-то опасного⁹:

Невозможным оказался перевод *depression* и в следующей реплике старца, обращенной к приезжей даме: *Я вполне верую в искренность вашей тоски*. Здесь существенным оказывается параметр *искренности*, который плохо применим к *depression*. В переводе использовано сочетание *the sincerity of your suffering*¹⁰.

Конечно, можно сказать, что переводы К. Гарнетт далеки от совершенства и к тому же в них нередко встречаются прямые кальки. Однако на тезис о разбросе переводов как показателе лингвоспецифичности это никак не влияет. В самом деле, наличие калек может свидетельствовать о желании хотя бы отчасти следовать стратегии формы; если, несмотря на это, одно и то же языковое выражение оригинала переводится по-разному, можно сделать вывод, что никакого однозначного выбираемого аналога исходному выражению в языке перевода нет и, следовательно, оно лингвоспецифично по крайней мере относительно языка перевода.

Обращение к переводам других русских текстов, выполненных другими переводчиками в русле стратегии смысла, обнаруживает едва ли не больший

⁸ Ср. [Wierzbicka 1999: 309]. Приведу также замечание из интервью, данного мною журналу *Esquire* в 2012: «Тоску можно разгонять, совершать от нее удалые действия, с тоски можно напиться. А в депрессии человек ничего не делает или идет к психотерапевту» (<https://esquire.ru/archive/2539-words/>).

⁹ Английское слово *apprehension* описано в книге [Wierzbicka 1999: 86–87]. Возможно, старые значения этого слова не обязательно предполагали страх или опасность.

¹⁰ Хотя русское слово *искренность* и английское *sincerity* тоже не вполне совпадают по смыслу [Wierzbicka 2002].

разброс переводов: *sadness, yearning, depression, restlessness, wretchedness, despondency, anguish, misery, distress, dismay, dejection* и т. д. Заметим также, что в переводах часто меняется часть речи и структура предложения, и при подобных изменениях в качестве соответствий могут использоваться такие слова, как *suffered, sadly, depressed, dejectedly, worried, miserable* и т. д.

Мы видим, что при переводе текстов, содержащих лингвоспецифичные слова, даже в тех случаях, когда переводчики следуют стратегии смысла, происходят определенные семантические сдвиги. В каждом конкретном случае переводчики решают (при условии, что они правильно понимают оригинальный текст), какими неявными компонентами смысла можно пожертвовать, а какие оказываются важными для смысла высказывания или всего текста¹¹. Именно эти сдвиги интересны для целей лингвистического анализа, поскольку они позволяют увидеть в значении исходного лингвоспецифичного выражения то, что определяет его специфику.

На этом фоне особенно обращает на себя внимание решение Ричарда Певеара и Ларисы Волохонской в переводе «Мастера и Маргариты» Михаила Булгакова. Слово *тоска* является одним из ключевых слов переводимого романа (оно встречается в нем 17 раз), и переводчики, в соответствии со стратегией формы, последовательно передают его как *anguish* (15 раз из 17; в двух случаях такой перевод уж совсем не годился бы) несмотря на то, что в некоторых случаях смысл ощутимо сдвигается¹². Слово *anguish* — весьма неточное переводное соответствие для *тоски*, оно указывает на душевную боль, муку или даже отчаяние, «острое и глубокое душевное страдание», которое «обычно бывает обусловлено не столько внешними, сколько внутренними, часто психологическими причинами» [Апресян и др. 1979: 403]. С *тоской* это чувство сближает некоторая иррациональность (субъект не знает, почему он чувствует нечто плохое), но, в отличие от *тоски*, *anguish* не предполагает компонента 'субъект чего-то хочет'. Показательно, что в англо-русском подкорпусе НКРЯ русские переводчики довольно редко переводят *anguish* как *тоска*, обычно выбирая слова со значением 'боль, мука, страдание'.

Сказанное побуждает обратиться еще к одному переводу, выполненному Певеаром и Волохонской, а именно — переводу «Преступления и наказания», пока не вошедшему в русско-английский параллельный подкорпус НКРЯ. Мы уже видели, что Певеар и Волохонская склонны к стратегии формы, и слово *тоска* они стремятся передавать единообразно (стандартное выбираемое ими соответствие *anguish*). Именно этой стратегии они следуют и в переводе

¹¹ Когда лингвоспецифичное слово оказывается ключевым для произведения, выбор стратегии смысла неизбежно затушевывает его роль. В связи с этим упомяну слово *тоска* в некоторых эпизодах «Ракового корпуса» Солженицына и его соответствия в английском и французском переводах, отсутствующих в НКРЯ [Шмелев 2019].

¹² Заметим, что и у К. Гарнетт есть перевод, в котором, в соответствии со стратегией формы, слово *тоска* переведено единообразно. Речь идет о рассказе Чехова «Тоска» (включенном в НКРЯ). В нем названная эмоция оказывается в центре повествования (в коротком рассказе оно встречается шесть раз, не считая заглавия), и разброс переводов полностью разрушал бы авторский замысел; К. Гарнетт перевела название рассказа как *Misery*, и точно так же она переводит это слово в пяти случаях из шести, а еще один раз несколько меняет синтаксическую структуру и использует родственное прилагательное *miserable*.

«Преступления и наказания», в котором, как мы помним, слово *тоска* встречается 22 раза. Приведем первые восемь примеров вхождения этого слова вместе с их переводом в версии Певеара и Волохонской:

- (2) он не знал, куда деться от *тоски* своей
he did not know where to flee from his *anguish*
- (3) Он так устал от целого месяца этой сосредоточенной *тоски* своей и мрачного возбуждения...
After a whole month of this concentrated *anguish*, this gloomy excitement of his, he was so tired out...
- (4) ... он был в беспокойстве, ерошил волосы и подпирал иногда, в тоске, обеими руками голову...
...he was agitated, kept ruffling his hair, and every once in a while leaned his head on his hands in *anguish*
- (5) Давным-давно как зародилась в нем вся эта теперешняя *тоска*...
Long, long ago this present *anguish* had been born in him...
- (6) целые полчаса бродил в *тоске* и тревоге
he have wandered for a whole half an hour in *anguish* and anxiety
- (7) *Тоска* проглянула в лице Лужина.
Anguish flitted across Luzhin's face.
- (8) с нетерпением и *тоской* поглядел на Настасью
looked at Nastasya with impatience and *anguish*
- (9) Даже *тоска* прошла...
Even his *anguish* had gone...

Выбранная переводчиками стратегия позволяет передать нагнетание тоски, характерное для «Преступления и наказания». При этом каждое отдельное вхождение слова *тоска* переведено с некоторым семантическим сдвигом: как уже говорилось, *anguish* указывает на эмоцию, вовсе не тождественную тоске.

Итак, и при выборе стратегии формы, и при выборе стратегии смысла неизбежны определенные потери и семантические сдвиги. Различие между указанными стратегиями чрезвычайно важно для теории перевода, а разброс переводов, возникающий при выборе стратегии смысла, свидетельствует о лингвоспецифичности переводимого языкового выражения. Однако с точки зрения проникновения в суть семантики лингвоспецифичного слова оригинала ни та, ни другая стратегия не может служить надежным инструментом.

2. Тоска в переводах на русский язык

Как выясняется, слово *тоска* используется переводчиками не так уж редко — скажем, переводчиками с английского языка текстов, включенных в НКРЯ, лишь менее чем в два раза реже, чем авторами оригинальных текстов: так, напр.,

англо-русский подкорпус НКРЯ (переводы на русский язык оригинальных текстов на английском языке) содержит (по состоянию на 31 августа 2018) 403 вхождения этого слова на 18 с небольшим миллионов словоупотреблений (т. е. 22 ipm), тогда как русско-английский подкорпус НКРЯ (оригинальные русские тексты с их переводами на английский язык) — 271 вхождение слова *toska* на более чем шесть с половиной миллионов словоупотреблений (т. е. 41 ipm). Возникает закономерный вопрос: какие же факторы обуславливают появление в русских переводах слова *toska*, ведь мы знаем, что в английском языке нет слова или выражения, для которого этот перевод был бы максимально точным соответствием.

Частично список слов и выражений, которые переведены как *toska*, совпадает со списком переводов русского слова *toska* на английский язык: *distress, depression, despair, anguish, grief, agony* и т. д., а также *sadness*; но часто слово *toska* соответствует другим оборотам. Так, некоторые переводчики регулярно используют его как перевод слова *melancholy*. Рассмотрение стимулов, которые побуждают русских переводчиков употребить слово *toska*, часто высвечивает важные компоненты в семантике этого слова, причем самым важным иногда оказывается общий контекст, в котором говорится о данной эмоции.

Слово *toska* часто появляется, когда в оригинале говорится о сильном желании чего-либо. Оно может ставиться в соответствие не только слову *yearning*, но и, напр., слову *longing*. Иногда объект желания вполне определен, но это не самая характерная ситуация для использования в переводе слова *toska*

- (10) I don't think I ever in my life, before or since, felt I wanted mustard as badly as I felt I wanted it then.

Никогда в жизни, ни прежде, ни потом, я не испытывал такой тоски по горчице, как в ту минуту.

Чаще всего слово *toska* используется, когда желание характеризуется как смутное и неопределенное (субъект затрудняется точно определить, чего именно он хочет). Именно в этом случае слово *toska* в переводе оказывается особенно уместным:

- (11) A sort of undefined longing crept upon them.

Какая-то смутная тоска напала на них.

- (12) It caused him to feel a vague, sweet gladness, and he was aware of wild yearnings and stirrings for he knew not what.

Бэк испытывал какую-то смутную радость, и беспокойство, и буйную тоску неведомо о чем.

Кроме того, когда в оригинале речь идет о подавленном, мрачном состоянии духа (напр., используются такие слова, как *gloom, tоре* или просто *excruciating feelings*), переводчики нередко бывают склонны охарактеризовать это состояние при помощи слова *toska*.

Часто *toska* связана с тем, что субъект чувствует себя одиноким:

- (13) I felt so lonesome I most wished I was dead. — Такая напала тоска, хоть помирай.

- (14) ...Dorothy built a splendid fire that warmed her and made her feel less *lonely*.
...Дороти разожгла большой костер, возле которого быстро согрелась и забыла уже подступившую к сердцу *тоску*.
- (15) He wanted to be alone—to be *lonely*.
Ему хотелось быть одному, отдалиться своей *тоске*.
- (16) During these days he got immensely *lonely*.
В эти дни его грызла *тоска*.
- (17) But he felt *lonely*.
Но на душе у него была *тоска*.
- (18) ...fear and *loneliness* goaded him.
...подстрекнули страх и *тоска*.

В целом, когда в английском оригинале говорится, что субъект чувствует нечто плохое, русские переводчики нередко характеризуют это чувство как *тоску*. При этом слово *тоска* часто появляется, когда в оригинале говорится о неудовлетворенном желании субъекта; само желание может быть смутным и безотчетным, но обычно таким, что оно не может быть удовлетворено. Мы видим, что это в основном подтверждает описание семантики слова *тоска*, данное А. Вежбицкой. Данные параллельного англо-русского подкорпуса НКРЯ позволяют уточнить компонент 'из-за этого субъект нечто чувствует', добавив, что он чувствует 'нечто плохое', а также высказать предположение, что к этому описанию может быть добавлен еще один компонент, связанный с тем, что субъект чувствует нечто плохое из-за того, что не может рассказать кому-либо о том, что именно он чувствует. Эта невозможность может быть обусловлена как смутностью и неопределенностью самого чувства, так и одиночеством субъекта (и, вероятно, поэтому упоминание одиночества в оригинале наводит переводчиков на мысль о *тоске*).

В связи со словом *тоска* и его английскими соответствиями можно также упомянуть статью [В. Ю. Апресян 2011], где высказывается предположение, что наиболее близким соответствием слову *тоска* является английское *blues* (в этой работе даже формулируется семантический инвариант для *тоски* и *blues*). Однако показательно, что в реальных переводах (в частности, по данным НКРЯ) *тоска* практически никогда не передается как *blues*, а *blues*, даже когда это эмоция, а не оттенки цвета или музыкальный жанр, обычно не переводят как *тоска*. Правда, Надежда Вольпин в переводе Голсуорси передает *blues* как *зеленая тоска*, забавным образом меняя цвет.

В целом представляется, что именно для таких слов, как *тоска*, анализ их появления в переводных текстах может помочь верифицировать или уточнить их описание и подчас оказаться более эффективным инструментом исследования, нежели анализ их соответствий в языке перевода, когда такие единицы появляются в языке оригинала. Перевод лингвоспецифичных слов оригинала представляет собою задачу, которая, как правило, бывает отрефлектирована переводчиком, и для решения этой задачи велика роль индивидуальных предпочтений переводчика. Кроме того, основанием для конкретного переводческого решения «могут оказаться поверхностное понимание, влияние переводных

словарей или даже желание придать тексту перевода некоторый налет «иностранности», чтобы текст воспринимался именно как перевод» [Михайлов 2005: 381]. Поэтому, когда русский язык выступает в качестве языка оригинала, данные параллельного корпуса часто дают нам недостаточную информацию о семантических особенностях русских лингвоспецифичных лексических единиц (хотя, разумеется, они чрезвычайно важны для переводоведения). Напротив того, появление лингвоспецифичных слов в переводе чаще всего представляет собою неосознанное решение переводчика как носителя языка и отражает его спонтанную речевую деятельность. Поняв, почему переводчик употребил данное лингвоспецифичное слово, мы получаем шанс увидеть какие-то его семантические характеристики, не замеченные при анализе оригинальных текстов.

Литература

1. Апресян Ю. Д. и др. 1979. Англо-русский синонимический словарь, Русский язык, Москва.
2. Апресян В. Ю. 2011. Опыт кластерного анализа: русские и английские эмоциональные концепты, Вопросы языкознания, 2, 63–88.
3. Добровольский Д. О., Шмелев А. Д. 2018. Русские лингвоспецифичные единицы, работа с ними при разных стратегиях перевода и русская конструкция что ни говори. Вопросы языкознания, 5, 34–48.
4. Зализняк Анна А. 2015. Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа, Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015, Москва, 651–662.
5. Михайлов М. Н. 2005. The particle in the text: is it possible to check correspondences of functional words in parallel corpora?, Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference 'Dialog 2005', Москва, 377–381.
6. Топоров В. Н. 1970. Поэтика Достоевского и архаичные схемы мифологического мышления («Преступление и наказание»), Проблемы поэтики и истории литературы, Мордовский государственный университет им. Н. П. Огарева, Саранск, 91–109.
7. Шмелев А. Д. 2012. Перевод «непереводимого», Логический анализ языка. Перевод художественных текстов в разные эпохи, Индрик, Москва, 44–54.
8. Шмелев А. Д. 2015. Русские лингвоспецифичные лексические единицы в параллельных корпусах: возможности исследования и «подводные камни», Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015, Москва, 561–570.
9. Шмелев А. Д. 2019. Языковое мастерство Солженицына как вызов переводчику, Александр Солженицын: взгляд из XXI века, Москва, 576–596.
10. Шмелев А. Д., Зализняк Анна А. 2017. Реверсивный перевод как инструмент лингвистического анализа дискурсивных слов, Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2017, Москва, 370–380.

11. *Nida E. A., Taber C. R.* 1969. *The Theory and Practice of Translation, With Special Reference to Bible Translating*, Brill, Leiden.
12. *Wierzbicka Anna* 1992. *Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations*, Oxford University Press, New York; Oxford.
13. *Wierzbicka Anna* 1999. *Emotions across Languages and Cultures: Diversity and Universals*, Cambridge University Press, Cambridge.
14. *Wierzbicka Anna* 2002. *Русские культурные скрипты и их отражение в языке (Russian cultural scripts and their reflection in language), Русский Язык в Научном Освещении (Russian Language and Linguistic Theory)*, 2 (4), 6–34.

References

1. *Apresjan Ju. D. et al.* (1979), *English-Russian Dictionary of Synonyms [Anglo-Russkii Sinonimicheskii Slovar']*, Russkii iazyk, Moscow.
2. *Apresjan V. Ju.* (2011), *An attempt at cluster analysis: Russian and English emotion concepts [Опыт кластерного анализа: русские и английские эмоциональные концепты]*, *Topics in the study of language [Voprosy yazykoznaniya]*, 2, 63–88.
3. *Dobrovol'skij D. O., Shmelev A. D.* (2018), *Russian language-specific items and their correlates in translation: The case of chto ni govori 'say what you will' [Russkie lingvospetsifichnye edinitsy, rabota s nimi pri raznykh strategiyakh perevoda i russkaya konstruktsiya chto ni govori]*, *Topics in the study of language [Voprosy yazykoznaniya]*, 5, 34–48.
4. *Mihajlov M. N.* (2005), *The particle in the text: is it possible to check correspondences of functional words in parallel corpora?*, *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference 'Dialog 2005'*, 377–381.
5. *Nida Eugene A. and Taber Charles R.* (1969), *The Theory and Practice of Translation, With Special Reference to Bible Translating*, Brill, Leiden.
6. *Shmelev A. D.* (2012), *Translating the untranslatable [Perevod «neperevodimogo»]*, *Logical Analysis of Language. Translating Fiction in Various Epochs [Logicheskij Analiz Yazyka. Perevod Khudozhestvennykh Tekstov v Raznye Epokhi]*, Indrik Publ., Moscow, 44–54.
7. *Shmelev A. D.* (2015), *Russian language-specific lexical units in parallel corpora: Prospects of investigation and “pitfalls”*, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue' 2015*, Moscow, pp. 584–594.
8. *Shmelev, A. D.* (2019), *Solzhenitsyn's language as a challenge that translators face [Iazykovoje masterstvo Solzhenitsyna kak vyzov perevodchiku]*, *Alexandr Solzhenitsyn: Looking Back from the 21st Century [Aleksandr Solzhenitsyn: Vzgliad iz XXI Veka]*, *Russkii put'*, Moscow, 576–596.
9. *Shmelev A. D., Zaluzniak Anna A.* (2017), *Reverse translation as a tool for analysis of discourse words*, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue' 2017*, Moscow, pp. 370–380.

10. *Toporov V. N.* (1970), *Dostoevsky's Poetics and Archaic Patterns of Mythological Thought: Crime and Punishment* [Poetika Dostoevskogo i arkhainnye skhemy mifologicheskogo myshleniya («Prestuplenie i nakazanie»)], *Problems of Poetics and the History of Literature* [Problemy Poetiki i Istorii Literatury], Saransk, 91–109.
11. *Wierzbicka Anna* (1992), *Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations*, Oxford University Press, New York; Oxford.
12. *Wierzbicka Anna* (1999), *Emotions across Languages and Cultures: Diversity and Universals*, Cambridge University Press, Cambridge.
13. *Wierzbicka Anna* (2002), *Russian cultural scripts and their reflection in language* [Russkie kul'turnye skripty i ikh otrazhenie v yazyke], *Russian language and linguistic theory* [Russkij yazyk v nauchnom osveshchenii], 2 (4), 6–34.
14. *Zalizniak Anna A.* (2015), *Russian language-specific words as an object of contrastive corpus analysis*, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue' 2015, Moscow*, pp. 651–662.

TAGGING AND PARSING OF MULTIDOMAIN COLLECTIONS

Sorokin A. A. (alexey.sorokin@list.ru)

Moscow Institute of Physics and Technology, Neural Networks and Deep Learning Lab, Dolgoprudny, Russia;
Moscow State University, Faculty of Mathematics and Mechanics, Moscow, Russia

Smurov I. M. (ivan.smurov@abbyy.com)

ABBYY, Moscow, Russia;
Moscow Institute of Physics and Technology, Computer Vision and Natural Language Processing Laboratory (ABBYY Lab), Dolgoprudny, Russia

Kirianov D. P. (denkirjanov@gmail.com)

SberDevices, Moscow, Russia

In this paper we describe our submission to GramEval2020 competition on morphological tagging, lemmatization and dependency parsing.

Our model uses biaffine attention over the BERT representations. The main feature of our work is the extensive usage of language model, tagger and parser fine-tuning on several distinct genres and the implementation of genre classifier. To deal with dataset idiosyncrasies we also extensively apply handwritten rules.

Our model took second place in the overall model performance scoring 90.8 aggregate measure over all 4 tasks.

Keywords: morphological tagging, lemmatization, dependency parsing, domain adaptation, fine-tuning, BERT

DOI: 10.28995/2075-7182-2020-19-670-683

МОРФОЛОГИЧЕСКИЙ И СИНТАКСИЧЕСКИЙ АНАЛИЗ МУЛЬТИЖАНРОВЫХ ТЕКСТОВЫХ КОЛЛЕКЦИЙ

Сорокин А. А. (alexey.sorokin@list.ru)

Московский Физико-технический Институт, Лаборатория нейронных систем и глубокого обучения, Долгопрудный, Россия; Московский Государственный Университет, механико-математический факультет, Москва, Россия

Смулов И. М. (ivan.smurov@abbyy.com)

АВВУ, Москва, Россия; Московский Физико-технический Институт, Лаборатория компьютерного зрения и обработки естественного языка (АВВУ Lab), Долгопрудный, Россия

Кириянов Д. П. (denkirjanov@gmail.com)

SberDevices, Москва, Россия

В данной статье описана наша модель, использованная в соревновании GramEval2020 по морфологическому анализу, лемматизации и синтаксическому анализу. Наша модель основана на бифинном механизме внимания и архитектуре БЕРТ. Её отличительной чертой является использование отдельной модели БЕРТ для каждого жанра и настройка базовой модели на доменных данных, а также использование правил для унификации разметки.

Наша модель заняла второе место в соревновании, показав среднее качество 90,8% по 4 заданиям и 6 предметным областям, в то время как результат победителя составил 91,7%.

Ключевые слова: морфологический анализ, лемматизация, синтаксический анализ, дообучение, БЕРТ

1. Introduction

Automatic processing of morphology and syntax have been a part of Natural Language Processing (NLP) for several decades. Introduction of end-to-end deep learning pipelines [7], embeddings pretraining ([17], [18]), and char-level features [16] have all contributed to rapid improvement of NLP in general and grammatical features extraction systems in particular. This trend was enhanced by a recent introduction of pretrained language models and context-dependent embeddings such as ELMo [22] and BERT [8] leading to a drastic improvement for an overwhelming majority of NLP tasks including grammatical tagging and parsing.

Dependency parsing has seen a similar evolution of its own in the last half-decade. This process was kick-started by [5] who introduced deep learning into transition-based parsing [20]. Two years later [12] were able to successfully utilize RNNs in both transition-based and graph-based dependency parsing. While their work allowed for a remarkable increase of graph-based parsing quality it wasn't until [9] model before graph-based parsing dethroned transition-based parsing as a state-of-the-art (SOTA) parsing approach. Most models introduced after 2017 follow the path tread by Dozat and Manning and implement biaffine attention with different feature sets, first utilizing ELMo and most recently BERT.

Currently the best performing dependency parsers for English are immediate successors of HPSG model introduced in [24]. The most important feature of this family of parsers is joint learning on dependency and constituency trees. Unfortunately, this makes utilizing HPSG-style parsing for Russian an extremely complicated task since there are no publicly available annotated corpora or constituency parsers. While we would like to explore the possibility of utilizing proprietary parsers such as Compreno [1], for now this remains for future work. Given these considerations we decided to base our model on biaffine attention of Dozat and Manning [9] with BERT-based token features. This approach is common in modern NLP and is utilized, e.g., in [13]

Section 2 describes GramEval-2020 shared task and the corpora made available for it. **Section 3** gives an overview of our model. **Section 4** describes the process of training and provides the evaluation results. **Section 5** contains analysis of our model performance as well as the discussion on the representativeness of the Shared Task results for the processing of Russian morphology and syntax. Finally, **Section 6** provides conclusion and outlines our plans for future work.

2. GramEval-2020 and Corpus Analysis

GramEval-2020 [21] is a shared task on part-of-speech (POS) and full morphological tagging, lemmatization and dependency parsing of Russian texts. Parsing was scored with labeled attachment score (LAS) while the other three tasks with accuracy. The participants systems were ranked by an aggregate measure on all four tasks.

For training GramEval-2020 organizers provided not a single corpus, but rather a collection of several disjoint corpora of different origins and genres.

Train set consisted of the following subcorpora:

1. SynTagRus [10] corpus of dependency parses from UD [19] (\approx 62k sentences). Mostly contains texts of general domain.
2. MorphoRuEval2017 [23] morphologically labeled corpus with semi-automatic syntactic annotation. Contains texts of general domain and of social networks from GIKRYA.
3. Poetry corpus (\approx 0.9k sentences). Contains poetic texts from Taiga.
4. Social networks corpus (\approx 2.3k sentences). Contains social media texts from Taiga.
5. Wikipedia texts from GSD (about 5k sentences). Contains texts of general domain and technical texts.

6. XVII century corpus(\approx 1.2k sentences). Contains Middle Russian texts both in original and in an adapted orthography (where symbols not present in modern Russian were substituted by their closest analogues e. g. ” was substituted with ‘E’; the rest of spelling remained unchanged). In total, approximately 60% texts of train set were in original orthography and 40% in adapted.

Development set consisted of subcorpora 3–6 as well as news subcorpora from Lenta.ru (each subcorpus contained 40–70 sentences). Test set included the same sources as the development one and a fiction subcorpus.

3. Model Overview

3.1. Pipeline

The Shared Task data clearly consists of 4 isolated segments whose syntax may differ significantly: social media, poetry, historical (XVII century) texts and general domain. For the XVII century subcorpus its morphology and even graphics is also specific. We expect that there is no single tagger and parser that works for all domains equally well. Therefore we apply a separate model to each domain. Consequently, we use the following pipeline (see subsequent subsections for the description of its components):

1. The classifier predicts a domain label given the input sentence.
2. The morphological tagger outputs a sequence of morphological tags given the tokenized sentence.
3. The lemmatizer yields the source form of the word based on its tag and the word itself.
4. The dependency parser reconstructs the dependency tree given the tokenized sentence. The parser does not take morphological tags into account.
5. Several rule-based postprocessors modify tag and lemma to match the annotation standards. Some of the postprocessors use only the word, its lemma and tag, several other are domain-specific and hence use the class label as well. Some postprocessors also utilize tags of the words in the neighbourhood to fill the missing morphological features.

3.2. Classifier

We experimented with three different architectures of domain classifier: fastText [3] classifier, BERT [8] classifier and a classifier based on logistic regression over character ngrams. The two latter models have shown comparable results (approximately 91% macro F-score). Since logistic regression is both less computationally expensive and easier to interpret and tune we decided to choose the latter one.

We trained the domain classifier on the provided training data. Since the development set is too small to measure performance on it, we left one-quarter of the original training set as the held-out data (this approach was used only to detect the genres).

3.3. Morphological tagger

Both part-of-speech and full morphological tagging can be considered as sequence labeling tasks. Thus one can approach both tasks using standard sequence labeling techniques. In our model we used BERTs for tagging in the same way as [8] treats named-entity recognition task (see figure below).

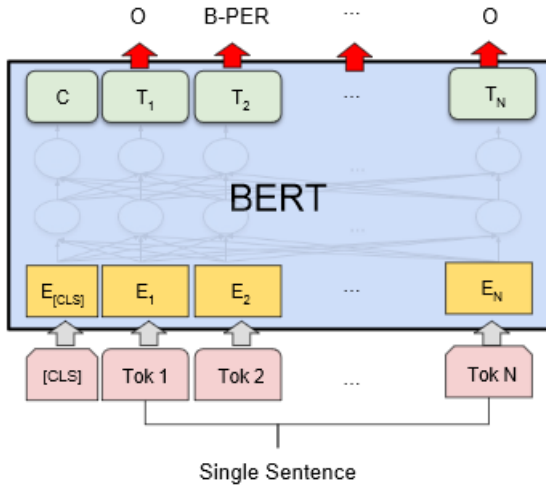


Figure 1: Illustration from [8]

3.4. Lemmatizer

In order to generate possible lemmas we used PyMorphy 2.0 [14] for initial lemma generation and several postprocessing rules to deal with format mismatch. Since PyMorphy outputs several possible variants, we select the one corresponding to the tag predicted by the morphological tagger using the DeepPavlov library [4].

3.5. Dependency parser

As mentioned earlier, for dependency parsing we used biaffine attention network [9] built over BERT contextualized word features (fine-tuned separately for each genre).

Going into more detail, given a contextualized token representation w_i , we generate 4 vectors $head_i^{(arc)}$, $dep_i^{(arc)}$, $head_i^{(rel)}$ and $dep_i^{(rel)}$ using multi-layered perceptrons. $head_i^{(arc)}$ and $dep_i^{(rel)}$ are used to predict the probability of a token being an arc head and arc dependent respectively. In order to generate score for arc from token w_i to token w_j we use biaffine attention as follows:

$$s_{ij}^{(arc)} = head_i^{(arc)} \cdot A \cdot dep_j^{(arc)} + head_i^{(arc)} \cdot b.$$

These scores were transformed to probabilities using standard SoftMax layer. These probabilities are used to produce a maximum spanning tree i. e. dependency tree with the highest probability using Chu-Liu/Edmonds algorithm [6], [11].

Given the dependencies, $head_i^{(rel)}$ and $dep_i^{(rel)}$ can be interpreted and combined in a similar way to generate a probability distribution over all possible dependency labels using a biaffine classifier.

3.6. Postprocessors

We apply several rule-based postprocessors. Their main goal is to manually transform the outputs of all models to the same format. For example, not all datasets annotate Animacy of the adjectives. This information can be copied from the parent node of the adjective, however, this requires the presence of syntactic tree. Other postprocessors use only the word itself or/and its lemma/tag. Our final model contains the following postprocessing stages:

1. Emoji postprocessor that uses the Emoji library¹.
2. Adjective animacy postprocessor.
3. Pronoun *что/который* “what/which” postprocessor, that fills the gender/number information for these words. It finds the antecedent of the pronoun by traversing the dependency tree using rule-based instructions and copies the relevant feature values from it.
4. Digit postprocessor. It explores whether the digit satisfies some frequent patterns (such as *3 июня* “3rd of June” or 1917–1920) and calculates the required features (NumType and Case/Number/Gender when applicable).

4. Models and results

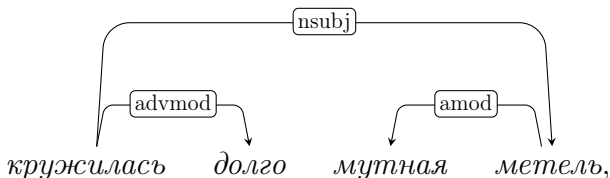
4.1. Training approaches

In our study we apply several approaches to model training including

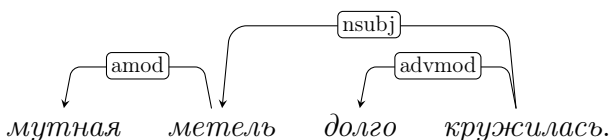
1. Standard supervised model training. In this case we initialize the embedder with the weights of pretrained BERT language model and train the whole network together. This mostly affects the task-specific head layers, however, the weights of the embedder are also altered.
2. BERT language model unsupervised finetuning [2], [15]. If we suspect that dataset domain (e. g., poetry) does not match the domain BERT was trained on, we additionally train the masked language model on the domain-specific raw data.

¹ <https://pypi.org/project/emoji/>

- Supervised model finetuning. We initialize the model weights using the weights of basic supervised model (see Approach 1) and tune them on the domain-specific annotated data. We follow this strategy when there is not enough such data for learning the weights from scratch, however, the basic model seems to be less suitable due to data peculiarity.
- Word order adaptation: the word order in verses often differs from the one in formal speech. Consider the verse



its “formal” version would be



More precisely, though the word order in Russian is flexible, for most syntactic dependencies the relative positions of the head and the dependent are rather predictable. In particular, the subject (nsubj) usually precedes the verb, the adjective modifier precedes the corresponding noun, etc. However, for poetic texts these patterns are less strict.

Hence, to make the formal prose more similar to the poetry, we randomly switch the ordering of the head and the dependent for those syntactic relations whose dependent-head order in poetic texts deviates from the one in formal corpus. The switch probability is set to match the order distribution in the poetic training set. Note that this dataset is not large enough to train the model only on it without utilizing the general domain in some fashion.

Graphic adaptation. The most challenging domain in the test data is the XVII century Russian language. Since its alphabet contains several symbols that are not present in the modern Russian, BERT tokenizer is not able to divide such words to sub-tokens in a proper way. Consider the word *человкъ* (the man), its tokenization will be

челов/##/##къ.

On the contrary, the modern Russian spelling of the same word (*человек*) produces a single subtoken for the whole word. It implies that after tokenization a word in old orthography loses its similarity to the corresponding word in modern orthography, hence the model is unable to utilize the unsupervised knowledge stored in subtoken embeddings. To overcome this obstacle we apply several handwritten rules, such as

- Removal of word-final -ѡ.
- $\rightarrow e, \rightarrow u$ and analogous modifications.
- Changing word-final *-ти* to *-ть* in verb infinitives, etc.

4.2. Model selection and evaluation

In this section we describe how we select the optimal model for different parts of the dataset. The selection procedure for the tagger and lemmatizer differs from the one for the parser, so we discuss them separately. However, they share several common steps, which we list below.

1. First, we found that data annotation is inconsistent through the training data, as different segments of it were annotated using different annotation standards. Moreover, inspecting the training data we found many errors and artefacts of automatic annotation, therefore we decided to rely only on some subsets of the dataset. Namely, we selected the SynTagRus v2.5 dataset to be the only training source for our basic model². The remaining parts of the training set are used only for validation and finetuning purposes.
2. Since our classifier has 4 possible domain labels ('17cent', 'poetry', 'social' and 'other'), we picked a subcorpus for each of these categories. For first three categories the choice is unambiguous, while the performance on 'other' was evaluated on Wikipedia GSD subcorpus and Lenta News subcorpus from 2017 MorphoRuEval competition [23].
3. Since we have no access to the correct annotation of the test set, we make conclusions using the official training data, as only the SynTagRus part of the training set was used for pure training. We used several parts of the training set for model finetuning, in this case we report performance on the development set.
4. During our preliminary experiments we found that altering BERT embedders and tuning the model has no significant effect on the tagger performance, therefore we performed experiments on BERT and model finetuning only for the syntactic parser.
5. All our models are based on ruBERT model [15] from DeepPavlov library [4]. This model was obtained by finetuning the multilingual BERT [8] on Russian language data.

4.3. Tagger performance

We present the performance of our basic model on 5 mentioned domains: '17cent', 'poetry', 'social', 'wiki' and 'news' (the two latter being the part of 'other'). We also present the scores of the finetuned 17th century model on '17 cent' segment. For comparison purposes we also train another variant of the basic model on the unified training set and evaluate it on development set. We present scores on development set only.

² The organizers informed us that SynTagRus guidelines does not match the annotation of the test set. However, they did not answer if there was any other subset better corresponding to the annotation of the test data. Using SynTagRus, we at least expect different parts of our training set not to contradict each other.

Table 1: The results of different morphological taggers on the training and development set. We report POS accuracy(P), lemmatization accuracy (L) and morphological feature recall (F) using the official evaluation script.

Model	wiki			news			poetry			social			17cent		
	F	L	P	F	L	P	F	L	P	F	L	P	F	L	P
Basic	96.1	96.6	94.5	95.9	98.4	96.7	94.1	97.1	92.8	95.1	98.4	93.8	87.1	51.7	92.8
1 + 17cent rules	—	—	—	—	—	—	—	—	—	—	—	—	92.6	86.2	94.4
2 + 17cent finetuned	—	—	—	—	—	—	—	—	—	—	—	—	96.0	86.8	98.0
joint	98.4	96.9	98.2	97.8	98.8	98.4	95.6	97.1	95.0	97.2	98.3	95.7	94.2	51.2	96.6
+ 17cent rules	—	—	—	—	—	—	—	—	—	—	—	—	94.2	85.4	96.7

We observe that data-specific rules for 17 century data actually help a lot, finetuning on more 17 century data also improves the model drastically. More surprising is the fact that the joint model beats the one trained only on SynTagRus by a notable margin. However, this question requires further investigation.

4.4. Parser performance

We expect the parser to depend more severely from the training domain. Therefore here we perform a much more detailed comparison, which is performed in several stages. First, we want to select an optimal BERT embedder for each of the domains. We test three BERT models.

1. The default ruBERT model from DeepPavlov library³.
2. The Conversational BERT model from DeepPavlov library⁴ finetuned on social network data from Taiga corpus.
3. The StihBERT model, finetuned on poetry data from Taiga corpus.

All these models are trained only on SynTagRus data. The results are presented in **Table 2**.

Table 2: The effect of different BERT embedders on syntactic parsing. We report Labeled Attachment Score (LAS) for training and development set.

Model	wiki		news		poetry		social		17cent	
	T	D	T	D	T	D	T	D	T	D
ruBERT	83.8	86.7	92.9	92.9	69.1	79.0	77.1	82.8	59.0	72.8
ConvBERT	83.3	86.5	93.0	92.0	71.4	80.2	78.4	83.2	54.1	72.8
StihBERT	—	—	—	—	72.5	81.5	—	—	—	—

³ <http://docs.deeppavlov.ai/en/master/features/models/bert.html>,
http://files.deeppavlov.ai/deeppavlov_data/bert/rubert_cased_L-12_H-768_A-12_v2.tar.gz

⁴ http://files.deeppavlov.ai/deeppavlov_data/bert/ru_conversational_cased_L-12_H-768_A-12.tar.gz

We observe that using domain-specific BERT improves results on social and poetry subsets, as expected. Consequently, we decide to use StihBERT for poetic data, ConvBERT for the social media and the basic ruBERT for the remaining.

We also evaluate the effect of fine-tuning on heritage, social and poetic data (for ‘other’ domain the results are controversial). **Table 3** contains the results for ‘social’ and ‘poetry’ domains. “No finetuning” means choosing the best BERT for a given domain between the models evaluated in **Table 2**. We also present the results on the ‘poetry’ data for the model trained on the dataset with switched order of heads and dependencies, as discussed in **Subsection 4.1**. For comparison we give the scores of the ‘joint’ model, as it can be viewed as the model fine-tuned on the concatenation of all training data available.

Table 3: The effect of fine-tuning on syntactic parsing. We report Labeled Attachment Score (LAS) for different parts of the development set.

Model	wiki		news		poetry		social	
	T	D	T	D	T	D	T	D
No finetuning	83.8	86.7	92.9	92.9	72.5	81.5	78.4	83.2
+FT(social)	—	87.0	—	91.7	—	78.9	—	85.7
+FT (poetry)	—	—	—	—	—	81.9	—	—
+switch	—	—	—	—	—	82.4	—	—
joint	—	87.0	—	91.2	—	74.0	—	81.5

We observe the positive effect of fine-tuning. Also note that fine-tuning the model on ‘social’ data decreases its performance on other domains. In contrast to morphological tagging, the model trained on joint data has significantly lower performance. We suppose that one of the reasons may be inconsistent annotation of syntactic phenomena in different training subsets.

We also compare the basic model with the fine-tuned one of the 17 century data. The results are given in **Table 4**. Here we again observe the positive influence of model fine-tuning.

Table 4: The effect of fine-tuning on syntactic parsing on 17 century data

Model	17 cent	
	T	D
No finetuning	59.0	72.8
+rules	63.5	73.6
+rules+FT	—	85.8
joint	—	78.3
+rules	—	78.3

5. Analysis and Discussion

Table 5 and **Table 6** contain the official GramEval-2020 results.

Table 5: GramEval results on historic texts (17) and fiction (fict)

corpus	all	17				fict			
team	overall	POS	morph	lemmas	LAS	POS	morph	lemmas	LAS
qbic	91.6	96.3	93.0	78.3	66.5	98.0	98.8	98.1	89.6
ADVance (our model)	90.8	96.0	93.0	79.7	61.9	98.0	98.6	97.7	87.0
lima	87.9	93.5	89.6	61.1	55.5	97.6	97.9	93.7	85.1
vocative	85.2	87.1	79.4	58.3	50.0	97.5	94.8	96.2	82.7

Table 6: GramEval results on news and poetry (poet)

corpus	all	news				poet			
team	overall	POS	morph	lemmas	LAS	POS	morph	lemmas	LAS
qbic	91.6	96.7	98.1	98.3	91.3	96.9	96.7	95.4	81.4
ADVance (our model)	90.8	96.5	98.2	98.2	91.2	96.1	96.0	95.3	78.1
lima	87.9	97.2	96.7	95.0	84.4	95.8	95.6	91.3	72.6
vocative	85.2	96.6	94.5	95.5	83.5	92.3	89.8	93.9	66.0

Table 7: GramEval results on social media (soc) and wikipedia (wiki)

corpus	all	soc				wiki			
team	overall	POS	morph	lemmas	LAS	POS	morph	lemmas	LAS
qbic	91.6	94.8	94.7	96.0	80.7	92.7	94.4	93.6	78.1
ADVance (our model)	90.8	93.8	95.9	95.4	78.5	92.2	92.3	92.2	76.1
lima	87.9	93.7	95.3	95.3	71.3	92.5	96.8	92.3	69.8
vocative	85.2	91.8	90.0	95.5	66.0	91.0	90.5	91.6	69.5

Analyzing them, one can notice that our model and the winner’s model significantly outperform the other two models. The main difference of the top two models is the usage of BERT. As expected using BERT provides for a significant advantage.

The overall gap between our model and the winner’s model is relatively small but consistent across domains. This is mostly due to dependency parsing performance: while on other tasks the scores of these two systems are comparable, our LAS scores are considerably lower on all corpora but news.

After the release of all systems into the open source we have spent some time analyzing the winner’s model. Both models use similar architecture and the same ruBERT embedder, however, there is a number of differences in training procedure. So far we were not able to isolate the decisive one, but we would like to explore it more in future work.

5.1. Data quality

Before criticising the datasets provided for the Shared Task, we would like to deeply thank the organizers for their work, which is an important contribution for future studies on computational syntax and morphology of Russian. However, the quality of the provided training data makes the conclusions not so reliable as they could be.

First, the annotation of some morphological phenomena is inconsistent. For example, the label assigned to foreign proper nouns is `NOUN`, `PROPN` or `SYM` depending from the segment. This problem holds for the annotation of proper nouns in general, a word may have controversial labels even in consecutive sentences.

Second, the annotation is inconsistent even inside segments. Namely, while the training data for XVII century contains a significant fraction of texts in original orthography, the development set is entirely in modern (adapted) one. Many specific syntactic relations (e.g., `nsubj:pass` or `det`) are occasionally replaced with their more general analogues (`nsubj` and `amod`). The same problem holds for fixed constructions, e.g. complex prepositions as *со стороны* “by”, which are not annotated in most of the training corpora, being present only in one of them. Additionally, the annotation of punctuation is also inconsistent, which produces many spurious errors that do not reflect the actual performance of the model.

Last but not the least, the significant amount of data is annotated automatically using a model of rather low quality. It sometimes yield nonsense errors, for example, the noun *джакузи* has the lemma **джакузить* in the training set.

6. Conclusion and Future Work

We have presented ADVance—a system performing part-of-speech and full morphological tagging, lemmatization and dependency parsing for Russian. Our system has participated in GramEval-2020 shared task and was able to reach second place. Our morphological tagger uses BERT as contextualized embedder and the parser system is based on biaffine attention over BERT representations. We release our system in open source⁵.

Our main scientific contribution is the relative success of domain adaptation and fine-tuning approaches, that goes in line with previous studies. However, the results on dependency parsing on challenging poetry and XVII century domains are well below the scores for more formal texts. Additionally, these scores are lower than the ones reported for Universal Dependencies datasets, where the Basic version of our model achieves LAS over 93%. That is partially due to train-test annotation mismatch, however, this is a common real world situation. We hope that our study will help to shed light on practical aspects of training morphological and syntactic analyzers on real-world data with imperfect annotation.

⁵ <https://github.com/AlexeySorokin/GramEval2020>

7. Acknowledgements

We thank the organizers of Shared Task Olga Lyashevskaya and Tatiana Shavrina for providing the data and holding the competition. We are also grateful to members of DeepPavlov team for their assistance in BERT fine-tuning. We thank the anonymous reviewers whose valuable comments helped to improve the paper.

References

1. *Anisimovich, K. et al.*: Syntactic and semantic parser based on abbyy compreno linguistic technologies. In: Computational linguistics and intellectual technologies: Proceedings of the international conference “dialog” [komp’iuternaia lingvistika i intellektual’nye tehnologii: Trudy mezhdunarodnoj konferentsii “dialog”]. pp. 90–103, Bekasovo, Russia (2012).
2. *Arkipov, M. et al.*: Tuning multilingual transformers for language-specific named entity recognition. In: Proceedings of the 7th workshop on balto-slavic natural language processing. pp. 89–93 (2019).
3. *Bojanowski, P. et al.*: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 5, 135–146 (2017).
4. *Burtsev, M. et al.*: Deeppavlov: Open-source library for dialogue systems. In: Proceedings of acl 2018, system demonstrations. pp. 122–127 (2018).
5. *Chen, D., Manning, C. D.*: A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp). pp. 740–750 (2014).
6. *Chu, Y.-J., Liu, T.-H.*: On the shortest arborescence of a directed graph. Scientia Sinica. 14, 1396–1400 (1965).
7. *Collobert, R. et al.*: Natural language processing (almost) from scratch. Journal of machine learning research. 12, Aug, 2493–2537 (2011).
8. *Devlin, J. et al.*: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR. abs/1810.04805, (2018).
9. *Dozat, T., Manning, C. D.*: Deep biaffine attention for neural dependency parsing. arXiv preprint arXiv:1611.01734. (2016).
10. *Dyachenko, P. et al.*: Sovremennoe sostoyanie gluboko annotirovannogo korpusa tekstov russkogo yazyka (syntagrus). [The current state of the deeply annotated corpus of russian texts (syntagrus)]. Trudy Instituta Russkogo Yazyka im. V. V. Vinogradova. 6, 272–300 (2015).
11. *Edmonds, J.*: Optimum branchings. Journal of Research of the national Bureau of Standards B. 71, 4, 233–240 (1967).
12. *Kiperwasser, E., Goldberg, Y.*: Simple and accurate dependency parsing using bidirectional LSTM feature representations. Transactions of the Association for Computational Linguistics. 4, 313–327 (2016).
13. *Kondratyuk, D.*: 75 languages, 1 model: Parsing universal dependencies universally. arXiv preprint arXiv:1904.02099. (2019).
14. *Korobov, M.*: Morphological analyzer and generator for russian and ukrainian languages. 542, 320–332 (2015).

15. *Kuratov, Y., Arkhipov, M.*: Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213. (2019).
16. *Lample, G. et al.*: Neural architectures for named entity recognition. In: Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies. pp. 260–270 Association for Computational Linguistics, San Diego, California (2016).
17. *Mikolov, T. et al.*: Efficient estimation of word representations in vector space. CoRR. abs/1301.3781, (2013).
18. *Mikolov, T. et al.*: Distributed representations of words and phrases and their compositionality. In: Burges, C. J. C. et al. (eds.) Advances in neural information processing systems 26. pp. 3111–3119 Curran Associates, Inc. (2013).
19. *Nivre, J. et al.*: Universal dependencies v1: A multilingual treebank collection. In: Proceedings of the tenth international conference on language resources and evaluation (lrec’16). pp. 1659–1666 (2016).
20. *Nivre, J. et al.*: Maltparser: A data-driven parser-generator for dependency parsing. In: LREC. pp. 2216–2219 (2006).
21. *Olga, L., Tatiana, S.*: GramEval 2020 Shared Task: Russian Full Morphology and Dependency Parsing. In: Computational linguistics and intellectual technologies: Papers from the annual conference “Dialogue”. (2020).
22. *Peters, M. E. et al.*: Deep contextualized word representations. CoRR. abs/1802.05365, (2018).
23. *Sorokin, A. et al.*: MorphoRuEval-2017: An evaluation track for the automatic morphological analysis methods for russian. 2, 297–313 (2017).
24. *Zhou, J., Zhao, H.*: Head-driven phrase structure grammar parsing on Penn treebank. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp. 2396–2408 Association for Computational Linguistics, Florence, Italy (2019).

VISUAL VS. AUDITORY PERCEPTION OF BULGARIAN STIMULI BY RUSSIAN NATIVE SPEAKERS

Stenger I. (ira.stenger@mx.uni-saarland.de),
Avgustinova T. (avgustinova@coli.uni-saarland.de)
Saarland University, Saarbrücken, Germany

This study contributes to a better understanding of receptive multilingualism by determining similarities and differences in successful processing of written and spoken cognate words in an unknown but (closely) related language. We investigate two Slavic languages with regard to their mutual intelligibility. The current focus is on the recognition of isolated Bulgarian words by Russian native speakers in a cognate guessing task, considering both written and audio stimuli. The experimentally obtained intercomprehension scores show a generally high degree of intelligibility of Bulgarian cognates to Russian subjects, as well as processing difficulties in case of visual vs. auditory perception. In search of an explanation, we examine the linguistic factors that can contribute to various degrees of written and spoken word intelligibility. The intercomprehension scores obtained in the online word translation experiments are correlated with (i) the identical and mismatched correspondences on the orthographic and phonetic level, (ii) the word length of the stimuli, and (iii) the frequency of Russian cognates. Additionally we validate two measuring methods: the Levenshtein distance and the word adaptation surprisal as potential predictors of the word intelligibility in reading and oral intercomprehension.

Keywords: online experiments, receptive multilingualism, linguistic factors, Bulgarian, Russian

DOI: 10.28995/2075-7182-2020-19-684-695

ПОНИМАНИЕ БОЛГАРСКИХ СЛОВ-СТИМУЛОВ В ПИСЬМЕННОЙ И УСТНОЙ ФОРМАХ НОСИТЕЛЯМИ РУССКОГО ЯЗЫКА

Штенгер И. (ira.stenger@mx.uni-saarland.de),
Августинова Т. (avgustinova@coli.uni-saarland.de)
Университет земли Саар, Саарбрюккен, Германия

В настоящей статье представлен материал исследования в области рецептивного мультилингвизма, в центре которого находится явление сходств и различий в процессе успешного понимания слов-когнатов незнакомого, но (близко)родственного языка в письменной и устной формах. Мы рассматриваем два славянских языка с точки зрения их взаимопонятности. В данной работе мы анализируем понятность изолированных болгарских слов-стимулов для носителей русского языка при выполнении экспериментов по свободному переводу слов-когнатов в письменной и устной формах. Результаты проведенных экспериментов свидетельствуют о достаточно высоком уровне понятности болгарских слов для носителей русского языка, а также указывают на трудности в процессе понимания в зависимости от формы восприятия. С целью объяснения полученных результатов мы анализируем, какие из лингвистических факторов играют более важную роль в понимании болгарских слов при чтении и на слух. Полученные результаты онлайн-экспериментов коррелируются со следующими переменными: а) идентичные и неидентичные соответствия на орфографическом и фонетическом уровнях, б) длина слова-стимула, в) частотность русских когнатов. Дополнительно производится оценка двух методов измерения: расстояние Левенштейна и мера неожиданности слова, в качестве потенциальных параметров, объясняющих понятность слов при чтении и на слух.

Ключевые слова: онлайн-эксперименты, рецептивный мультилингвизм, лингвистические факторы, болгарский язык, русский язык

1. Background

Globalization and migration processes as well as the rapid development of new technologies in recent years pose new challenges in communication. In this regard, a special mode of language use known as *receptive multilingualism* [Braunmüller, Zeevaert 2001] or *intercomprehension* [Doyé 2005] is becoming more and more topical. It shows the robustness of human language competence by taking on various types of information—e.g. contextual, situational, socio-cognitive—and employing multisource compensatory and guessing strategies. This unconventional communication form relies both on *inherent* intelligibility, i.e. when speakers of one language can understand unknown but related languages because of structural linguistic similarities, and on *acquired* intelligibility, i.e. when they have learnt the respective language [Gooskens 2019]. An acquired *lingua receptiva* can apply to less related or unrelated languages, too [Muikku-Werner 2014], and this is also known as *mediated* receptive multilingualism [Branets et al. 2019] which eases understanding between typologically distant languages through the medium of a language closely related to the target.

In receptive multilingualism, the reading activity is expected to be essentially different from the listening one. While in listening the time available for input processing is limited, in reading one can jump back at will [Möller, Zeevaert 2015]. In foreign language teaching and learning, for example, the listening ability is generally considered more difficult than reading comprehension. This can certainly be true of receptive multilingualism, too. Examining the mutual intelligibility of some West

and South Slavic languages, [Golubović 2016] reports slightly lower intelligibility scores in the spoken word translation task than in the written one, as well as generally lower values for the spoken cloze test in comparison to the written¹ one.

Our key assumption is that the recognition of individual cognates² is a precondition for any further text understanding, be it via reading or listening. In (closely) related languages, the recognition of isolated words is certainly not to be equated with the reading or listening of actual texts. In addition to the identification of orthographic or phonetic similarities, further assumptions about the text—based on external characteristics, textual context, or morphological and syntactic analysis—can play a significant role in text exploration [Möller, Zeevaert 2010]. An additional fact to consider here is that a given context is only useful for inferring purposes if it has been understood, which is not always the case in an intercomprehension scenario. While the possibility for intercomprehension is strongly linked to the amount of common vocabulary in the respective languages, the respective cognates are neither identical nor transparent enough [Möller, Zeevaert 2015]. Therefore, it is justified to scrutinize cognate recognition in isolation, i.e. without context, and based on these findings to proceed with the exploration of entire texts.

For our online experiments, we use the INCOMSLAV platform³, concentrating on the spoken and written translation of cognates. We want to see whether the degree of similarity between the source Bulgarian words and their cognate Russian targets could predict the expected mutual intelligibility, and how this might be different in visual and auditory perception. After presenting the experimental setup, we introduce linguistic factors that may influence and explain the cross-lingual intelligibility of visual and auditory linguistic input. We then present the experimental results in both modalities and analyze the influence of the linguistic factors on human performance, before we draw some general conclusions.

2. Cognate guessing and free translation

After completing a background questionnaire, participants were asked to translate randomized 120⁴ written and spoken⁵ Bulgarian (BG) stimuli into their native Russian (RU) language in two series of 60 stimuli each. The items were taken from parallel lists consisting of internationalisms, Pan-Slavic vocabulary, and cognates

¹ Only participants who indicated that they could read Cyrillics were presented with written tasks in Bulgarian [Golubović 2016: 75].

² I.e. historically (etymologically) related word pairs that still bear the same meaning in both languages.

³ The website includes a large number of different online experiments in 11 Slavic languages (as well as in German and English) carried out as challenges in a linguistic game (<http://intercomprehension.coli.uni-saarland.de>).

⁴ 118 nouns and 2 numerals.

⁵ BG stimuli were read aloud by a female native speaker and recorded in a professional sound studio.

from Swadesh lists (for more details cf. [Fischer et al. 2015], [Stenger et al. 2019]). As we are interested in *inherent* intercomprehension, only people who speak RU natively and who do not know the stimulus language BG have been included in the analysis. In order to avoid any learning effects we consider here the results of the initial experiment in each modality.

The number of participants in the *written* word translation task is 40, aged between 18 and 71 years (i.e. average age 33) with 32 women and 8 men⁶. They saw the stimuli on their screen one at a time, and had 10 seconds to translate each word. The number of participants in the *spoken* word translation task is 29, aged between 16 and 48 years (i.e. average age 32)⁷ with 23 women, 5 men, and 1 not specified. They listened to the stimuli one by one (each word was played twice), and had to provide a written translation within 10 seconds. The time limit was chosen based on the experience from other intercomprehension experiments, and the results were automatically categorized as ‘correct’ or ‘wrong’ via pattern matching with pre-defined correct answers and acceptable alternatives. The responses were then manually checked for typographical errors in the final analysis.

3. Explaining variables in linguistic stimuli

3.1. Orthographic and sound correspondences

We distinguish between *identical* and *mismatched* correspondences and assume that identical correspondences positively affect intelligibility while that mismatched correspondences affect it negatively [Stenger 2019]. We have calculated identical and mismatched correspondences by means of the Levenshtein algorithm for 120 BG–RU cognate pairs on the orthographic and phonetic level. For example, the BG–RU cognate pair *автомобил* [əftomo'bił]—*автомобиль* [ɛftəme'bil'] ‘car’ has 9 identical orthographic correspondences and 1 mismatched orthographic correspondence and 4 identical and 5 mismatched sound correspondences.

3.2. Word length

Word length has been shown to influence the intelligibility of individual words. For example, [Kürschner et al. 2008] found that longer words are more easily recognized than shorter words in an intercomprehension listening scenario. We calculated word lengths in terms of the number of characters and sound segments. The longest word in BG is *автомобил* [əftomo'bił] ‘car’ and consists of 9 characters and 9 sound segments. The shortest BG word is *еж* [ɛʃ] ‘hedgehog’, consisting of 2 characters and 2 sound segments.

⁶ For more details of the written word intelligibility test see [Stenger 2019].

⁷ In total 30 participants took part at the experiments, one male participant noticed the knowledge of the BG and was excluded from the analysis.

3.3. Word frequency

Word frequency may also influence the correct understanding of cognates, since speakers are exposed more often to frequent words [Kürschner et al. 2008]. Word frequencies of RU cognates we use are based on frequency lists from the Russian national corpus [Lyashevskaya, Sharov 2009]. With regard to the stimuli, the most frequent RU cognate is *один* ‘one’ (2245.7 ipm⁸) and the least frequent one is *кельнер* ‘waiter’ (0.4 ipm). So, the initial hypothesis is that the more frequent the RU cognate, the easier it would be to understand the equivalent BG stimulus.

4. Predictors of mutual intelligibility

4.1. Levenshtein distance

Linguistic similarity is commonly measured on cognates (cf. [Gooskens 2019], [Stenger 2019], [Vanhove 2014]) by computing the *Levenshtein distance* (LD), i.e. taking into account—for two corresponding items—the minimum number of symbols that need to be inserted, deleted or substituted in order to transform a word in one language into the corresponding word in another language. Employing a modified Levenshtein algorithm [Levenshtein 1965], which disallows matching between a vowel and a consonant; we have calculated the orthographic and the phonetic⁹ distances between 120 BG–RU cognate pairs. This objective measure, we calculated automatically using the *incompy* tool of [Mosbach et al. 2019]. While in the basic form of the algorithm all string operations have the same cost, we use 0 for the cost of mapping a character/sound to itself, e.g. *a:a* or *[a]:[a]*, and 1 for the cost of aligning it to a character/sound of the same kind (vowel vs. consonant), e.g. *a:o* or *[a]:[o]*.¹⁰ To compensate for word length effects we normalize dividing the total sum of costs by the number of slots in the alignment. For example, the normalized LD of BG–RU cognate pair in the written condition *язык–язык* ‘tongue/language’ is $2/4 = 0.5$ or 50%, while in the spoken condition *[ɛˈzik]–[ɪˈzik]* ‘tongue/language’ it is $3/5 = 0.6$ or 60%. The assumption is that the higher the normalized LD, the more difficult it is to understand another language in an intercomprehension scenario.

⁸ instances per million words

⁹ For the transcription in IPA (International Phonetic Alphabet) we used the following online resources: <https://de.glosbe.com/>; https://en.wiktionary.org/wiki/Wiktionary:Main_Page; [Dellert, Jäger 2017] <http://www.northeastalex.org/> (18.11.2019). The missing IPA transcriptions for a few BG words were added according to Ternes, [Vladimirova-Buhtz 2010].

¹⁰ For orthographic distance we assign only the pair *e:ë* a substitution cost of 0.5. For phonetic distance we assign only the pair *[ɪ]:[ɣ]* a substitution cost of 0.5. The semi-vowel *[ɪ]* may correspond to a vowel and to a consonant.

4.2. Word adaptation surprisal

We take an information-theoretic approach [Shannon 1948] using the concept of *Word Adaptation surprisal* (WAS) to approximate the difficulty humans encounter when mapping one orthographic or phonetic system. Essentially, it reveals how predictable a given correspondence is in the respective pair of languages. Technically, WAS is calculated in bits according to the character or sound transformation probabilities, summing up values for character adaptation surprisal (CAS) or sound adaptation surprisal (SAS), respectively. For example, CAS is defined as in (1).

- (1) $CAS(L1 = c1|L2 = c2) = -\log_2 P(L1 = c1|L2 = c2)$
 L1—native language, c1—character of the native language
 L2—stimulus language, c2—character of the stimulus language

Since WAS between two words is computed by summing up the CAS and the SAS values of the contained characters and sounds in the aligned word pair, it strongly depends on the number of available word pairs (for more details see [Mosbach et al. 2019], [Stenger 2019]). Finally, we normalize the WAS based on the set of 120 BG–RU cognates. For example, the normalized WAS of the BG–RU cognate pair *език–язык* ‘tongue/language’ is 0.799 bits in the written condition, while in the spoken condition [eˈzik]–[ɹɹˈzik] ‘tongue/language’ it is 1.488 bits. The assumption is that the higher the normalized WAS value, the more difficult it is to comprehend the unknown language on the word level.

5. Results of intercomprehension experiments

The mean percentage of correctly translated items constitutes the intercomprehension score in a given modality (Table 1). The results show no great difference between the visual and the auditory perception of BG words by RU native speakers: the RU participants understand a slightly larger number of written BG words (71.33%) than spoken ones (68.42%).

Table 1. The results of written and spoken word translation tasks

Native language	Bulgarian word translation task	
	written	spoken
Russian	71.33%	68.42%

The next question is whether RU subjects are more successful in cognate guessing when they perceive the BG stimuli as a visual or as an auditory input. Fig. 1 presents quantitative data on the 62 successfully guessed spoken cognates (left side), the 56 successfully guessed written cognates (right side) and the two cognate pairs yielding identical scores (middle). The vertical axis expresses the difference as the percentage of correct translations.

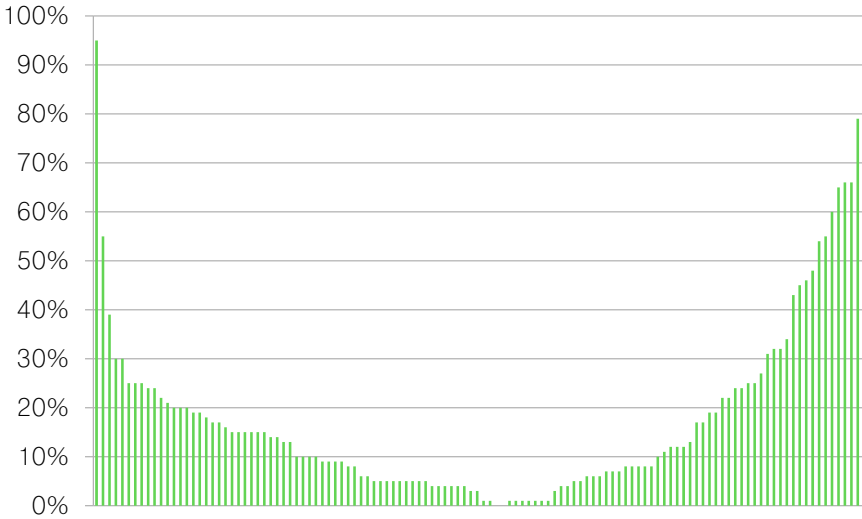


Fig. 1. Number of successfully guessed cognates while listening (left) vs. reading (right)

In **Table 2** we split the differences between the intercomprehension scores into three categories: high (more than 60%), medium (between 60% and 30%), and low (less than 30%). For 6 BG words in the high-percentage group, reading gave much better results than listening, e.g., the BG stimulus *еж* [ɛʃ] was correctly understood as *ёж* [jɵʂ] ‘hedgehog’ to 90% in the written translation task and only to 6% in the spoken translation task. For 10 BG words at the middle group, reading caused fewer difficulties than listening, e.g., the BG stimulus *лен* [lɛn] was correctly translated in the written task as *лѐн* [lʲɵn] ‘flax’ to 60% and in the spoken task only to 5%. These examples show that for RU subjects the BG–RU mismatched orthographic correspondence *e:ë* is more transparent than the respective mismatched sound correspondences. There were a total of 40 BG stimuli in the low-percentage group, where RU subjects performed better in the written than in the spoken task. Only 1 BG stimulus *ръка* [rə'ka] was much better understood by RU subjects in auditory perception as *рука* [rʊ'ka] ‘hand, arm’ (100% intelligibly vs. 5%).

It is remarkable that the most frequent incorrect answer in the written word translation task was the RU word *река* ‘river’ which can be considered an orthographic neighbor to the BG stimulus and to the RU cognate word¹¹. Furthermore, 4 spoken BG stimuli were in the middle group and 57 spoken BG words fell into the low-percentage group of differences with regard to intelligibility. To conclude, the slightly better

¹¹ Previous research shows that the neighborhood density may play a significant role in spoken and written word recognition without context [Kürschner et al. 2008], [Stenger 2019]. For example, Bulgarian and Serbian written intelligibility to Russian native speakers shows that the higher the neighborhood density, the lower is the number of successful translations, although this is not the case for Ukrainian, Belarusian, and Macedonian stimuli when presented to Russian readers [Stenger 2019].

performance in the written task as opposed to the spoken one (71.33% vs. 68.42%) is due to the higher number of correctly translated words in the first two groups.

Table 2. Grouping cognate pairs according to differences in correct translations

% difference	visual perception	auditory perception
60–100	N = 6 BG <i>еж</i> [ɛʃ] RU <i>ёж</i> [jɔʂ] ‘hedgehog’	N = 1 BG <i>ръка</i> [rə'ka] RU <i>рука</i> [rʊ'ka] ‘hand, arm’
30–60	N = 10 BG <i>лен</i> [lɛn] RU <i>лён</i> [lʲɛn] ‘flax’	N = 4 BG <i>кръв</i> [krɔf] RU <i>кровь</i> [krɔfʲ] ‘blood’
1–30	N = 40 BG <i>дъб</i> [dɔp] RU <i>дуб</i> [dʊp] ‘oak’	N = 57 BG <i>език</i> [ɛ'zik] RU <i>язык</i> [ɹɪ'zik] ‘tongue’

6. Correlation of intercomprehension scores with explaining variables

We correlated the experimentally obtained intercomprehension scores with the explaining variables introduced in [Section 3](#)—the identical and mismatched correspondences (ic vs. mc), the word length (wl) and the word frequency (wf)—in order to determine their impact on word-level written and spoken intelligibility. An overview of the statistical results (Pearson’s r and p -value) is provided in [Table 3](#).

Table 3. The correlations between intercomprehension scores and explaining variables

Translation task	Explaining variables			
	ic	mc	wl	wf
visual perception	$r = 0.3480$ $p < 0.0001$	$r = -0.3830$ $p = 1.61e-05$	$r = 0.3210$ $p < 0.0005$	$r = -0.0160$ $p = 0.09$
auditory perception	$r = 0.2359$ $p < 0.01$	$r = 0.0550$ $p = 0.55$	$r = 0.3620$ $p < 0.00005$	$r = 0.1610$ $p = 0.08$

In visual perception, intercomprehension scores correlate significantly with the identical and mismatched orthographic correspondences, as well as with the word length. The correlation between the written intelligibility and word frequency is not positive as assumed, but insignificant.

In auditory perception, only two variables have an impact on word recognition: identical sound correspondences and the word length. The correlation between the intercomprehension scores and the mismatched sound correspondences is not negative as assumed, but yet insignificant. It might be the case that qualitative characteristics of the mismatched sound correspondences are more decisive in spoken word recognition than their number. We also found a positive but low and not significant correlation between

the intercomprehension scores and the word frequency in the spoken modality. This was already shown by [Kürschner et al. 2008] and could be replicated in our experiment.

7. Correlation of intercomprehension scores with predictors of mutual intelligibility

7.1. Levenshtein distance

To investigate how word intelligibility depends on orthographic and phonetic similarities, the intercomprehension scores were correlated with the respective normalized LDs (Fig. 2).

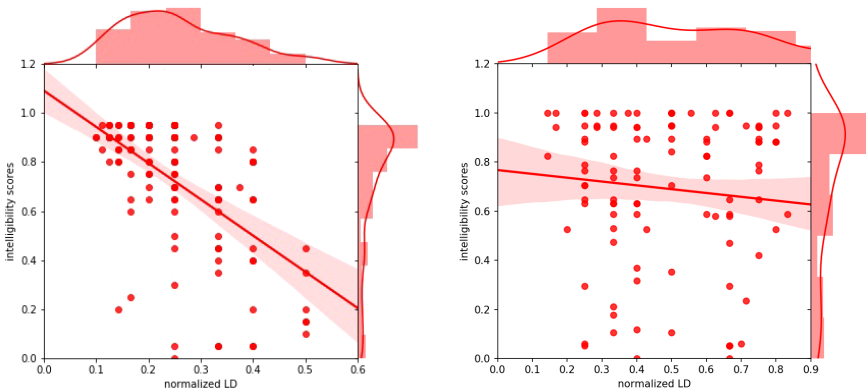


Fig. 2. Normalized LD as predictor of written (left) and spoken (right) intelligibility

There is a negative and significant correlation of -0.566 ($p = 1.47e-11$) in the written modality (Fig. 2 left), which means that on the word level the written intelligibility can be predicted well from the normalized LD. This is in contrast to the spoken modality (Fig. 2 right), where we can see that the spoken intelligibility is not well predictable from the normalized LD (Pearson's $r = -0.1185$, $p = 0.197$). This can be explained by the fact that the normalized LD depends on the number of mismatched sound correspondences as well as on their weights and thus increases with more mismatched correspondences. The RU listeners seem to have a high tolerance towards the phonetic distance and are able to process sensitive sound differences quite well. To model the perception of spoken BG stimuli by RU listeners more successfully, we decided to integrate into the LD algorithm sensitive sound distances with a substitution cost of 0.5 between non-palatalized and palatalized consonants, as well as between the following similar correspondences: [ɛ]:[e], [i]:[i], [ɔ]:[o], [u]:[ʊ], [ʃ]:[ʒ], [z]:[ʒ], [ʧ]:[ʧ]. As a result, we found a significant negative correlation between the intercomprehension scores and the phonetic distance: (Pearson's $r = -0.2019$, $p < 0.05$).

7.2. Word adaptation surprisal

The correlations between the normalized WAS and the intercomprehension scores of both translation tasks (Fig. 3) show that written and spoken intelligibility cannot be predicted reliably from the normalized WAS on the word level. Both correlations are very low and not significant (written: Pearson's $r = -0.1350$, $p = 0.14$, spoken: Pearson's $r = -0.1492$, $p = 0.1$).

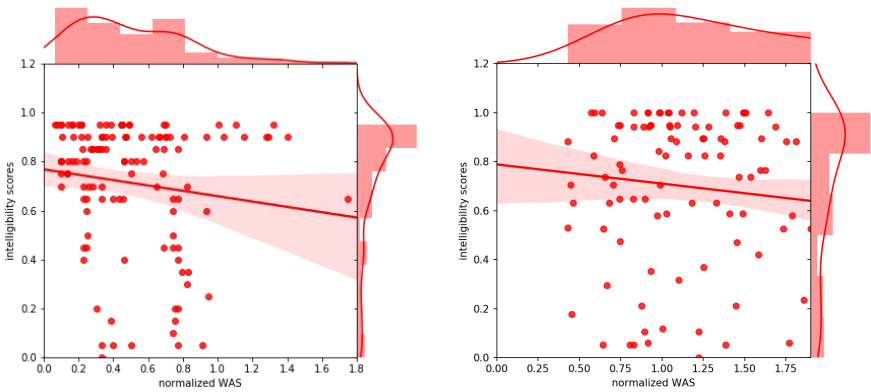


Fig. 3. Normalized WAS as predictor of written (left) and spoken (right) intelligibility

The low and insignificant correlations can be explained by the fact that identical orthographic and sound correspondences may still have a small adaptation surprisal value, which increases the whole WAS value. Hence, we decided to modify our normalized WAS calculation in such a way that all identical orthographic and sound correspondences were measured with 0 bits. The calculated CAS values for mismatched orthographic correspondences and the SAS values for mismatched sound correspondences remained unchanged in the modified calculation. After the modification of the WAS method, we found a negative and significant correlation between the modified normalized WAS and the intercomprehension scores in both conditions (written: Pearson's $r = -0.210$, $p < 0.05$, spoken: Pearson's $r = -0.181$, $p < 0.05$).

8. Conclusions

In a series of online experiments, we investigated how well RU natives could understand written and spoken BG words in an intercomprehension scenario. Overall, RU subjects are good at recognizing both written and spoken BG stimuli: 71.33% and 68.42% respectively. The obtained intercomprehension scores in the written task are only slightly higher than in the spoken one. We considered which linguistic factors can determine the mutual intelligibility and to what degree. First, we examined the individual variables that are known to affect intelligibility and found that identical correspondences and word length of the BG stimuli have a positive and significant

impact on human performance in both perception modes. However, while the statistical analysis reveals that mismatched orthographic correspondences could be significant in negatively affecting written intelligibility, the number of mismatched sound correspondences cannot be an explanatory variable for the difficulties in spoken word recognition. The frequency of cognates is not a reliable predictor in listening as well as in reading and seems not to have any effect on word recognition. Second, we validated two predictors of written and spoken intelligibility: the normalized LD and normalized WAS. While the written intelligibility was well predictable from the orthographic distance, the explanatory potential of the phonetic distance was very small and not significant. The incorporation of sensitive sound distances into the LD algorithm showed a significant influence of the phonetic distance in spoken word comprehension, but still to a low degree. After the modification of the WAS method, we found that predictability of both orthographic and sound correspondences have slight effects on subject's ability to understand an unknown but related language. The knowledge gained from this study will serve as a useful point for further research in understanding of the specific linguistic factors that contribute to (un)successful intercomprehension.

9. Acknowledgements

We wish to thank Marius Mosbach for his support in the implementation of our data in the *incompy* tool. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102.

References

1. Branets A., Bahtina D., Verschik A. (2019), Mediated receptive multilingualism: Estonian- Russian-Ukrainian case study, in *Linguistic Approaches to Bilingualism*, available at: <https://doi.org/10.1075/lab.17079.ver> (18.11.2019).
2. Braunmüller K., Zeevaert L. (2001), Semicommunication, receptive multilingualism and related phenomena. A bibliographical overview, [Semikommunikation, rezeptive Mehrsprachigkeit und verwandte Phänomene. Eine bibliographische Bestandsaufnahme], Working papers in multilingualism [Arbeiten zur Mehrsprachigkeit], Series B, No. 19, University Hamburg [Universität Hamburg].
3. Dellert J., Jäger G. NorthEuraLex (version 0.9). available at: <http://www.northeuralex.org/> (18.11.2019).
4. Doyé P. (2005), Intercomprehension. Guide for the development of language education policies in Europe: from linguistic diversity to plurilingual education, Reference study, Strasbourg, DG IV, Council of Europe.
5. Fischer A., Jágrová K., Stenger I., Avgustinova T., Klakow D., Marti R. (2015), An Orthography Transformation Experiment with Czech-Polish and Bulgarian-Russian Parallel Word Sets, in Sharp B., Lubaszewski W., Delmonte R. (eds.), *Natural Language Processing and Cognitive Science 2015 Proceedings*, Libreria Editrice Cafoscarina, Venezia, pp. 115–126.
6. Golubović J. (2016), Mutual intelligibility in the Slavic language area, PhD thesis, University of Groningen (Netherlands).

7. *Gooskens C.* (2019), Receptive multilingualism, in Montanari S., Quay S. (eds.), *Multidisciplinary perspectives on multilingualism: The fundamentals*, De Gruyter Mouton, Berlin, pp. 149–174, http://www.let.rug.nl/gooskens/pdf/publ_multidisciplinary_perspectives_2019.pdf.
8. *Kürschner S., van Bezooijen R., Gooskens C.* (2008), Linguistic determinants of the intelligibility of Swedish words among Danes, *International Journal of Humanities and Arts Computing* 2(1/2), pp. 83–100.
9. *Levenshtein V. I.* (1965), Binary codes capable of correcting deletions, insertions, and reversals [Dvoičnye kody s ispravleniem udaleniy, vstavok i zamen simvolov], *Doklady of the Soviet Academy [Doklady Akademii Nauk SSSR]*, 1965, Vol. 163, No. 4, pp. 845–848.
10. *Lyashevskaya O. N., Sharov S. A.* (2009), Frequency dictionary of the contemporary Russian language [Chastotnyy slovar' sovremennogo russkogo yazyka], Azbukovnik, Moskva.
11. *Muikku-Werner P.* (2014), Co-text and receptive multilingualism Finnish students comprehending Estonian, *Journal of Estonian and Finno-Ugric Linguistics* 5(3), pp. 99–103.
12. *Möller R., Zeevaert L.* (2010), „There I spontaneously think of a blackboard“—For word recognition in related Germanic languages [„Da denke ich spontan an Tafel“—Zur Worterkennung in verwandten germanischen Sprachen], *Journal for foreign language research* 21(2) [Zeitschrift für Fremdsprachenforschung 21(2)], pp. 217–248.
13. *Möller R., Zeevaert L.* (2015), Investigating word recognition in intercomprehension: Methods and findings, *Linguistics* 2015 53(2), pp. 313–352.
14. *Mosbach M., Stenger I., Avgustinova T., Klakow D.* (2019), *incom.py*—A Toolbox for Calculating Linguistic Distances and Asymmetries between Related Languages, in Angelova G., Mitkov R., Nikolova I., Temnikova I. (eds.), *Proceedings of Recent Advances in Natural Languages Processing, RANLP 2019*, pp. 811–819.
15. *Shannon C. E.* (1948), A mathematical theory of communication, *Bell System Technical Journal* 27 (379–423), pp. 623–656.
16. *Stenger I.* (2019), On the role of orthography in Slavic intercomprehension with a special focus on the Cyrillic script [Zur Rolle der Orthographie in der slavischen Interkomprehension mit besonderem Fokus auf die kyrillische Schrift], PhD thesis, Universaar, Saarbrücken.
17. *Stenger I., Avgustinova T., Belousov K., Baranov D., Erofeeva E.* (2019), Interaction of linguistic and socio-cognitive factors in receptive multilingualism, in Selegej V. et al. (eds.), *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference 'Dialogue' (2019)*, Issue 18, Supplementary volume, pp. 211–226.
18. *Ternes E., Vladimirova-Buhtz T.* (2010), Bulgarian, in *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press, Cambridge, pp. 55–57.
19. *Vanhove J.* (2014), Receptive multilingualism across the lifespan. Cognitive and linguistic factors in cognate guessing, PhD thesis, University of Fribourg (Switzerland).

AN EMPIRICAL INVESTIGATION OF LANGUAGE MODEL BASED REVERSE TURING TEST AS A TOOL FOR KNOWLEDGE AND SKILLS ASSESSMENT

Tarasov D. (dtarasov3@gmail.com),
Matveeva T., Galiullina N.

Meanotek, Kazan, Russia

Automating assessment of person's skills is an important area of study in artificial intelligence and natural language processing. In this work we conduct empirical study of a recently proposed Reverse Turing Test for Knowledge Assessment approach—a completely automated domain agnostic method of knowledge assessment that can operate completely without human assessor involvement. Our study involved 53 participants and three different knowledge domains. We conclude that this method can reliably differentiate between expertise levels and therefore can be a compelling alternative to human grading and multiple-choice tests in many domains.

Keywords: Knowledge assesment, Reverse Turing test, language model

DOI: 10.28995/2075-7182-2020-19-696-707

ЭМПИРИЧЕСКОЕ ИССЛЕДОВАНИЕ ОБРАТНОГО ТЕСТА ТЬЮРИНГА НА ОСНОВЕ ЯЗЫКОВОЙ МОДЕЛИ КАК ИНСТРУМЕНТА ОЦЕНКИ ЗНАНИЙ И НАВЫКОВ

Тарасов Д. (dtarasov3@gmail.com),
Матвеева Т., Галиуллина Н.

ООО «Меанотек», Казань, Россия

Автоматическая оценка знаний и навыков человека является важной задачей искусственного интеллекта и обработки естественного языка. В этой работе мы описываем эмпирическое исследование недавно предложенного метода обратного теста Тьюринга для оценки

знаний—полностью автоматизированного метода оценки знаний, независимого от предметной области, который может использоваться без участия человека. В нашем исследовании приняли участие 53 участника и три разные области знаний. Мы пришли к выводу, что этот метод может надежно дифференцировать уровни знаний и, следовательно, может стать альтернативой человеческим оценкам и тестам с выбором вариантов ответа.

Ключевые слова: оценка знаний, тест обратного Тьюринга, языковая модель

1. Introduction

Testing student's knowledge is a cornerstone of every educational system, as no educational process can function efficiently without a way to assess that students possess certain knowledge and are capable to perform reasoning based on that knowledge.

Today, traditional ways to test knowledge, such as oral exam with a teacher, or written test, with manual verification of the results, can no longer satisfy practical needs for objective standardized and automated process.

Furthermore, the quality of human assessment of knowledge can only be as good, as knowledge of a given human assessor. As the population of the world ages, old experts leave organizations, which often leads to a catastrophic loss of critical operational knowledge [DeLong et al, 2004]. As a consequence, organizations lose not only operational efficiency, but also the ability to accurately assess knowledge of new employees and students, which opens the door to further knowledge loss.

These and other factors create high demand for computerized knowledge assessment tools and many methods have been developed to automate this process, none of them being completely satisfactory, despite receiving wide adoption. In fact, tools such as multiple choice tests, constructed answer tests, and automatic essay grading are often considered as harmful to learning process [Ryan and Weinstein, 2009]; [Groothuis, 2018], as they favor rote learning and can undermine both student engagement and best teaching practices. Better knowledge assessment tools are needed to solve these problems.

The Reverse Turing Test (RTT) is a variation of the Turing Test, in which the roles of computer and person are reversed [Baird et al, 2003]. In a typical implementation of the Reverse Turing Test, a computer algorithm should determine if it is dealing with a person, or with another computer algorithm. A widely known version of Reverse Turing Test is CAPTCHA—a test whose task is to determine whether a website visitor is a bot or a person. CAPTCHA presents the agent being evaluated with a task specially selected in such a way that it is easy for a person to solve, but is practically unsolvable for the algorithm. Many variations of this CAPTCHA test [Kochanski et al, 2002]; [Lopresti, 2005]; [McInerney et al, 2019] are known.

In the context of medical diagnostics, Reverse Turing Test have been proposed for diagnosing cognitive disorders such as autism spectrum disorders (ASD) and Alzheimer's disease [Montenegro et al, 2017]; [d'Arc et al, 2018]. In particular, the

papers argue that those suffering from ASD lack “the theory of mind”—the ability to predict the internal states of another person. The authors therefore propose to detect such disorders through a competitive dyad game against a computer or human opponent. Persons suffering from cognitive impairment are not able to use knowledge about the nature of the opponent (human or computer) to adapt their strategy.

Quite recently, another interesting variation of Reverse Turing Test was proposed that can be theoretically viewed as nearly universal knowledge assessment tool [Tarasov, 2019]; [2020]. In that specific version of the Reverse Turing Test a generative domain model is used to produce an object that imitates the result of human intellectual activity and knowledge is assessed based on assessing the difference between a person’s interaction with a real and algorithmically generated object.

The key idea here is that algorithmically generated object will contain certain flaws due to limitation of generative model used, and these flaws should be apparent to a human who possesses specialized knowledge to identify them. Such approach circumvents the biggest barrier to automatic knowledge assessment—the lack of human expert level AI, making possible for lesser intelligence to assess knowledge of the higher one.

One practical implementation of such test is language model-based reverse Turing test (LM-RTT). In this implementation a high-capacity neural language model is trained on a set of texts containing specialized knowledge, such as scientific publications on a given subject. An examinee is then given a mix of generated and real text fragments from with the task of distinguishing real and fake texts. The hypothesis is that since modern language models are known to be capable of generating text, that can not be reliably identified as fake by humans using only grammatical cues and common sense [Graefe et al, 2018], passing such a test will require examinee to catch world-modeling failures in their respective fields of study. The task therefore, will require construction of the mental model of certain situation and checking it for consistency, and be impossible to complete with just rote learning.

Can such a procedure really be a reliable indicator of student’s knowledge? A positive answer to that question can lead to development of a completely new knowledge assessment method, possessing unique properties of being fully automated and capable to assess deep knowledge and reasoning abilities.

To answer this question, we conducted a number of experiments where we compare LM-RTT scores of established domain experts in certain subject area versus non-experts, or students. Our results indicate that experts score significantly higher, and that it is possible to distinguish member of expert from non-expert group with high accuracy.

2. Related work

A number of attempts to automate knowledge assessment were made. A multiple-choice test (MTT) and its numerous variations received a wide adoption after early computer systems allowed for automated scoring of such tests. They, however, have a well-known disadvantages. MTT and its variants favor rote learning, can not assess higher-order reasoning skills and still need human to develop it. They are also prone to ambiguous interpretation problem [Ryan et al, 1998]; [Roediger, 2015].

There are many known approaches to the computer generation of new MTTs using the knowledge base (ontology) [Papasalouros et al, 2008], as well as more modern solutions allowing to rephrase key sentences of a text into a question using semantic analysis [Kantor et al, 2018] or deep neural networks [Subramanian et al, 2017] and the generation of wrong answers. However, automatically generated MTTs are prone to errors, usually require human supervision and even less suited to assess higher-order reasoning and thinking capacity than human-authored tests.

Another branch of research aims to apply natural language understanding for verification of free-form answers. Automatic essay grading [Dong et al, 2017], theme adherence check [Tikhomirov et al, 2019], and plagiarism detection [Zubarev et al, 2019] are examples of such attempts. However reported performance even for relatively simple tasks like topic adherence is low, and checking higher-order reasoning in free-form answers outside of a few very specialized cases remains currently out of reach for present day technology. In fact, it can be argued that these forms of knowledge assessment require AI to possess human or even super-human level of general intelligence, since assessing knowledge of an agent can only be done by an agent that possesses the same or better level of understanding of the subject.

3. Algorithms and Methods

3.1. Test groups

The lack of a gold standard knowledge assessing method prevents us from directly comparing gold standard student grades with RTT grades. Lack of correlation between RTT and human grades or multiple-choice test scores can result from both invalidity of LM-RTT or failure of commonly used methods to capture real knowledge score. We opt therefore for a more objective approach. In our study we compare the score of a group of students or non-experts with the score of a group of experts where experts are selected on the basis of a) having real working experience in the field of interest for at least 2 years b) being active in the field of interest at the moment of the study. A good method of knowledge assessment should be able to differentiate between students and experts.

All participants were recruited on condition of anonymity and gave permission to publish aggregate results statistics. However, we were not able to obtain permissions from corresponding institutions to mention these institutions' names, therefore institutions are described here in generic terms.

Due to the preliminary nature of this study, the total number of participants was relatively small. We believe that this is justified because preliminary evidence of method validity is needed to justify expensive and labour-intensive larger studies.

3.1.1. Computer science domain

The topics of the test were "HTML and OSI model". The main group consisted of 28 students (last year of study) of an information technology college. The comparison group consisted of 5 software developers with 2 years of work experience.

3.1.2. Biomedical domain

For this domain we compare scores of 3 groups. Main group consisted of 7 students of department of biochemistry at a major university. Comparison group #1 consisted from 5 lecturers in the same department, including two PhDs. Comparison group #2 (non-biomedical experts) consisted of 5 software developers (the same as previously). The aim for having comparison group #2 was to assess whenever RTT measures specialized knowledge or just the level of general intelligence or ability to identify machine-generated text by looking for specific flaws. The topic of the test was “Proteins secondary and tertiary structure, structure of collagen and insulin”.

3.1.3. Food safety domain

Unlike two previous groups this test measured the knowledge of internal document (Restaurants rules for personnel). Main group consisted of 3 new hires, comparison group #2 consisted of 3 managers with 4 years+ experience, comparison group #3 consisted of 2 software developers

3.2. Language models and test construction

3.2.1. Computer science domain

16 paragraphs were sampled from English Wikipedia pages on HTML and OSI model. Paragraphs were required to contain keywords “HTML” or “OSI” in first 10 words. After that 16 paragraphs were generated with GPT-2 large model [Radford et al, 2019] using top k random sampling with k = 2, using first 10 words as context for generation (to make model follow desired topic). All texts were then translated into Russian using Google Translate API. Such approach makes it impossible for students to find real paragraphs by searching Internet.

A complete test, containing with 32 text fragments and the task was to distinguish between real and generated fragments on the basis of their logical consistency and factual correctness.

3.2.2. Biomedical domain

For biomedical domain, custom language model was trained on corpus of freely available pubmed abstracts (https://www.nlm.nih.gov/databases/download/pubmed_medline.html), data from 2019 baseline were used. We used character-level LSTM-based model [Hochreiter and Schmidhuber, 1997] with 7 LSTM layers, with 3,192 LSTM units in first two layers and 2,500 units in remaining layers. The model was trained for 1 month using two 2080Ti GPUs, achieving 0.96 BPC on test set.

For construction of the test, 200 abstracts were found with pubmed search API using keywords “collagen structure” and “proteins structure and function” out of which 14 abstract where sampled randomly and 14 abstracts were generated with language model using nucleus sampling method and corresponding article title as a context. We then used first three sentences from each abstract as a text fragment for construction of complete test (total 28 text fragments). All texts were translated with into Russian using Google Translate API.

3.2.3. Food safety domain

Custom language model was pre-trained on complete Russian Wikipedia dump (8Gb of text). We used character-level LSTM-based model with 8 LSTM layers, with 3,192 LSTM units in first two layers and 2,500 units in remaining layers. The model was trained for 2 weeks using two Tesla V100 16GB GPUs, achieving 1.04 BPC on test set. The model was then fine tuned on internal document, that described the food safety rules to which restaurant personnel must obey. Total document size was 260KB. We then randomly selected 14 paragraphs from this document and sampled corresponding 14 paragraphs from the model, using first 10 words of real paragraphs, as context.

3.3. Analysis of results

Statistical analysis was conducted by means of exact binomial test. The null hypothesis was that probability of giving a correct answer is the same in all groups.

4. Results and discussion

4.1. Computer science domain

For students group, average number of correct answers was 16.1 (50.3%), maximum 23 correct answers, minimum—11 correct answers. For experts group, average number of correct answers was 21.2 (66.25%), maximum was 24, minimum—18. Threshold of 21 correct answers allowed to distinguish student from expert with 98% accuracy with one false positive (student identified as expert) and one false negative (expert identified as student). The differences between means in two groups were found to be statistically significant at $p=0.95$ level using exact binomial test.

No correlation was found between RTT scores and student's annual grades. In fact, student with highest mark RTT (23) had the smallest annual grade (3). The teacher explained his grade was for low attendance and poor behavior. This leaves out 3 possible explanations:

- Administered RTT test was not sensitive enough to distinguish between students knowledge
- Test scores reflected higher general intelligence in expert group, not specialized knowledge
- Human annual grades were subjective, not reflecting actual knowledge

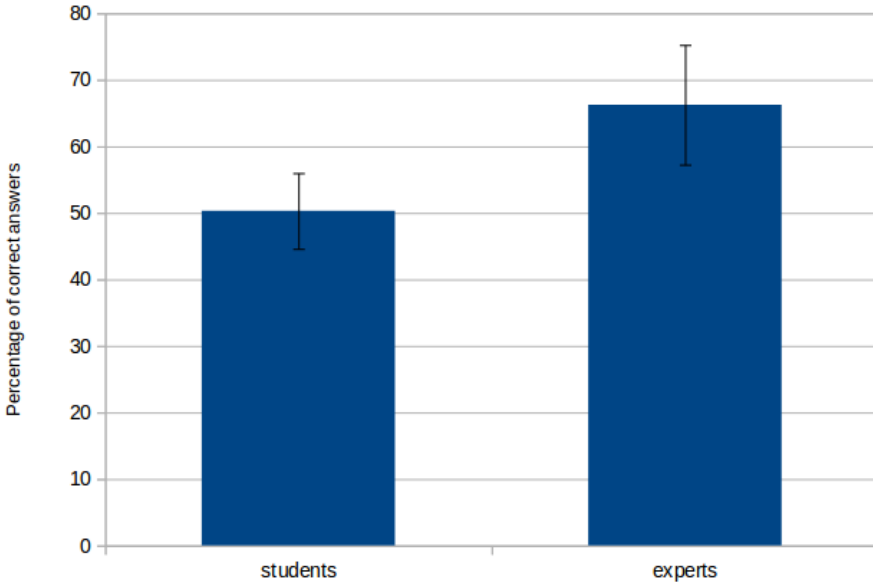


Figure 1. Percentage of correct answers in students and experts groups (Computer science domain). Error bars indicate 95% binomial confidence interval

4.2. Biomedical domain

Average score was 52% for student group, 67.5% experts group and 56.2% in no biomedical expertise group. The difference was statistically significant at $p=0.95$ level between student and expert groups and between no biomedical expertise and expert group, using exact binomial test. These results suggest that specialized knowledge gives advantage to biomedical experts over both students and software developers groups, even though software developers group was previously exposed to LM-RTT test on different topic and generally familiar with artifacts of neural text generation.

Another interesting point is that software developers average score on in-domain test, described above was significantly higher than on biomedical test. Taken together these results indicate that LM-RTT score is improved when examinee possesses specialized knowledge. These results also indicate that humans can't easily adapt to solve such tests without having necessary knowledge.

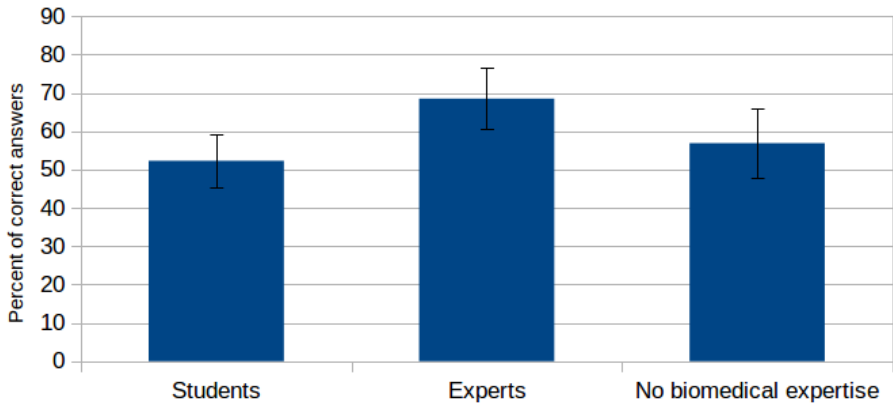


Figure 2. Percentage of correct answers in students and experts groups and group without biomedical expertise (Biomedical domain). Error bars indicate 95% binomial confidence interval

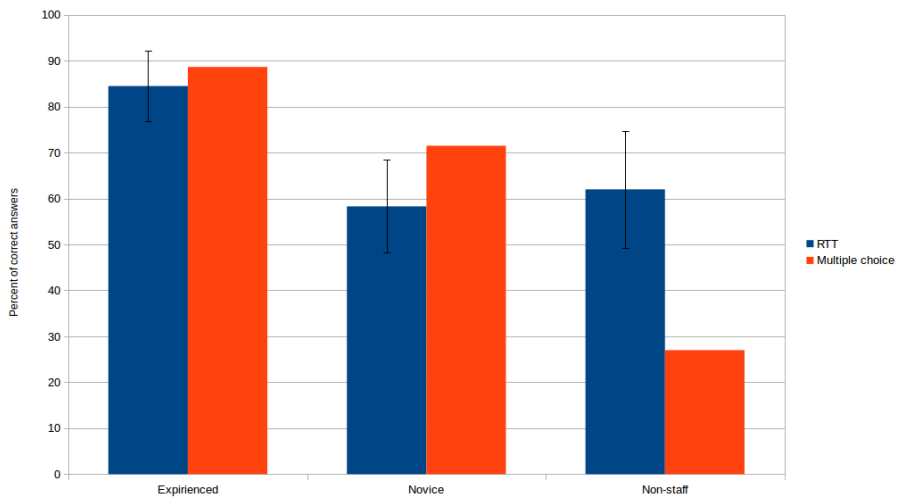


Figure 3. Percentage of correct answers in experienced, novice staff groups and non-staff group (Food safety domain) for RTT score and handcrafted multiple-choice test score (regularly used by restaurant to assess personnel knowledge). Error bars indicate 95% binomial confidence interval

4.3. Food safety domain

For experienced personnel group, average number of correct answers was 23.3 (84.5%), maximum 24 correct answers, minimum—23 correct answers. For novice group, average number of correct answers was 16 (57%), maximum was 18, minimum—14. The differences between means in experienced and novice groups were found to be statistically significant at $p=0.95$ level using exact binomial test. There were no significant difference between novice and non-staff groups.

The unexpected finding here was that while standard multiple-choice test, regularly used by restaurant to assess personnel knowledge, demonstrated clear difference between novice and non-staff groups, there were no such differences using RTT. At the same time, both tests are showing difference between experienced and novice. Restaurants rules require all new staff to pass multiple choice test and thus people repeatedly attempt to pass it and test content is not being changed between attempts, thus people in novice group could learn correct answers but their actual knowledge is low. In contrast, experienced staff group possesses real knowledge about operational procedures.

4.4. Qualitative analysis

Table 1 shows some example of generated fragments in two domains and shows several types of typical errors.

Table 1. Examples of generated fragments

Domain	Text	Error type
Computer science	HTML markup consists of several key components, commonly known as the head element, the body element, and the script element. Body element is often referred to as the document root (or document head), as that is where HTML documents are created. The body element contains information about the document’s contents, such as the document title and document meta data, which are not normally presented elsewhere on the page.	World modelling failure: HTML body is described as a place where HTML documents are created. Factual error: document metadata are located in document head, not body.
	HTML documents imply a structure of nested HTML elements. This is called a tree. If each node contains an attribute and some value, this might look like this: <pre><html> <head> <title>hibernate generator example</title> </head> <body> <h1>test</h1> </body> </html></pre> A node in this tree has an attribute named “title”, and a value of “test”.	Logical consistency error: element “title” does not have value of “test” in this example

Domain	Text	Error type
Biomedical	The aim of this study was to investigate the effect of BMP-2 on osteoclast formation and osteoclast formation in vitro.	Neural text degeneration (repeated phrase “osteoclast formation and osteoclast formation in vitro”)
	Collagen is the most abundant protein in the blood serum of patients with chronic renal diseases.	Factual error—collagen is not most abundant protein in blood serum, serum albumin is.
	The aim of this study is to identify, in terms of the collagen content, the proteoglycan content of the collagen fibrils in the developing human and rabbit condylar cartilage. The collagen fibrils were isolated from bovine cartilage by differential centrifugation.	Logical inconsistency. First sentence speaks about human and rabbit subjects, while second tells that collagen fibrils were isolated from bovine cartilage
	The aim of this study was to evaluate the effect of collagen type I on the production of collagen in rat skin by measuring the changes in the cell content of the serum proteins and the activity of the enzyme collagenase. Twenty-four hours after the induction of chronic progressive renal failure, the collagen content of the renal cortex was measured by radioimmunoassay. The results showed that the protein content of the collagen fibrils was significantly increased in the regenerated renal tissue.	World model error: collagen production is rat skin can not be measured by changes in «cell content of the serum proteins». Logical inconsistency (first part talks about collagen in skin, and then about collagen in renal cortex). Described overall sequence of events is biologically implausible. Grammar and overall structure, however are realistic.

We found a mix of surface mistakes, that stem from neural text degeneration, logical inconsistencies, factual errors and complex world modelling errors. Clearly, there is a room for improving test quality by using better language models and taking measures to prevent samples with degenerated text from appearing in the test.

5. Conclusions

We have established that:

1. Language model based reverse Turing test can reliably distinguish experts from non-experts in all 3 domains studied in this paper.
2. We found that specialized professional knowledge result in higher LM-RTT scores then using grammatical and logical consistency cues alone.
3. We repeatedly observed certain discrepancies between LM-RTT and traditional knowledge assessment methods, such as human grading and multiple choice test. We can't confidently establish reason for these discrepancies.
4. No significant adaptation was observed due to repeated exposure to LM-RTT testing format.

Overall LM-RTT seems to be a promising method for knowledge assessment. Undoubtedly, due to novelty of the method and controversial nature of RTT hypothesis, a lot of large studies in different knowledge domains will be needed to establish

method validity with enough certainty. However, such studies are difficult and costly to conduct as they need to involve a lot of real human subjects and some prior evidence is required to justify these expenses, and our findings justify further investigations of method capabilities using larger groups.

We believe that our results are important, because, if confirmed by further studies, they can lead to development of relatively cheap but powerful automated method to assess depth of person's knowledge in many domains.

References

1. *Baird, H. S., Coates, A. L., & Fateman, R. J.* (2003). Pessimism: a reverse turing test. *International Journal on Document Analysis and Recognition*, 5(2–3), 158–163.
2. *d'Arc, Baudouin Forgeot, Devaine Marie, and Jean Daunizeau.* "A reverse Turing-test for predicting social deficits in people with Autism." *bioRxiv* (2018): 414540.
3. *DeLong, David W., and J. Storey.* *Lost knowledge: Confronting the threat of an aging workforce.* Oxford University Press, 2004.
4. *Dong, Fei, Yue Zhang, and Jie Yang.* "Attention-based recurrent convolutional neural network for automatic essay scoring." *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017).* 2017.
5. *Graefe, A., Haim, M., Haarmann, B., & Brosius, H. B.* (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, 19(5), 595–610.
6. *Groothuis, Douglas.* "How Multiple-Choice Tests and Machine-Graded Essays Undermine Learning." *Academic Questions* 31.1 (2018): 70–76.
7. *Hochreiter, Sepp, and Jürgen Schmidhuber.* "Long short-term memory." *Neural computation* 9.8 (1997): 1735–1780.
8. *Kantor, Arthur, Jan Kleindienst, and Martin Schmid.* "Automatic question generation from natural text." U.S. Patent No. 9,904,675. 27 Feb. 2018
9. *Kochanski, Greg, Daniel Lopresti, and Chilin Shih.* "A reverse turing test using speech." *Seventh International Conference on Spoken Language Processing.* 2002.
10. *Lopresti, D.* (2005, May). Leveraging the CAPTCHA problem. In *International Workshop on Human Interactive Proofs* (pp. 97–110). Springer, Berlin, Heidelberg.
11. *McInerney, Michael James, Mark Evans Brighton, Sevag Demirjian, and Blair Livingstone Hotchkies.* "Turing test via failure." U.S. Patent 10,262,121, issued April 16, 2019.
12. *Montenegro, Juan Manuel Fernandez, and Vasileios Argyriou.* "Cognitive evaluation for the diagnosis of Alzheimer's disease based on turing test and virtual environments." *Physiology & behavior* 173 (2017): 42–51.
13. *Papasalouros, A., Kanaris, K., & Kotis, K.* (2008, July). Automatic Generation of Multiple Choice Questions From Domain Ontologies. In *e-Learning* (pp. 427–434)
14. *Radford, Alec, et al.* "Language models are unsupervised multitask learners." *OpenAI Blog* 1.8 (2019): 9.
15. *Roediger III, Henry L., and Elizabeth J. Marsh.* "The positive and negative consequences of multiple-choice testing." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31.5 (2005): 1155

16. Ryan, Ann Marie, and Gary J. Greguras. "Life is not multiple choice: Reactions to the alternatives." *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (1998): 183–202.
17. Ryan, Richard M., and Netta Weinstein. "Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing." *Theory and research in education* 7.2 (2009): 224–233.
18. Subramanian, Sandeep, et al. "Neural models for key phrase detection and question generation." arXiv preprint arXiv:1706.04560 (2017).
19. Tarasov D. S. A method for creation of tests. Patent application (Russia) N 2020103928 from 29.01.2020
20. Tarasov D. S. A way to create test items to test the depth of knowledge and the ability to reason students and specialists. Patent application (Russia) N 2019127875 from 04.09.2019
21. Tikhomirov M. M., Loukachevitch N. V., Dobrov B. V. (2019). [Assessing Theme Adherence in Student Thesis](#). Computational Linguistics and Intellectual Technologies: Proceedings of Annual International Conference "Dialogue", Issue 18
22. Zubarev D. V., Sochenkov I. V. (2019). [Cross-Language Text Alignment for Plagiarism Detection Based on Contextual and Context-Free Models](#). Computational Linguistics and Intellectual Technologies: Proceedings of Annual International Conference "Dialogue", Issue 18.

СЕМАНТИКА ОБРАТНО: ВОЗВРАЩЕНИЕ В ПРЕВАННОЕ СОСТОЯНИЕ

Татевосов С. Г. (tatevosov@gmail.com)

МГУ им. М. В. Ломоносова, Москва, Россия

Киселева К. Л. (xkisseleva@gmail.com)

ИРЯ им. В. В. Виноградова РАН, Москва, Россия

DOI: 10.28995/2075-7182-2020-19-708-723

SEMANTICS FOR *ОБРАТНО*: A RE-ENTRY INTO A DISCONTINUED STATE

Tatevosov S. G. (tatevosov@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

Kisseleva X. L. (xkisseleva@gmail.com)

Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

This paper explores the meaning and distribution of *obratno*, one of the Russian repetitive and restitutive morphemes. We identify three essential characteristics of *obratno*: obligatoriness of the restitutive reading, narrow scope with respect of indefinites, and incompatibility with eventuality descriptions that entail a result state in the sense of [Kratzer 2000]. We argue that like garden-variety repetitive and restitutive morphemes (e.g., Russian *opjat'*), *obratno* denotes a partial identity function with a presupposition. Unlike such morphemes, however, the presuppositional content of *obratno* involves a return to the same state in which an entity had been before. We capture this characteristic relying on [Landman's 2008] notion of cross-temporal identity of eventualities and the derivative notion of a cross-temporal substate. This makes the repetitive reading of *obratno* unavailable, forces identity of the holders of a state, deriving the narrow scope effect, and guarantees that *obratno* is only compatible with target state descriptions.

Keywords: event semantics, repetitive and restitutive morphemes, scope relations, state descriptions

1. Введение

Эти заметки посвящены репетитивно-реститутивным показателям (РПП), или, в другой терминологии, рефактивным морфемам ([Иванова 1993], [von Stechow 1996], [Пайар 1998], [Beck & Johnson 2004], [Bale 2007], [Beck 2005], [Стойнова 2012] и ряд других работ, [Alexiadou et al. 2014], [Lechner et al. 2015], [Pedersen 2015], [Бондаренко 2017] и многие другие), и одному из их частных воплощений в русском языке — слову *обратно*.

Классический пример РПП — английское *again* и его аналоги в других языках, в частности, русское *опять*. Вклад *again* и *опять* в интерпретацию — пресуппозиция, в соответствии с которой ситуация, удовлетворяющая событийной дескрипции, имела место в прошлом по отношению к утверждаемой ситуации:

$$(1) \quad || \text{опять} || = \lambda P_{\langle v, t \rangle}. \lambda e: \exists e' [e' \prec_T e \wedge P(e')]. P(e)^1$$

где v — логический тип событий, а « \prec_T » — отношение временного предшествования.

При таком понимании предложение (2) ниже верно описывает действительность, если, во-первых, Волга замерзла (ассерция) и, во-вторых, до этого Волга была замерзшей хотя бы единожды (пресуппозиция).

(2) *Волга, было, тронулась. Но потом опять замерзла. Я ходил в Астрахани смотреть... Белая, широкая, льдины сверкают на солнце.* [Л. Вергинская. Синяя птица любви (2004)]²

Предложение (3) с наречием *обратно*, на первый взгляд, имеет аналогичные условия истинности: оно осмысленно, только если удовлетворяется пресуппозиция, согласно которой снег уже был замерзшим. Ассерция, как и в предыдущем случае, состоит в том, что снег замерз.

(3) *Фирн — это снег, который подтаял и замерз обратно, снова подтаял и снова замерз.* [www.alpinabook.ru]

¹ Следуя практике, сложившейся, в частности, в [Heim, Kratzer 1998], в пределах этой статьи мы понимаем пресуппозицию как условие, ограничивающее область определения функции. В (1) условие — это наличие в актуальном мире предшествующего события из экстенционала предиката P . (Пресуппозиция записывается через двоеточие после переменной e , пробегающей по событиям из области отправления функции.)

² Анонимный рецензент «Диалога» пишет: «Безупречно с реститутивным *обратно*, на наш взгляд, только примеры с предикатами, в семантике которых присутствует возврат объекта в исходную точку или к исходному обладателю». Суждения по поводу РПП, безусловно, подвержены значительному варьированию и часто не отличаются четкостью: носители склонны оценивать материал не в терминах «можно» и «нельзя», а в терминах «лучше» и «хуже». Где возможно, мы отражаем такой характер суждений, используя градуальные пометы типа «^{2/3}». Для выяснения более объективной картины необходимо полномасштабное экспериментальное исследование с извлечением достаточного для статистической оценки количества суждений. Читатель может самостоятельно решить, насколько его суждения совпадают с системой, изложенной далее.

Если исходить из внешнего параллелизма (2)–(3), возникает желание анализировать (3) в том же духе, что и (2), а различия между ними искать в тонких ограничениях на содержание пресуппозиции, как это обычно и делается для разных групп РПИ [Alexiadou et al. 2014].

Цель этой статьи — представить аргументы в пользу того, что пресуппозиционное содержание *обратно* имеет принципиально другой характер. *Обратно*, в отличие от *опять*, сигнализирует не о повторном наступлении ситуации ‘замерзнуть’, а о возвращении в уже существовавшее состояние ‘быть замерзшим’.

В **разделе 2** мы обсудим три важнейшие семантические характеристики *обратно*. Это, во-первых, **невозможность для *обратно* иметь репетитивную интерпретацию**; во-вторых, **ограничения на сферу действия *обратно*** по отношению к неопределенным именованным группам; в-третьих, **лексические ограничения на дистрибуцию *обратно***, связанные с характером результирующего состояния.

В **разделе 3** мы обсудим и отклоним анализ *обратно* как аналога *опять* с более узкой сферой действия. В **разделе 4** будет предложен анализ, опирающийся на идею, что *обратно* нуждается не в повторении некоторого состояния, а возвращение в то состояние, которое имело место ранее и было прервано. В **разделе 5** суммируются основные результаты исследования³.

2. Реститутивность, сфера действия и результирующее состояние

Первое важное отличие *обратно* от *опять* состоит в диапазоне доступных для них интерпретаций. *Опять*, будучи РПИ, допускает два основных прочтения, характерных для многих единиц такого рода во множестве языков — репетитивное и реститутивное⁴. Характерный пример — интерпретация *опять* в комбинации с глаголами результата типа *открыть*:

- (4) Репетитивный контекст. [До прихода Али-Бабы Сезам был закрыт. Али-Баба открыл его. Затем пришли разбойники и закрыли Сезам. И тогда, недовольный этим, ...] ... *Али-Баба опять открыл Сезам.*
- (5) Реститутивный контекст. [Сезам был открыт от начала времен. Злой джинн, желающий навредить людям, закрыл Сезам. Но появился отважный Али-Баба, и ...] ... *он опять открыл Сезам.*

³ Мы полностью разделяем мысль анонимного рецензента «Диалога»: *обратно* «было бы хорошо рассмотреть не только в сравнении с *опять*, но и в сравнении с *назад*». Однако реализация этой идеи невозможна по соображениям объема в пределах одной статьи.

⁴ Согласно Н. М. Стойновой ([Стойнова 2012] и другие работы), помимо репетитивного и реститутивного значений, типологически засвидетельствовано и значительное количество других единиц, организованных в семантическую зону рефактива. В пределах этой статьи мы не имеем возможности обсуждать структуру и семантическое содержание значений из этой зоны. Мы надеемся, что сказанное ниже совместимо более чем с единственным представлением о ее устройстве. Ниже, однако, мы коснемся одного из выделяемых Н. М. Стойновой значений — редитивного.

Различие двух прочтений сводится к тому, что в первом случае повторяется ситуация целиком — и действия агенса, и результирующее состояние 'быть открытым', которое приобретает пациенс. Во втором случае повторяется только результирующее состояние.

При определении диапазона интерпретаций РПП, однако, требуется осторожность. Как показывают В. Лехнер и соавторы [Lechner et al. 2015], проблема состоит в том, что реститутивное прочтение в (5) — это логическое следствие репетитивного в (4). Это легко увидеть в (6a)–(6b), которые представляют собой неформальное толкование (4)–(5).

- (6) а. Повторно произошло следующее: Али-Баба совершил над дверью некоторые действия, результатом которых стало состояние 'дверь открыта'
 б. Али-Баба совершил над дверью некоторые действия, результатом которых стало повторное состояние 'дверь открыта'

Если верно, что дважды произошла деятельность и наступило состояние 'быть открытым', как в (6a), верно и то, что дважды наступило состояние 'быть открытым', как в (6b). Обратное неверно. Так происходит в силу того, что истинность конъюнкции ('произошло действие' ∧ 'наступило состояние') влечет истинность любого из конъюнктов (в частности, 'наступило состояние'), но не наоборот.

С асимметричным следованием репетитивного и реститутивного прочтений возникает следующее затруднение: невозможна ситуация, когда предложение с *опять* истинно в репетитивном прочтении и ложно в реститутивном. Это означает, что если *опять* выражает только более слабое реститутивное значение, оно заведомо может употребляться и в более сильном репетитивном контексте. Точно так же любой объект, описываемый как *холостяк*, описывается и как *мужчина*. Но даже малоопытные лингвисты не склонны думать, что слово *мужчина*, кроме значения 'мужчина', выражает еще и значение 'холостяк'.

Таким образом, примеры типа (5) несомненно показывают реальность реститутивного прочтения у *опять*. Однако примеры типа (4) не показывают, что у *опять* есть репетитивное значение, отличное от реститутивного.

Чтобы снять эту проблему, необходимо создать конфигурацию, разрушающую асимметричное следование. В. Лехнер [Lechner et al. 2015] предлагает для этого конструкции с немонотонными кванторами типа 'ровно n NP':

- (7) [Студенты А, В и С занимаются в библиотеке. Они хотят, чтобы окно было открыто, а библиотекарь не любит сквозняки. Студент А открывает окно, но библиотекарь закрывает его. Затем то же самое происходит со студентами В и С. Наконец, А открывает окно еще раз.]
Ровно один студент опять открыл окно.

При репетитивном прочтении предложение истинно: имеется ровно один студент, который открыл окно два раза. Это студент А. При реститутивном прочтении, однако, предложение ложно. Окно повторно входило в состояние 'быть открытым' трижды, и имеется ровно три студента, которые своими действиями способствовали этому. Это (в хронологической последовательности) студенты В, С и А.

Таким образом, в этой конфигурации асимметричного следования нет, и если предложение в заданном контексте расценивается как правильно описывающее то, что произошло, мы имеем свидетельство реальности репетитивного значения, отдельного от реститутивного. По всей видимости, для *опять* это действительно так: по оценке носителей, (7) в описанной ситуации истинно.

Опять, таким образом, имеет и репетитивное и реститутивное прочтение.

В этом месте обнаруживается первое различие *опять* и *обратно*: у последнего репетитивного значения нет.

- (8) [Студенты А, В и С занимаются в библиотеке. Они хотят, чтобы окно было открыто, а библиотекарь не любит сквозняки. Студент А открывает окно, но библиотекарь закрывает его. Затем то же самое происходит со студентами В и С. Наконец, А открывает окно еще раз.]

#Ровно один студент открыл окно обратно.

В том же контексте, что и (7), предложение с *обратно* оценивается как ложное. Студентов, которые открыли окно обратно, в этой ситуации имеется три. Таким образом, *опять* имеет и репетитивное и реститутивное прочтение, *обратно* — только реститутивное⁵.

Второе свойство, разводящее *опять* и *обратно*, — их сфера действия по отношению к неопределенным именованным группам. Рассмотрим следующий сценарий:

- (9) Володя находится в комнате с двумя закрытыми окнами. В комнате слишком жарко, и Володя открывает окно. Это не помогает: по-прежнему слишком жарко, и тогда Володя открывает второе окно.

В момент открытия второго окна сторонний наблюдатель может описать происходящее как (10):

- (10) *Володя опять открыл окно.*

Предложение (11) в той ситуации, которая описана в (9), по-видимому, невозможно:

- (11) *#Володя открыл окно обратно.*

В рассматриваемом контексте *опять* имеет широкую сферу действия по отношению к именно группе *окно*:

⁵ Дополнительный аргумент в пользу обязательности реститутивного прочтения *обратно* — предикаты, описывающие смену обладателя, типа *отдать обратно*. Предикат *отдать книгу Феликсу* описывает действие агенса, каузирующее состояние 'книга у Феликса'. Предпочтительная интерпретация *обратно* в комбинации с такими глаголами — именно реститутивная: *отдал книгу обратно* означает 'совершил действие, в результате которого книга еще раз оказалась у Феликса', но не 'еще раз совершил действие, в результате которого книга оказалась у Феликса'. Последняя, репетитивная интерпретация реализуется с *опять*; см. обсуждение в [Бондаренко 2017]. Сделать это уточнение нас побудил комментарий анонимного рецензента «Диалога».

(12) РРП > Э

а. опять [ле. Эх [окно(х) ∧ открыл(х)(Володя)(е)]

б. Еще раз произошло событие такое, что в этом событии имеется окно, которое открыл Володя.

Если *опять* имеет узкую сферу действия по отношению к именной группе *окно*, то утверждается, что имеется такое окно, которое подверглось открытию дважды:

(13) Э > РРП

а. Эх [окно(х) ∧ опять [ле. открыл(х)(Володя)(е)]

б. Имеется окно такое, что еще раз произошло событие, в котором Володя его открыл.

Прочтение (13) допускается и с *опять* и с *обратно*:

(14) [В комнате с единственным окном было жарко, и Володя открыл его.

Однако из-за сквозняка окно закрылось, и тогда...]

а. *Володя опять открыл окно.*б. *Володя открыл окно обратно.*

Дополнительное подтверждение этим наблюдениям дают глаголы созидания. Поскольку один и тот же объект, как правило, нельзя создать дважды, интерпретация с широкой сферой действия неопределенной именной группы исключается. Ожидаемым образом *обратно*, но не *опять* с большинством глаголов созидания оказывается семантически аномальным.

(15) *Надя сварила суп, но Володя нечаянно опрокинул на пол кастрюлю, и тогда...*а. ... ^{??/*}*Надя сварила суп обратно.*б. ... *Надя опять сварила суп.*

Отдельную группу среди глаголов созидания образуют единицы типа *собрать*, *сложить*, *склеить*, описывающие процесс соединения частей в качественно новое целое⁶. В той степени, в которой это целое можно разять на исходные части, а части затем еще раз соединить, такие глаголы ожидаемо допускают сценарии с повторным созиданием, а *обратно* в этом случае получает существенно более высокие оценки приемлемости, чем в (15а).

(16) *Володя собрал модель броневика, но Феликс уронил его на пол, и модель рассыпалась. И тогда...*а. ... ^{2/ОК}*Володя собрал модель обратно.*

Таким образом, если *опять* допускает любую сферу действия по отношению к неопределенной именной группе, *обратно* имеет исключительно узкую сферу.

⁶ Мы признательны анонимному рецензенту «Диалога», обратившему наше внимание на примеры типа (16).

Мы предполагаем, что обсуждаемые Н. М. Стойновой [Стойнова 2012b: 81] примеры, которые воспроизводятся в (16)–(17), — реализация этой же закономерности.

(17) *Я приехала из Вятки в Москву и, пробыв там всего час, снова уехала.*

- а. ^{ок}‘в Вятку’ б. ^{ок}‘не в Вятку’

(18) *Я приехала из Вятки в Москву и, пробыв там всего час, уехала обратно.*

- а. ^{ок}‘в Вятку’ б. *‘не в Вятку’

Глаголы движения имеют валентности на актанты, обозначающие исходную и конечную точку перемещения, или пути. Синтаксическая реализация этих актантов необязательна, и в отсутствие выраженных предложных групп они получают экзистенциальную интерпретацию (‘имеется такая исходная / конечная точка, что ...’). Глагол *уехать*, в частности, обозначает в первом приближении перемещение из дейктического центра в локализацию, находящуюся за его пределами. Предложение типа *Наташа уехала* содержит предикат в (19a), неформально описанный в (19b):

- (19) а. || Наташа уехала ||^g = || Наташа уехала из *pro*₅ в DP ||^g = λe. ∃!
 [move(Наташа)(e) ∧ source(g(5))(e) ∧ CAUSE(s)(e) ∧ BE(Наташа)(l)(s) ∧ l ≠ g(5)]
 б. Множество событий, в которых происходит перемещение Наташи из дейктического центра и в результате которых она приобретает состояние ‘находиться в некоторой локализации, отличной от дейктического центра’.

Если синтаксически нереализованная локализация (в (19a) такой материал показан более бледным шрифтом) интерпретируется экзистенциально, это открывает ровно те же две возможности, что и в обсуждавшемся выше случае с неопределенным дополнением. Квантор существования может иметь широкую сферу действия по отношению к РПП, либо наоборот. В (19a) именно такую интерпретацию имеет конечная точка, в которой оказывается агент в результате перемещения; ей соответствует переменная *l*, связанная квантором существования.

В (20a) показан только стативный компонент семантического представления ‘уехать’, поскольку при реститутивном прочтении РПП взаимодействует именно с ним:

- (20) РПП > ∃
 а. ... опять [λs. ∃! [... BE(Наташа)(l)(s) ∧ l ≠ g(5) ...]
 б. Повторно возникло состояние такое, что Наташа находится в локализации, отличной от дейктического центра.

При таком взаиморасположении РПП и квантора, естественно, не требуется, чтобы локализации были идентичны — в точности как в (12) с открываемым окном. Этой возможности соответствует случай ‘не в Вятку’ в (17b) и (18b).

- (21) ∃ > РПП
 а. ∃! ... l ≠ g(5) ... опять [λs. [... BE(Наташа)(l)(s) ...]
 б. Имеется такая локализация, отличная от дейктического центра, в которой Наташа оказалась повторно.

При широкой сфере действия квантора в (21) агенс повторно оказывается в той же самой локализации, что и в первый раз. Это прочтение, которое соответствует (17а) и (18а).

Очевидно, что (17)–(18) образуют тот же самый паттерн, что и (9)–(14), и их анализ сводится к соотношению сферы действия РРП и квантора существования. *Опять* совместимо с любой сферой действия, *обратно* нуждается в широкой сфере действия квантора⁷.

Наконец, последнее свойство *обратно*, которое не воспроизводится у *опять* — наличие значительного количества лексических ограничений. Вот один пример:

- (22) a. Володя накачал лодку обратно.
b. ??/*Эндрю доказал теорему Ферма обратно.

- (23) a. Володя опять накачал лодку.
b. Эндрю опять доказал теорему Ферма.

Опять, по всей видимости, имеет минимальные лексические ограничения, если вообще имеет. В (22)–(23), чтобы избежать возмущающего фактора, связанного со сферой действия неопределенных именных групп, в качестве прямого дополнения используется определенная дескрипция *теорема Ферма*. Предположим сценарий в (24):

- (24) В 1993 году Эндрю предложил доказательство теоремы Ферма, безоговорочно принятое научным сообществом. Однако в 1994 году в доказательстве обнаружилась не замеченная сразу критическая ошибка. Эндрю переработал доказательство, и в 1995 году опубликовал окончательный вариант.

При таком сценарии (23b) вполне пригодно для описания того, что произошло в 1995 году. (22b), однако, выглядит аномальным.

Кажется правдоподобным, что в этих и аналогичных примерах задействуется противопоставление, описанное А. Кратцер [Kratzer 2000] как противопоставление целевых (target states) и результирующих состояний (result states). Целевое состояние — элемент лексического значения глагола; оно составляет неотъемлемую часть дескрипции ситуации, или, что то же самое, толкования глагола. Целевые состояния могут характеризоваться большей или меньшей временной стабильностью; в общем случае целевые состояния могут прекращаться. Результирующие состояния не являются элементом лексического представления. Результирующее состояние имеет место просто в силу осуществления предшествующего действия или процесса и как таковое наступает навечно. Это новое состояние мироздания, которое имеет место после осуществления любой ситуации. Соответственно, применительно к любому

⁷ Для Н. М. Стойновой примеры (17)–(18) создают ключевое эмпирическое свидетельство реальности редитивного значения 'возвратиться в исходную точку', отличного от реститутивного. Как представляется, предложенные выше рассуждения делают этот материал полностью совместимым с анализом в реститутивных терминах.

глаголу можно говорить о результирующем состоянии. У некоторых глаголов имеется также целевое состояние.

Целевые состояния соответствуют стативному компоненту сложных событийных дескрипций класса свершений. В событийно-семантических представлениях им соответствует стативная переменная ($\lambda s.\lambda e....$). Результирующие состояния А. Кратцер анализирует как временные объекты: это интервал, который начинается немедленно по завершении описываемой ситуации; ему соответствует временная, а не событийная переменная ($\lambda t. \exists e.... \tau(e) <_T t \dots$).

Важное свойство, разводящее дескрипции с целевым и с результирующим состоянием,— совместимость первых, но не вторых с обстоятельствами типа ‘еще’:

(25) *Лодка еще надута*

(26) *#Володя еще мертв* (приемлемо при определенном устройстве мира, отличном от нашего)

(27) *??Теорема еще доказана* (аномально при любом устройстве мира)

Примеры такого рода показывают, что допустимость *еще* коварирует с допустимостью *обратно*. Дополнительный пример такого же рода — предикаты с глаголом *написать*, которые в зависимости от типа прямого дополнения проявляют свойства дескрипций или с результирующим, или с целевым состоянием (что коррелирует, как справедливо заметил рецензент «Диалога», с различиями в интерпретации самого глагола: ‘сочинить, зафиксировав графически’ и ‘нанести на поверхность надпись того или иного содержания’):

(28) [В тот момент, когда Володя закончил писать «Материализм и эмпириокритицизм», огонь уничтожил дом вместе с рукописью, но за два месяца ...]

*???*Володя написал эту книгу обратно.

(29) [На заборе был написан антиправительственный лозунг. Дворник смыл его, но на следующий день ...]

Володя написал этот лозунг обратно.

Ограничения на *обратно* в (28)–(29) коррелируют с дистрибуцией *еще*, как и ожидается.

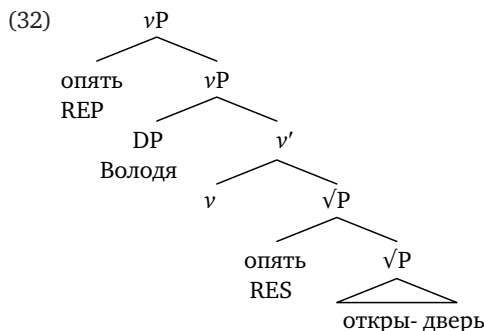
(30) *???*Роман еще написан.

(31) *Лозунг еще написан (на заборе)*.

Таким образом, возможность употребления *обратно* зависит от характера стативного компонента описания ситуации: состояние должно быть целевым, но не результирующим в кратцеровском смысле. *Опять* не подвержено этому ограничению.

3. *Обратно как опять с узкой сферой действия*

Напрашивающийся, но эмпирически неадекватный вариант анализа *обратно* — рассматривать его как вариант *опять* с более узкой сферой действия. Классические работы А. фон Штехова [von Stechow 1996], [Rapp, von Stechow 1999] и практически вся последующая литература выводят репетитивно-реститутивную полисемию *опять*, опираясь на структурные соображения. *Опять* может присоединяться более чем в одной синтаксической позиции. При низком присоединении в его сферу действия попадает только результирующее состояние. При высоком — вся событийная дескрипция. Вариант анализа в этом духе показан в (32):



В такой системе можно было бы предположить, что для *обратно* доступна лишь одна из этих двух возможностей — «низкая».

Такой анализ объясняет один из трех фактов, описанных выше — то, что для *обратно* доступно только реститутивное прочтение, как в (8). Однако для двух других фактов — узкой сферы действия по отношению к неопределенным именным группам и ограничений, связанных с характером результирующего состояния, — объяснение не выстраивается.

РПП при реститутивном прочтении в общем случае не требуют узкой сферы действия по отношению к неопределенным именным группам. Это можно увидеть на примере *опять*: достаточно создать контекст, форсирующий одновременно реститутивную интерпретацию и широкую сферу действия *опять* по отношению к неопределенной именной группе:

- (33) Володя входит в комнату с двумя окнами — открытым окном №1 и закрытым окном №2. Через некоторое время порыв ветра закрывает окно №1. И тогда Володя открывает окно №2.

В таком контексте открытие окна №2 можно описать как *Володя опять открыл окно* — ‘Володя сделал так, чтобы в комнате еще раз было открытое окно’ Это означает, что в общем случае низкое присоединение РПП не влечет за собой низкую сферу действия по отношению к квантору существования. Однако эта возможность недоступна *обратно*: в описываемой в (33) ситуации предложение *Володя открыл окно обратно* ложно. Следовательно, одно только низкое присоединение не объясняет обязательность узкой сферы действия *обратно*; необходимы какие-то дополнительные допущения, но не очень ясно, какие именно.

Схожая проблема возникает и при объяснении ограничений на характер состояния. *Опять* в реститутивном прочтении совместимо не только с целевыми, но и результирующими состояниями. Рассмотрим следующий контекст:

(34) [Эндрю доказал теорему Ферма и вскоре умер. Затем Раджеш опубликовал статью, в которой показал, что в рассуждения Эндрю вкралась ошибка. Прошел еще год, и] Володя опять доказал теорему Ферма.

В (34) *опять* имеет реститутивное прочтение: два раза наступает только состояние ‘быть доказанным’. Деятельность Володи по построению доказательства находится вне сферы действия *опять*, поскольку происходит только один раз; в предыдущий раз это была деятельность Эндрю.

Таким образом, если принять анализ, согласно которому *обратно* — это *опять*, ограниченное реститутивным прочтением, семантическая аномальность примеров типа (9)–(11) остается без объяснения.

Дополнительное сомнение по поводу этого анализа возникает, как только мы обращаем внимание на аномальность *обратно* в комбинации с дескрипциями непроеизводных состояний, таких, как (35):

(35) ^{???}Володя болен *обратно*.

Вероятно, предложения типа (35) возможны в некоторых диалектных вариантах русского языка — таких, где возможны, например, и предложения типа (36):

(36) ^{???}*Обратно он глупости говорит*.

В таких диалектах, по-видимому, *обратно* эволюционирует в полноценный аналог *опять*. В литературном русском, однако, предложения (35) и (36) аномальны, хотя, по-видимому, в несколько разной степени для разных носителей.

Однако если *обратно* присоединяется к дескрипции стативного компонента сложной событийной структуры, как в (32), и более того, *только* к стативному компоненту, становится решительно непонятно, что предотвращает его присоединение к дескрипции непроеизводного состояния в (35).

Таким образом, анализ, предполагающий, что *обратно* — это как бы ополовиненное *опять*, не предсказывает наблюдаемые факты и нуждается в серьезной модификации. В следующем разделе мы предложим альтернативу, которая, как мы надеемся, окажется более работоспособной.

4. Семантика *обратно*

Мы исходим из предположения, что *обратно*, как и *опять*, — тождественное отображение с пресуппозицией. Это функция, аргументом которой выступает отношение между индивидами и событиями. Если пресуппозиция удовлетворяется, значением выступает это же отношение; в противном случае значение не определено.

(37) || *обратно* || = $\lambda R_{\langle v, \langle vt \rangle \rangle} \cdot \lambda s. \lambda e: \dots \cdot R(s)(e)$

Однако этим исчерпывается сходство *обратно* и *опять*. Мы предполагаем, что основной компонент семантики — пресуппозициональное содержание — у *обратно* устроено принципиально иначе, чем у *опять*. Пресуппозиция *обратно* — не наступление еще одного состояния, удовлетворяющего нужной дескрипции, а возвращение в то же самое состояние, которое уже имело место ранее и было прервано.

Семантика *обратно* выписана в (38):

$$(38) \quad || \text{обратно} || = \lambda R_{\langle v, \langle v, t \rangle \rangle}. \lambda s. \lambda e. : \exists s' \exists s_0 [s' \prec_T s \wedge \neg s' \infty_T s \wedge s' \leq_s s_0 \wedge s \leq_s s_0 \wedge Q_C(s_0)]. R(s)(e)$$

где « \prec_T » — отношение временного предшествования,

« ∞_T » — отношение временного примыкания и

« \leq_s » — отношение межвременного подсостояния

Согласно (38), *обратно* применяется к отношению между событиями и состояниями R логического типа $\langle v, \langle v, t \rangle \rangle$, и если пресуппозиция удовлетворяется, возвращает это же отношение. Пресуппозиция требует, чтобы для любого стативного элемента s отношения R нашлось состояние s' такое, что, во-первых, s' предшествует s и не примыкает к нему во времени, а во-вторых, s' и s оба выступают (межвременными) подсостояниями некоторого состояния s_0 . Наконец, s_0 должно удовлетворять контекстно-зависимой дескрипции Q_C . Иными словами, s — продолжение состояния s_0 , другая часть которого, s' , уже имела место в прошлом; s_0 — состояние с временным разрывом.

(39) Отношение межвременного подсостояния

Состояние s — это межвременное подсостояние состояния s' , $s \leq_s s'$, ровно в том случае, когда s — это мереологическая часть s' и $s \sim s'$ характеризуются межвременной идентичностью.

Понятие межвременной идентичности событий или состояний вводится, хотя и по другому поводу, в работах Ф. Ландмана [Landman 2008], [Landman, Rothstein 2010]. Это понятие отражает интуицию, согласно которой во многих случаях мы готовы воспринимать некоторые положения дел как одну и ту же ситуацию. Межвременная идентичность позволяет нам, в частности, рассматривать замерзшую воду вчера утром и вчера вечером как находящуюся в одном и том же состоянии. В общем случае вопрос о межвременной идентичности возникает, когда есть эмпирическая потребность рассматривать положения дел, имеющие место на разных временных интервалах, как одну и ту же сущность.

Межвременная идентичность — теоретический примитив, содержание которого далее ограничивается семантическими аксиомами. Не имея возможности входить в подробности, мы отсылаем читателя к первоисточнику [Landman 2008] и ограничимся лишь одним семантическим постулатом, важным для содержания (38):

(40) Свойство замкнутости для стативных глаголов

$$\text{If } s_1 \in || \alpha || \text{ and } s_1 \sim s_2 \text{ then } s_2 \in || \alpha ||$$

где α — стативный глагол, $|| \alpha || \subseteq \text{STATE}$ — его интерпретация,

« \sim » — отношение межвременной идентичности.

В соответствии с (40), если, например, s — это состояние ‘быть замерзшим’, то любое состояние, характеризующееся временной идентичностью s — это также состояние ‘быть замерзшим’. Для *обратно* в (38) это свойство гарантирует, что состояния s_0 и s' в пресуппозициональной части будут удовлетворять тому же предикату, что и состояние s в ассертивной части. Это определяет дескриптивные свойства s_0 и s' через целевое состояние s исходного глагола.

Итак, *обратно* эксплуатирует идею межвременной идентичности и именно этим отличается от *опять*. *Опять* допускает, чтобы при повторной реализации имело место другое состояние, пусть даже описываемое так же, как состояние в первой реализации. *Обратно* вводит более сильное требование: чтобы наступило продолжение того же самого состояния, которое уже имело место. *Обратно* дает нам одно и то же состояние, наблюдаемое в разное время, *опять* — два разных состояния, которые, однако, имеют абсолютно одинаковое описание.

С эпистемологической точки зрения непросто развести случаи, когда мы дважды наблюдаем один и тот же объект и когда мы имеем дело с разными объектами, которые только выглядят одинаково. Однако удачным образом именно из этого различия следуют те факты, которые отличают *опять* от *обратно* и не объясняются при анализе *обратно* как реститутивного *опять*.

Первый факт — узкая сфера действия по отношению к неопределенным именованным группам, когда требуется, чтобы в повторной реализации состояния участвовал тот же самый объект, что и при первичной реализации. Это ограничение следует из (38) самым непосредственным образом. У одной и той же ситуации не могут разные участники с одной семантической ролью. Этот принцип известен в событийной семантике как принцип уникальности участников.

(41) Уникальность участников

$$\forall R \forall e \forall x \forall y [R(x)(e) \wedge R(y)(e) \rightarrow x = y]$$

Поэтому если повторная реализация состояния — это реализация того же самого состояния, в нем по необходимости задействуется тот же самый участник. Благодаря (41) одно и то же состояние не может иметь двух разных носителей. Так возникает эффект широкой сферы действия.

Второй факт — ограничения, связанные с противопоставлением кратцевровских целевых и результирующих состояний. По этому поводу выстраивается следующее рассуждение.

Обратно требует, чтобы одно и то же состояние было прервано и возобновлено. Чтобы выполнить это требование, состояние в принципе должно быть прерываемым и возобновляемым. Таковы целевые, но не результирующие состояния. Последние, как мы видели, неотменимо наступают в момент завершения описываемой ситуации. Невозможность прийти *обратно* в результирующее состояние, которую мы наблюдаем в (22b) и (28), следует из этого.

Иное дело *опять*, которое нуждается в двух разных состояниях, удовлетворяющих одинаковой дескрипции. Два разных состояния могут располагаться друг по отношению к другу во времени как угодно. Они могут примыкать, но могут быть и разделены «пустым» интервалом, и это не зависит от того, является ли состояние целевым или результирующим. Поэтому для *опять* безразличен характер состояния.

В обоих случаях единичность состояния мы видим по косвенным признакам: через единичность участников и невозможность прерывания, если состояние результирующее. Это дает нам самые важные свидетельства в пользу анализа в духе (38).

Отметим роль контекстно-доступной дескрипции Q_C в этом описании. Смысл ее в том, что прерываемое и возобновляемое состояние должно определяться не только как целевое состояние глагола, с которым соединяется *опять*, но и обладать собственными положительными дескриптивными свойствами. Мы предполагаем, что именно это ограничение отвечает за невозможность или сомнительность (по меньшей мере в нулевом контексте) предложений типа (42a)–(42b):

- (42) а. *??Володя нагулялся обратно.*
 б. *??Володя выспался обратно.*

Если префиксально-суффиксальные глаголы типа (42a)–(42b) описывают целевые состояния, то это такие состояния, для которой в русском языке нет языковой дескрипции, отличной от той, которую дают сами эти глаголы. Состояния «нагулятости» или «выспатости» никак иначе не называются. Удовлетворить требование, согласно которому у прерываемого состояния должна быть независимая дескрипция, становится невозможно.

Мы ожидаем, соответственно, что чем более в текущем контексте доступна дескрипция прерываемого и возобновляемого состояния, тем более уместно *обратно*. По нашему предположению, именно с этим связана асимметрия, которой интересуется анонимный рецензент «Диалога»: почему *заснул обратно* дается с меньшим усилием, чем *проснулся обратно*. Заснуть — это войти в состояние сна, и дескрипция этого состояния содержится в самом глаголе. Но проснуться — это не войти в состояние бодрствования, а выйти из состояния сна; глагол дает результирующему состоянию не положительную дескрипцию, а отрицательную. Однако минимальной контекстной поддержки достаточно в этом случае для идентификации интересующего нас состояния, и *проснулся обратно* становится приемлемым. На это, в частности, указывают примеры типа (43), которые не выглядят ни аномальными, ни даже искусственными.

- (43) а. *Змеенок тоже задумался, и даже почти заснул от своей задумчивости, но начал совсем коченеть — и проснулся обратно.* [kulichki.net]
 б. *Я попил с ними немного, чтобы хоть чуть-чуть захотеть спать, но всё равно часа через полтора проснулся обратно и глядел на дорогу.* [М. Уткин. Особенности национальной командировки (2017)]

В завершение отметим два технических аспекта (38). Аргумент *обратно* — отношение между событиями и состояниями. Это, в частности, исключает присоединение *обратно* к дескрипциям производных состояний и объясняет невозможность предложений типа (35) выше. Во-вторых, пресуппозиция *обратно* вводит условие только на содержание стативного компонента сложной событийной дескрипции. Это делает невозможным появление у предложений с *обратно* репетитивной интерпретации.

Мы предсказываем, таким образом, что помимо перечисленных выше, *обратно* не имеет принципиальных ограничений на сочетаемость с глаголами. И *согреться обратно*, и *повеселеть обратно*, и *поумнеть обратно* и множество других сочетаний, вызывающих сомнения у анонимного рецензента «Диалога», засвидетельствованы в текстах на русском языке.

(44) *Запустили кондей — ништяк, пошел холод, но через 15 секунд воздух согрелся обратно.* [vwts.ru]

(45) *Младший немного погрустнел, ведь людям снова будет плохо. Впрочем, вспомнив какой сладкий вкус у человеческого горя, быстро повеселел обратно.* [vk.com/@aka_gst_gr-letuny]

(46) *От дурачков умный ушел уже дурачком. Он вернулся к умным, и вскоре поумнел обратно.* [proza.ru]

5. Заключение

Основной результат этого исследования — анализ наречия *обратно* как семантического элемента, описывающего возвращение участника ситуации в имевшее место состояние, прерванное до наступления события, составляющего ассертивный компонент предложения с *обратно*. Такой анализ позволяет объяснить комбинацию трех свойств *обратно*: реститутивное прочтение, узкую сферу действия по отношению к неопределенным именным группам и ограничения на сочетаемость с предикатами, в семантику которых входит указание на кратцовое результирующее состояние.

Важнейшие отличия *обратно* от более изученных репетитивно-реститутивных морфем — требование, чтобы повторная ситуация представляла собой то же самое состояние, что и первичная ситуация, наличие которой составляет presupпозициональное содержание *обратно*. Это требование формулируется с опорой на понятие межвременной идентичности ситуаций и производного от него отношения межвременного подстояния.

Литература

1. Alexiadou A., Anagnostopoulou E., Lechner W. (2014), Variation In Repetitive Morphemes: Some Implications For The Clausal Architecture, Paper presented at the Workshop on the State of the Art in Comparative Syntax, University of York, September 25, 2014.
2. Bale A. C. (2007), Quantifiers and verb phrases: An exploration of propositional complexity, *Natural Language and Linguistic Theory*, Vol. 25, pp. 447–483.
3. Beck S. (2005), There and back again: A semantic analysis, *Journal of Semantics*, Vol. 22, № 1, pp. 3–51.
4. Beck S., Johnson K. (2004), Double objects again, *Linguistic Inquiry*, Vol. 35, № 1, pp. 97–124.

5. *Bondarenko, T.* (2017), Repetitives in dative argument constructions [Repetitiv v konstruksijax s dativnymi argumentami]. MA thesis, Lomonosov Moscow State University, Moscow.
6. *Heim I., Kratzer A.* (1998), *Semantics in Generative Grammar*, Blackwell, Oxford.
7. *Ivanova S. A.* (1993), Semantics of repetition in Russian: evidence from adverbial and verbal synonyms [Semantika povtornosti v russkom jazyke: na materiale narechnykh i glagolnykh sinonimicheskikh rjadov], MA thesis, Lomonosov Moscow State University, Moscow.
8. *Kratzer A.* (2000), Building stative, Proceedings of the 26th Annual Meeting of the Berkeley Linguistic Society, ed. by L. J. Conathan, J. Good, D. Kavitskaya, A. B. Wulf and A. C. L. Yu, Berkeley Linguistic Society, Berkeley, pp. 385–399.
9. *Landman F.* (2008), 1066. On the differences between the tense-perspective-aspect systems of English and Dutch, Theoretical and Crosslinguistic Approaches to the Semantics of Aspect, ed. by S. Rothstein, Benjamins, Amsterdam, pp. 107–166.
10. *Landman F., Rothstein S.* (2010), Incremental homogeneity in the semantics of aspectual for-phrases, Syntax, Lexical Semantics and Event Structure, ed. by M. Rapaport Hovav, I. Sichel and E. Doron, John Benjamins, Amsterdam, pp. 229–251.
11. *Lechner W., Spathas G., Alexiadou A., Anagnostopoulou E.* (2015), On Deriving The Typology Of Repetition And Restitution, Paper presented at the 38th Generative Linguistics in the Old World colloquium, Paris, April 5, 2015.
12. *Paillard D.* (1998), *Opjat* ('again'), or there was a precedent [*Opjat*, ili pretsedent byl], Discourse markers in Russian [Diskursivnyje slova russkodo jazyka], ed. by D. Paillard and X. Kisseleva, Metatext, Moscow, pp. 115–121.
13. *Pedersen W.* (2015), A scalar analysis of again-ambiguities, *Journal of Semantics*, Vol. 32, pp. 373–424.
14. *Rapp I., von Stechow A.* (1999), Fast 'almost' and the visibility parameter for functional adverbs, *Journal of Semantics*, Vol. 16, pp. 149–204.
15. *Stojnova N. M.* (2012a), A typology of verbal markers of the refactive: Patterns of polysemy and the structure of the semantic domain [Tipologiya glagolnyh pokazatelej refaktiva: modeli polisemii i struktura semanticheskoi zony], Ph.D. dissertation, Lomonosov Moscow State University, Moscow.
16. *Stojnova N. M.* (2012b), Refactive: a typological study [Refaktiv: tipologicheskie dannye], *Topics in Linguistics [Voprosy Jazykoznanija]*, № 2, pp. 61–92.
17. *von Stechow A.* (1996), The different readings of 'wieder' ('again'): A structural account, *Journal of Semantics*, Vol. 13, pp. 87–138.

PRETRAINING AND AUGMENTATION IN NAMED ENTITY RECOGNITION TASK FOR CYBERSECURITY DOMAIN IN RUSSIAN¹

Tikhomirov M. M. (tikhomirov.mm@gmail.com),

Loukachevitch N. V. (louk_nat@mail.ru),

Sirotnina A. Yu. (overnastuhed@yandex.ru),

Dobrov B. V. (dobrov_bv@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

The paper presents the results of applying the BERT representation model in the named entity recognition task for the cybersecurity domain in Russian. Several variants of the model were investigated. The best results were obtained using the BERT model, trained on the target collection of information security texts. This model achieved results, which were 15 percentage points of F1-macro measure greater than results of CRF, the best method in previous experiments for the same task and data. We also explored a new form of data augmentation for the task of named entity recognition.

Keywords: Cybersecurity, Named entity recognition, Pretraining, Augmentation

DOI: 10.28995/2075-7182-2020-19-724-735

ПРЕДОБУЧЕНИЕ И АУГМЕНТАЦИЯ В ЗАДАЧЕ ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ ПО ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Тихомиров М. М. (tikhomirov.mm@gmail.com),

Лукашевич Н. В. (louk_nat@mail.ru),

Сиротина А. Ю. (overnastuhed@yandex.ru),

Добров Б. В. (dobrov_bv@mail.ru)

МГУ имени М. В. Ломоносова, Москва, Россия

¹ The research was supported by Russian Science Foundation (project No. 20-11-20166). Computational experiments were carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University.

1. Introduction

Automatic named entity recognition (NER) is one of the basic tasks in natural language processing. The NER methods are usually tested on well-known datasets such as CONLL-2003 for English and some other European languages [18]. For Russian, such known datasets are Gareev's dataset [7], Persons-1000 [21], Collection3 [16], FactRuEval [2]. The majority of well-known datasets consist of news documents with three types of named entities labeled: person (people's names), organization (names of organizations), location (places, mostly geographical objects). For these types of named entities, the state-of-the-art NER methods usually give impressive results.

Nevertheless, if some other types of texts are being processed or some other types of named entities are being extracted, various difficulties arise. In such cases, one has to establish new principles of annotation and to ensure that these principles are applied consistently. However, even this being done, one can still face such a problem as insufficient amount of entities of a certain type, which leads to decrease of recognition quality.

In this paper we discuss the NER task in the cybersecurity domain [19]. Several additional types of named entities for this domain were annotated if compared to general datasets such as software programs, devices, technologies, hackers, and malicious programs (vulnerabilities). The most important entities for this domain are names of malicious software and hackers. However, the annotated dataset contains a modest number of entities of these types. This could be explained by the fact that usually names of viruses and hackers are not known at the time of an attack and are revealed later.

To improve NER quality in such conditions, we suggest using BERT transformers [5] as well as an automatic dataset augmentation method, by which we mean extending a training dataset with sentences containing automatically labeled named entities.

Our paper's contribution is as follows:

- We study how quality of a NER system changes depending on variants of the BERT model used. We experimented with the following models: a multilingual model, a model fine-tuned on Russian data, and a model fine-tuned on cybersecurity texts. We compare these results with the CRF-model that previously achieved the best performance on the cybersecurity dataset.
- We introduce a new method of dataset augmentation for NER tasks and study the parameters of the method.

The remainder of the paper is organized as follows. **Section 2** presents related work. **Section 3** describes the labeled data in the cybersecurity domain used in the study. **Section 4** presents the BERT-based models and the augmentation approach specially intended for NER tasks. Section 5 describes the results of the experiments.

2. Related Works

2.1. Named Entity Recognition in Information-Security Domain

The information extraction task in cybersecurity domain has been discussed in several works. However, most works consider information extraction only from

structured or semi-structured English texts. For instance, Bridges et al. [3] used training corpora consisting of Microsoft Bulletins and National Vulnerability Database descriptions mainly. The training corpus presented in [9] does contain unstructured blog posts, but those comprise less than 10% of the corpus.

The proposed NER systems are based on such methods as principle of Maximum Entropy [3], Conditional Random Fields (CRF) [12], [9]. Gasmi et al. [6] explored two different NER approaches: the CRF-model and a neural network (NN) based model LSTM-CRF (as suggested by Lample et al. [13]). The NN-based model combined bidirectional LSTM, the word2vec representation as a source of pre-trained word embeddings and CRFs as an output layer.

In [19], the Sec_col² cybersecurity corpus for Russian named entity recognition was described. The corpus contains unstructured texts, it was collected from journal articles, news reports, and forum posts. All these data can provide additional details on cybersecurity problems. The authors compared different models for cybersecurity NER including CRF and several variants of neural networks.

2.2. Using BERT in Named Entity Recognition

The state-of-the-art models for named entity recognition utilize various contextualized vector representations. One such a popular model is BERT [5]. BERT is an implementation of a statistical language model based on deep neural networks; the task of the BERT pretraining is to predict the word in a given place in the text. The BERT architecture consists of a 12-layer transformer-encoder that forms contextualized token representations, thus converting a sequence of tokens into a sequence of vectors.

Using BERT made it possible to achieve better results in various natural language processing tasks [5], including named entity recognition. Such results are due to the high information content of vector representations, which, unlike static vector representations, such as word2vec [14], depend on the context. In addition, an important point is the use of transfer learning techniques. BERT is pretrained on a large amount of unlabeled data on the language modeling tasks, and then it is finetuned for a specific task.

Initially, BERT is multilingual, trained on multilingual data. The paper [11] describes an approach to further training of the multilingual model on the Russian-language data (Russian Wikipedia and the Russian news corpus). The new model, called RuBERT, showed an improvement in quality in three NLP tasks in Russian in comparison with previous results and multilingual BERT. The use of RuBERT in the NER task on the Russian dataset Collection3 [16] also gave a significant improvement [4].

In 2019, the named entity recognition shared task for Slavic languages was organized [17]. Most participants and the winner used BERT as the main model. An interesting detail of this competition was that there was a significant imbalance among the types of entities in the data. For example, the entity type “product” (PRO) was annotated only for 8% of all entities in the Russian data. The results of extracting this type of entities were significantly lower than for other entities, which raises the question of improving the quality of rare entity recognition in unbalanced datasets.

² <https://github.com/LAIR-RCC/InfSecurityRussianNLP>

2.3. Methods of Data Augmentation

Methods of data augmentation for natural language processing are mainly discussed for such tasks as machine translation and automatic text classification. The simplest augmentation method is to replace source words with their synonyms from manual thesauri (for example, WordNet [15]) or with words that are close to the source words according to a distributional model trained on a large text collection [24].

In [10] it was claimed that synonyms may not fit into the context, therefore the replacement words should be those that are the most probable according to a language model.

The authors of [22] used four simple augmentation techniques for the classification tasks: replacing words with their synonyms (WordNet), occasional word insertion, occasional word deletion and occasional word order changing. This method was applied to five datasets for the text classification task. Quality evaluation was presented for RNN and CNN neural networks. The average improvement of 0.8% for F-score was achieved. The study showed that all four operations contributed to the obtained improvement.

In this paper we discuss a specialized method of data augmentation for named entity recognition. We obtain additional annotated data by inserting named entities in appropriate sentences and contexts.

3. Data

We use a renewed version of Sec_col corpus [19] as a training dataset for the NER task. The final corpus contains 861 unstructured texts (more than 400,000 tokens), which are posts and comments extracted from several sources on cybersecurity.

The set of corpus labels includes four general types: PER (for persons excluding hackers), ORG (for organizations excluding hacker groups), LOC, and EVENT; and five domain-specific types such as PROGRAM (for computer programs excluding malware), DEVICE (for various electronic devices), TECH (for technologies having proper names), VIRUS (for malware and vulnerabilities), and HACKER (for single hackers and hacker groups). The corpus was pre-annotated automatically, then multi-pass manual annotation took place. The annotation principles are described in detail in [19]. The quantitative characteristics for each tag are presented in **Table 1**.

Table 1: Tag distribution

Type on entity	Description	Number of entities
ORG	organizations (not including hacker groups)	3,791
PROGRAM	programmes (software products and their parts: codes, procedures)	3,497
TECH	technologies (named methods and approaches)	2,962
LOC	locations (geographical locations)	1,376
PER	persons (names of people that are not hackers)	1,015
DEVICE	devices (various electronic devices and computer programs)	539
VIRUS	viruses (malicious software and vulnerabilities)	480
EVENT	events	301
HACKER	hackers (individual hackers and hacker groups)	60

According to the table, one of the labels, HACKER, is severely underrepresented in the dataset. One of the reasons for that could be that at the time when an attack happens, hackers are unknown, therefore their names are not mentioned. Another important type of label, VIRUS, is represented better than HACKER, but its frequency is still lower than for the other tags.

4. Models Used in Cybersecurity NER

4.1. BERT Models

As part of this study, we evaluated BERT in the cybersecurity NER task with the following pretrained weights:

- multilingual-bert-base model (BERT),
- model trained on Russian general data RuBERT,
- RuCyBERT, which was obtained by additional training RuBERT on information-security texts.

Training RuCyBERT was similar to training RuBERT [11], but without creating a new vocabulary. To do this, the pretraining procedure was launched on 500K cybersecurity texts with the initialization of all weights from RuBERT. The training lasted 500k steps with batch size 6.

All three models have the same architecture: transformer-encoder [20] with 12 transformer blocks, 12 self-attention heads and $H = 768$ hidden size. The models are fine-tuned for 6 epochs, with $B = 16$ batch size, with learning rate $5e-5$ and $T = 128$ maximum sequence length. When forming input for the model, only the first token of a word gets a real word label, the remaining tokens get a special label X. At the prediction step, the predicted label of the first token is chosen for the whole word.

4.2. Training Data Augmentation

The important classes of named entities in the cybersecurity domain are names of viruses and hackers (including hacker groups). The Sec_col collection, however, includes a quite small number of hackers' names. This could be due to the fact that names of many hackers and hacker groups are not known, therefore many texts related to cybersecurity include only unnamed descriptors (such as *hacker*, *hacker group*, *hacker community*).

Analysis of some extra texts revealed that additional manual annotation is not a reasonable solution to the problem, as most texts mention almost the same well-known hacker groups and their attacks. During the analysis, a new augmentation approach for the NER task was proposed. The core idea of the NER augmentation is as follows: in most contexts where an entity descriptor is mentioned, some other variants of mentions are possible. For Russian, such variants can be: 1) a descriptor followed by a name or 2) just the name alone. The first above-indicated variant of entity mentioning is language-specific, depends on language-specific grammar rules. Consequently, we could augment the collection by adding names after descriptors or by replacing descriptors with names.

Tables 2 and 3 show the examples of the proposed augmentation procedure in English translation. In the first pair of sentences, the descriptors were replaced with the names; in the second pair of sentences, the names were inserted after the descriptors *хакер* (*'hacker', hacker*) and *зловред* (*'zlovred', malware*). It should be noted that the sentences are given in translation into English, and for English the correct insertion of a name is before a descriptor. In parentheses, we give fragments with initial Russian augmentation.

Table 2: Augmentation examples for HACKER

Original	Modified
Replacement	
The absence of vulnerabilities on the site and its willingness to resist hacker attacks is an important issue, but often stubbornly ignored by site owners.	The absence of vulnerabilities on the site and its willingness to resist Pwn2Own attacks is an important issue, but often stubbornly ignored by site owners.
Insertion	
And the number of installed software protection tools against hackers is lower—71% of those who installed a firewall.	And the number of installed software protection tools against Sandworm hackers (<i>хакеры Sandworm</i> in Russian) is lower—71% of those who installed a firewall.

Table 3: Augmentation examples for VIRUS

Original	Modified
Replacement	
Almost 30% are seriously concerned about this issue, another 25% believe that the danger of spyware is exaggerated, and more than 15% do not consider this type of threat to be a problem at all.	Almost 30% are seriously concerned about this issue, another 25% believe that the danger of Remcos is exaggerated, and more than 15% do not consider this type of threat to be a problem at all.
Insertion	
The malware described above is unique and can create big problems for both an individual and the whole company.	The Locker malware (<i>Зловред Locker</i> in Russian) described above is unique and can create big problems for both an individual and the whole company.

The suggested augmentation includes two subtypes: inner and outer. The inner augmentation involves extracting sentences that contain relevant descriptors within the existing training data. If a sentence meet augmentation restrictions, then the descriptor is replaced with a name or a name is added after the descriptor with equal probability. In both cases, we require that the descriptor must not be followed by a labeled named entity and it must not be preceded by words that agree with the

descriptor in gender, number or case, such as adjectives, participles, ordinal numbers, and others.

For the outer augmentation, we look for sentences with relevant descriptors in a collection of unannotated cybersecurity texts. There also must not be any evident named entities (words starting with a capital letter) in a window of certain width around the descriptor. As for this purpose an unannotated collection is used, we do not know the classes of potential named entities, thus we have to exclude sentences with such entities. Besides, we also require the absence of adjectives before the descriptor. The selected sentences also undergo the procedure of inserting a name after a descriptor or replacing the descriptor with a name with equal probability.

The augmentation has been implemented for two types of named entities: malicious software (VIRUS label) and hackers (HACKER label). 24 virus descriptors and 6 hacker descriptors were used. By means of inner augmentation, 262 additional annotated sentences for viruses and 165 annotated sentences for hackers were created.

The outer augmentation can be of an unlimited size. In this paper we study how the size of the outer augmentation affects the NER quality.

Inserted named entities are obtained in the following way. We took a large cybersecurity text collection and used it to extract names and sequences of names that follow target descriptors. We created the frequency list of extracted names and chose those names for which frequency was higher than a certain threshold (5). Then we excluded the names that appeared in the annotated training collection and belonged to classes that are different from the target class. The rest of the names were randomly used for insertion into the augmented sentences.

5. Experiments

We compare several variants of the BERT model on the NER task for information security domain. In addition, the results of using augmentation of the labeled data are investigated.

The CRF method was chosen as the baseline model for comparison, since in previous experiments with the Sec_col collection, this method showed better results than several variants of neural networks that are usually used for the NER task (BiLSTM with character embeddings) [19]. The CRF model utilizes the following features: token embeddings, lemma, part of speech, vocabularies of names and descriptors, word clusters based on their distributional representation, all these features in window 2 from the current token, tag of the previous word. The detailed description of CRF features, vocabularies, and implementation is given in [19]. Also, for comparison, the LSTM-CRF model based on Flair³ realization was added. The LSTM-CRF model used fasttext embeddings⁴ and the first capital letter feature for training.

Table 4 shows the classification results for four models for all labels used, as well as the averaged macro and micro F-measures. It can be seen that the use of the

³ <https://github.com/flairNLP/flair>

⁴ araneum_none_fasttextcbow_300 from <https://rusvectors.org/ru/models/>

multilingual-bert-base (BERT in the table) gives better results than the CRF model for all types of named entities. The use of the pretrained models on the Russian data (RuBERT) and information security texts (RuCyBERT) gives a significant improvement over previous models. The LSTM-CRF model with the described features showed weak results, therefore, did not participate in further experiments.

Table 4: Results of basic models

	LSTM-CRF	CRF	BERT	RuBERT	RuCyBERT
DEVICE	13.92	31.78	34.04	43.13	46.77
EVENT	28.79	42.70	60.38	64.49	67.86
HACKER	5.70	26.58	42.69	52.43	61.03
LOC	83.10	82.30	90.00	91.28	90.01
ORG	62.82	68.15	76.10	78.95	78.58
PER	58.71	67.10	80.99	84.32	84.56
PROGRAM	44.22	62.15	63.15	64.77	66.57
TECH	47.14	60.65	67.08	67.60	69.24
VIRUS	14.39	40.90	40.21	46.92	54.72
F-micro	53.12	63.95	69.37	71.61	72.74
F-macro	39.87	53.59	61.63	65.99	68.82
F-macro std	2.63	—	1.52	0.93	0.86

Since models based on neural networks due to random initialization can give slightly different results from run to run, the results in the tables for all BERT models are given as averaging of four runs. The last row of **Table 4** indicates (F-macro std) the standard deviation of the results from the mean. It can be seen that the better the model fits the data, the better the results are, and the standard deviation decreases.

The following tables show the use of the proposed data augmentation approach to extract two types of named entities HACKER and VIRUS with inner and outer augmentations. For the outer augmentation, options for adding 100, 200, 400, 600 augmented sentences for each entity types (HACKER and VIRUS) were considered. However, the outer augmentation of 600 sentences gave a stable decrease in the results for all models, and therefore these results are not given in the tables. The “mean F1” column shows the averaging of the values of the F1 measure over all types of entities. The best achieved results are in bold. The results improving the basic results (without augmentation) are underlined.

Table 5 presents the results of applying augmentation to the CRF model. All types of the augmentation improved the results of extracting target entities. The best augmentation was inner augmentation, which gave an increase in the average quality of extracting named entities HACKER and VIRUS named entities by 10 percentage points (almost a third). Macro F1 measure for all types of entities (last column) is also significantly improved.

Table 6 shows the performance of the augmentation procedure for the multilingual BERT base. The table shows how unstable the multilingual BERT model behaves, demonstrating a very high standard deviation on the two types of entities that interest us. Any variant of augmentation reduces the standard deviation, which, however,

remains quite high (column F1 std). Two models of outer augmentation increase the quality of extraction of target entities while significantly reducing the standard deviation compared to the original model.

Table 7 presents the results of the RuBERT model, trained on the Russian data. The results are significantly higher than for the previous model, the standard deviation is lower. And in this model, the augmentation in all cases reduces the standard deviation of F measures for target and all types of entities. The results on the target entities increased with outer augmentation of 200 sentences for both entities. Also, for some reason, the outer augmentation only with viruses positively influenced the extraction of both of them (100 and 200 sentences). The study of this phenomenon is planned to continue.

Table 5: CRF with augmentation

	HACKER_VIRUS			macro
	P	R	F1	F1
base (no augmentation)	66.31	24.21	33.73	53.59
inner	42.08	47.31	43.58	57.39
outer 100	47.36	32.63	38.20	54.98
outer 200	48.12	35.36	40.21	55.18
outer 400	40.58	35.27	36.97	54.21

Table 6: BERT with augmentation

	HACKER_VIRUS				macro	
	P	R	F1	F1 std	F1	F1 std
base (no augmentation)	46.43	38.14	41.45	7.23	61.63	1.52
inner	36.81	45.44	39.92	3.53	61.26	0.86
outer 100	39.13	44.96	41.04	2.18	62.02	0.55
outer 200	39.32	48.24	42.51	4.33	62.21	0.74
outer 400	40.23	45.97	42.53	4.59	62.12	1.08

Table 7: RuBERT with augmentation

	HACKER_VIRUS				macro	
	P	R	F1	F1 std	F1	F1 std
base (no augmentation)	53.65	47.38	49.67	4.65	65.99	0.93
inner	45.01	55.74	48.87	3.48	65.92	0.68
outer 100	47.46	53.29	49.38	3.10	65.88	0.79
outer 200	47.83	55.34	50.71	2.96	66.24	0.59
outer 400	45.57	53.45	48.46	2.36	65.77	0.67
outer viruses 100	57.14	51.67	53.79	3.05	66.85	0.64
outer viruses 200	55.33	52.55	53.34	3.90	66.68	0.77

Table 8 presents the results of the RuCyBERT model, trained on the information-security texts. The basic quality of this model is much higher, and there is no improvement from the augmentation. The augmentation on average reduces the standard deviation of F-measure, which leads to the fact that the performance of models with augmentation and the basic model is comparable.

It can be also seen from **Tables 5–8** that in almost all experiments the proposed augmentation significantly increases recall, but decreases precision.

Table 8: RuCyBERT augmentation

	HACKER_VIRUS				macro	
	P	R	F1	F1 std	F1	F1 std
base (no augmentation)	61.33	55.89	57.87	3.75	68.82	0.86
inner	52.51	62.57	56.03	2.54	68.61	<u>0.53</u>
outer 100	50.78	59.69	53.79	2.36	67.78	0.43
outer 200	52.82	59.61	54.82	3.94	68.06	<u>0.74</u>
outer 400	52.42	<u>61.31</u>	55.64	2.16	67.93	<u>0.71</u>

6. Conclusion

In this paper we present the results of applying BERT to named entity recognition for cybersecurity Russian texts. It is shown that the multilingual model performs better than the CRF model, which uses a substantial number of token features. Further tuning of the model (first on the Russian data and then on the cybersecurity collection) has significantly improved the NER quality. The highest macro F-score shown by BERT model (RuCyBERT) is 15 percent higher than macro F-score of the CRF model.

For each model, we have also presented a new form of augmentation of labeled data for the NER task, that is adding names after or instead of a descriptor of a certain type. A considerable improvement is recorded for relatively weak CRF and multilingual BERT models. For the fine-tuned models, the quality has barely grown. Nevertheless, if in some cases it is impossible to fine-tune BERT on a specialized collection, the presented augmentation for named entities could be of great use while extracting named entities of non-standard types. Besides, the proposed augmentation approach can be used in automated creation of a domain-specific NER annotated dataset from general datasets such as CONLL-2003, or Collection3. The described `Sec_col` collection and the trained RuCyBERT model can be obtained from the repository⁵.

References

1. Bahdanau, D., Cho, K., Bengio, Y. (2014), Neural machine translation by jointly learning to align and translate, available at <https://arxiv.org/abs/1409.0473>.

⁵ <https://github.com/LAIR-RCC/InfSecurityRussianNLP>

2. *Bocharov, V., Starostin, A., Alexeeva, S., Bodrova, A., Chunchunkov, A., Dzhumaev, S., Efimenko, I., Granovsky, D., Khoroshevsky, V., Krylova, I., Nikolaeva, M., Smurov, I., Toldova, S.* (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 2016” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2016”], Moscow, pp. 702–720.
3. *Bridges, R., Jones, C., Iannacone, M., Testa, K., Goodall, J.* (2013), Automatic labeling for entity extraction in cyber security, available at <https://arxiv.org/abs/1308.4941>.
4. *DeepPavlov* documentation, <http://docs.deeppavlov.ai/en/master/>. Last accessed 25 Dec 2019
5. *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.* (2018), Bert: Pre-training of deep bidirectional transformers for language understanding, available at <https://arxiv.org/abs/1810.04805>.
6. *Gasmi, H., Bouras, A., Laval, J.* (2018), LSTM Recurrent Neural Networks for Cybersecurity Named Entity Recognition, ICSEA-2018, Vol 11.
7. *Gareev, R., Tkachenko, M., Solovyev, V., Simanovsky, A., Ivanov, V.* (2013), Introducing baselines for Russian named entity recognition, International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Heidelberg, pp. 329–342.
8. *Howard, J., Ruder, S.* (2018), Universal language model fine-tuning for text classification, available at <https://arxiv.org/abs/1801.06146>.
9. *Joshi, A., Lal, R., Finin, T., Joshi, A.* (2013), Extracting cybersecurity related linked data from text, 2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA , pp. 252–259.
10. *Kobayashi, S.* (2018), Contextual augmentation: Data augmentation by words with paradigmatic relations, 2018 Conference of the North American Chapter of the Assoc. for Computational Linguistics, NAACL-2018, New Orleans, pp. 452–457.
11. *Kuratov, Y., Arkhipov, M.* (2019), Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language, available at <https://arxiv.org/abs/1905.07213>.
12. *Lafferty J., McCallum A., Pereira F.* (2001), Conditional random fields: Probabilistic: models for segmenting and labeling sequence data, International Conference on Machine Learning ICML-2001.
13. *Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.* (2016), Neural architectures for named entity recognition, available at <https://arxiv.org/abs/1603.01360>.
14. *Mikolov, T., Chen, K., Corrado, G., & Dean, J.* (2013), Efficient estimation of word representations in vector space, available at <https://arxiv.org/abs/1301.3781>.
15. *Fellbaum, Ch.* (1998), WordNet: An Electronic Lexical Database, MIT-press
16. *Mozharova, V., Loukachevitch, N.* (2016), Combining knowledge and CRF-based approach to named entity recognition in Russian, International Conference on Analysis of Images, Social Networks and Texts, Springer, Cham, pp. 185–195.

17. *Piskorski, J., Laskova, L., Marcinczuk M., Pivovarova, L., Priban P., Steinberger, J., Yangarberger, R.* (2019), The Second Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages, 7th Workshop on Balto-Slavic Natural Language Processing BSNLP-2019, Florence, pp. 63–74.
18. *Sang, E., Meulder, F.* (2003), Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, Proc. of the 7th conference on Natural language learning at HLT-NAACL 2003, Vol 4., pp. 142–147.
19. *Sirotnina, A., Loukachevitch, N.* (2019), Named Entity Recognition in Information Security Domain for Russian, Proceedings of RANLP-2019, Varna, pp. 1115–1122.
20. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I.* (2017), Attention is all you need, Advances in neural information processing systems, pp. 5998–6008.
21. *Vlasova, N. A., Suleymanova, E. A., and Trofimov, I. V.* (2014), The message about Russian collection for named entity recognition task [Soobshchenie o russkoyazychnoj kollekcii dlya zadachi izvlecheniya lichnyh imen iz tekstov], Proceedings of computational and cognitive linguistics TEL, Kazan, pp. 36–40.
22. *Wei, J. W., Zou, K.* (2019), Eda: Easy data augmentation techniques for boosting performance on text classification tasks, Conference on Empirical Methods in Natural Language Processing EMNLP-2019, Hong Kong, pp. 6381–6387.
23. *Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J.* (2016), Google’s neural machine translation system: Bridging the gap between human and machine translation, available at <https://arxiv.org/abs/1609.08144>.
24. *Yang Wang, W., Yang, D.* (2015), That’s so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets, 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, pp. 2557–2563.

COMBINED APPROACH TO HYPERNYM DETECTION FOR THESAURUS ENRICHMENT

Tikhomirov M. M. (tikhomirov.mm@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

Loukachevitch N. V. (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russia,
Kazan Federal University, Kazan, Russia

Parkhomenko E. A. (parkat13@yandex.ru)

Lomonosov Moscow State University, Moscow, Russia

This paper describes a combined approach to hypernym detection task. The approach combines the following techniques: distribution semantics, rule-based patterns, and modern neural networks (BERT). An important feature of our solution is that hypernyms are extracted only from a single text collection provided by the organizers. The described approach obtained the fourth result on the private nouns track. It was found out that the use of the rule-based patterns can significantly improve the results. Also, using the BERT model as an additional factor always helps to improve the performance.

Key words: hypernym extraction, thesaurus, embeddings, patterns, BERT

DOI: 10.28995/2075-7182-2020-19-736-746

КОМБИНИРОВАННЫЙ ПОДХОД К ИЗВЛЕЧЕНИЮ ГИПЕРОНИМОВ ДЛЯ ОБОГАЩЕНИЯ ТЕЗАУРУСА

Тихомиров М. М. (tikhomirov.mm@gmail.com)

МГУ имени М. В. Ломоносова, Москва, Россия

Лукашевич Н. В. (louk_nat@mail.ru)

МГУ имени М. В. Ломоносова, Москва, Россия,
Казанский Федеральный Университет, Казань, Россия

Пархоменко Е. А. (parkat13@yandex.ru)

МГУ имени М. В. Ломоносова, Москва, Россия

В данной статье описывается комбинированный подход к решению задачи извлечения гиперонимов. В работе используются: методы дистрибутивной семантики, шаблоны на основе правил и современные нейронные сети (BERT). Важной особенностью описанного решения является то, что извлечение гиперонимов происходит только на основе коллекции текстов, предоставленной организаторами. Описанный подход получил четвертый результат в задаче предсказания гиперонимов существительных. Было показано, что использование шаблонов на основе правил может значительно улучшить результаты. Кроме того, использование BERT в качестве дополнительного фактора всегда улучшает результаты.

Ключевые слова: извлечение гиперонимов, тезаурус, векторные представления, шаблоны, BERT

1. Introduction

Knowing relationships between words such as a hypernym (more general word) or hyponym (more specific word) can be useful in various tasks of natural language processing such as question answering, sentiment analysis, textual entailment, etc. Lexical relationships can be found in lexical-semantic resources such as WordNet [13], BabelNet [14] and others. However, development and maintenance of large semantic resources is a very difficult, time-consuming activity. Besides, such resources are never comprehensive: some relations significant for a specific domain or a task can be absent in the used resource. Numerous studies have been devoted to the problem of hypernym detection from text collections or definitions in conventional vocabularies [7], [22], [6], [16]. The researchers utilized such methods as the use of linguistic patterns, unsupervised distributional approaches, machine learning, and combinations of several methods.

For evaluation, the task of hypernym detection can be formulated as a two-class classification of relations. Some approaches attempted to evaluate the whole structure of the taxonomy created on the basis of extracted hypenymy [3], [19]. The organizers of the SemEval 2018 hypernym detection task [4] considered the task of hypernym detection as a ranking problem when correct hypernyms had to be located closer to the top of the hypernym candidate list.

In 2020, the first evaluation for extracting hypernyms for Russian RUSSE-2020 Taxonomy Enrichment was organized [15]. The task was to provide the most appropriate hypernym synsets from the published version of RuWordNet [11] for new words. Previously, a new, extended version of RuWordNet had been prepared manually but not published. It was proposed to use extended RuWordNet as a gold standard in evaluation of hypernym extraction.

The organizers provided a news collection to participants ensuring that the practically all novel words under consideration occurred in the collection with frequency more than 50, but no restrictions were placed on the use of additional resources. Solving such a task, the participants could consider two possible approaches in utilizing data. The first approach is to use various available resources (text collections, vocabulary definitions, pre-trained embeddings) to detect hypernym synsets for a give word. The second approach can be to use mainly the provided corpus trying to extract all possible relations

from it. In the current paper, we consider the second approach: in our experiments we used only the given text collection. We supposed the following practical scenario: for tuning a general resource to a specific domain, only a specialized text collection can be given, other types of available data can be too general or irrelevant to the domain.

We consider a combined approach to hypernym extraction from a given text collection including distributional representation of words and phrases, patterns, accounting named entities, which have to be distinguished from ordinary words, and the BERT model [5] trained for hypernym prediction.

2. Related Work

Approaches for hypernym extraction from texts can be subdivided into: pattern-based (linguistic) approaches, unsupervised distributional approaches, machine learning methods, and combined approaches.

Linguistic approaches are based on obtaining relations between words revealing their joint occurrences in a certain set of patterns [7]. The use of these methods presupposes that if some specified context is detected between entities X and Y, then, there is a given relationship between these entities. For the taxonomic relationship, such a context can be the context “X is a kind of Y”. The pattern-based approach is highly accurate and simple, and can be applied to various types of relationships. Later pattern approaches exploited syntactic structures [21], [18]. In [20], the authors trained the LSTM neural network to generate embeddings for syntactic paths for hypernym extraction. In Russian, Sabirova and Lukanin [17] described patterns for hypernym extraction from Russian texts.

Distributional approaches are based on vector representations of words. Several methods were used to represent term relationships as a combination of each term’s embedding vector: concatenation [1], difference [16], and dot-product. The best results using distributional methods were achieved with supervised approaches. Supervised projection learning models are trained to project the embedding of the target word such that its projection is close to that of its hypernym [6], [23]. Fu et al. [6] proposed also to learn multiple projection matrices representing different kinds of hypernymy relationships. Using this approach, Yamane et al. [25] presented a model that jointly learns word clusters and corresponding projection matrices.

However, Levy et al. [9] revealed that the results achieved by supervised distributional models are mainly due to so-called “lexical memorization phenomenon”, that is the models learn that y is a prototypical hypernym regardless of x. For example, if a classifier obtains many positive examples with the word y = animal, it may learn that anything that appears with y = animal should generate the positive answer. Based on these results, Camacho-Collados et al. [4] proposed to change the evaluation scheme of hypernym extraction from two-class classification to the ranking scheme.

The advantages of different methods are combined in incremental taxonomy generation [22], [26], in which all of the above characteristics (occurrence in templates, similarities in different types of contexts (linear and syntactic, local and global), joint occurrence) serve as features for a classifier that decides to join the next word to the created taxonomy.

The CRIM system, the winner of the SemEval 2018 Hypernym Discovery task, exploited the combined approach to hypernym recognition [2]. Their approach comprised hypernym patterns, co-hyponym patterns, and supervised projection learning. Held and Habash [8] practically repeated and partially improved the results of the CRIM system using a significantly simpler combined method based on hypernym patterns and distributional similarity of a term from the test collection to the most similar term from the training collection.

3. Data and Task

The task of RUSSE-2020 taxonomy enrichment is to find the closest synset from the published RuWordNet version for a novel word. The gold standard data are taken from prepared but not currently published extended RuWordNet version. Only the closest hypernym from RuWordNet for a given word or its direct hypernyms are considered as correct answers.

The task comprises nouns and verbs. There were 762 nouns and 175 verbs in the public set for developing and tuning algorithms, and 1525 nouns and 350 verbs in the private (test) data. The evaluation was organized in the CodaLab evaluation system¹. Besides, the organizers provided the training dataset constructed on published RuWordNet, a news collection of 2017 (further, News2017 corpus) gathered from more than 1000 news sources, in which new words under evaluation mainly occurred at least 50 times, and some other resources.

This news text collection consists of 8 million articles, a total of 2.2 billion tokens. Statistics of the target words can be seen in Table 1. It can be seen that the corpus provides more than 50 occurrence for almost all words in the datasets. Therefore it is possible to attempt to obtain hypernyms for words in the current task from this specific corpus without using additional sources of candidates.

Table 1: Target words statistics

	median	max	freq < 50
public nouns	188.5	133,683	10
public verbs	128.5	16,657	7
private nouns	183.0	140,173	28
private verbs	140.5	45,807	14

Participating system have to generate a list of the most relevant hypernym synsets for novel words. The top 10 synsets are considered as answers. The main measure is mean average precision MAP, which is equal to 1 as maximum when all the correct answers are at the top of the candidate list. The second used measure is MRR, which is calculated on the basis of the first correct answer.

¹ <https://competitions.codalab.org/competitions/22168>

4. Method for Hypernym Extraction

Our method of hypernym extraction is intended to extract maximum information from a given text collection, in the current case from the News2017 text collection provided by the organizers. The method comprises the following components:

- distributional (vector) representation of words and phrases under analysis,
- linguistic patterns for hypernyms and co-hyponyms,
- special processing of named entities to remove their contexts from consideration,
- application of BERT in a supervised manner to approve hypernym candidates.

4.1. Word embeddings

In this study word embeddings were trained only on the News2017 corpus. As a training method we used not word2vec, but the traditional approach in distributional semantics, which includes the following steps:

1. The matrix of co-occurrence frequencies of words in the corpus is calculated,
2. Scores of word co-occurrences are recalculated using the positive point-wise mutual information (PPMI) measure [10],
3. SVD method over PPMI matrix is applied, which allows reducing the dimension of the matrix from the vocabulary size to a chosen smaller value,
4. When processing the corpus, phrases from the thesaurus were merged into united tokens, for example **rice_grain**. The models were calculated with 600 vector size and different window sizes: 1, 3, 5.

The PPMI + SVD approach demonstrated comparable quality with word2vec in a number of experiments [10] and showed better results in our internal studies.

For all target words, top 100 word candidates from each model were considered, based on cosine similarity. Also, words must be present in the thesaurus. Then, the list of candidate synsets was formed by extracting hypernyms for synsets and hypernyms of hypernyms (with penalty weight) corresponding to candidate words. For each candidate synset, the average cosine similarity of the initial words (*cos_sim_list*) and the number of times the algorithm returned this synset (*count*) were calculated; the candidates were ranked by the following formula:

$$base_{score} = mean(cos_sim_list) \cdot \log_2(1 + count) \cdot \alpha, \quad (1)$$

where α is equal to 0.3, if estimated synset is hypernym of hypernym.

4.2. Excluding Named Entity Contexts

It was found in previous works that named entities can include general words and in such a way to distort the contexts of ordinary words [2], [12]. For example, Loukachevitch [12] found that Russian word *mistral*, which means “a strong, cold, north-westerly wind. in the northern Mediterranean” and is linked to hypernym *wind* in Ru-WordNet, but in current news articles in Russian, this word mainly means the class of French helicopter carriers. We assumed that the RUSSE evaluation dataset contains mainly ordinary words and decided to train word embeddings excluding named entity

contexts from consideration. Examples of entries in the RUSSE dataset that coincide with named entities are as follows: *trepak* (dance vs. family name), *chub* (strand of hair vs. family name), *ryabinovka* (alcoholic beverage vs. names of villages), and others.

Currently, we applied the following simple preprocessing procedure: if a word occurrence is capitalized and it is not the first word in a sentence then such an occurrence obtains a special prefix, which means that the occurrence and its contexts are excluded from calculating embeddings.

4.3. Use of Patterns

In addition to the basic method, two kinds of patterns were extracted and used. All patterns were applied to the pairs of the target word and all similar words extracted by a distributional model. The appearance of candidates in the patterns increases the similarity weight of the candidate words to the target word, because pattern matching is an additional evidence of semantic similarity. For example, it is examined whether pattern matching exist for the target word *peony* and the candidate word *rose*. If it exists, then it is evidence, that *rose* and *peony* can be synonyms or co-hyponyms.

All patterns were automatically extracted using regular expressions on the News2017 corpus. Two types of patterns were considered:

1. co-hyponym patterns, whose successful matching leads to an increase in the weight of hypernym synsets for words from top most similar words,
2. synonym-hypernym patterns, whose successful matching leads to both: an increase in the weight of hypernym synset and an additional inclusion in the candidate list of a direct synset of a candidate word, but not just its hypernym as in the base model.

The following are sample patterns (splitted by ; character), where X is a target word, Y is a candidate word, and W is any word.

- Examples of co-hyponym patterns: X , W , Y; Y, W, X; Y , X; X, Y; Y and X; X and Y; Y or X; X or Y;
- Examples of synonym—hypernym patterns: X (Y; Y (X; Y—X; X—Y; X—is Y; Y—is X; X , W and another Y; X and W—is type of Y.

Examples of co-hyponyms extracted with the described patterns for word *ice-cream* (with frequencies in the corpus) are as follows: *chocolate* (145), *sweets* (106), *candy* (93), *cookie* (69), *yogurt* (61), .. Examples of pattern-based hypernyms for *ice-cream* include the following words: *delicacy* (70), *dessert* (21), *cake* (13), *sweets* (10), ...

When applying a co-hyponym pattern, the frequency of being in one sentence for target and candidate words, is calculated. The modified formula is as follows:

$$upd_pattern_score = base_score \cdot (1 + shp_hit) \cdot (1 + \frac{2 \cdot chp_count}{os_count + 2}) \quad (2)$$

where *shp_hit* is equal to 1 if the synonym-hypernym pattern was matched, *chp_count* is a count of times when co-hyponym pattern was matched, and *os_count* is a count when words were in one sentence.

In addition to modifying the formula, direct synsets (synsets of the most similar words) were also added as candidates, when the synonym-hypernym pattern was matched.

4.4. Use of BERT to Assess Hypernym Candidate

A neural network architecture such as BERT [5] can be used for the hypernym prediction task [27]. BERT is a transformer encoder [26], and one of its key features is that it is trained on a large amount of unlabeled data. After this procedure, the model shows strong results on a wide range of specific tasks.

For the hypernym prediction task, the following approach was implemented:

1. The binary classification problem was considered;
2. The BERT input was a pair of words with special characters: [CLS] word1 [SEP] word2 [SEP]; word2 can be a multi-word expression. If word2 is a hypernym of word1, then the label is 1, and 0 otherwise;
3. The training data was created from RuWordNet. For each positive example, three negative examples were added. The negative examples were uniformly sampled from: random synsets, hypernyms of hypernyms, hyponyms and hyponyms of hypernyms of word1.

The model was fine-tuned on 5 epochs with $2e-4$ learning rate for the classification layer and $2e-5$ for the BERT layers. The training dataset consisted of 1.5 million examples.

Further, for each candidate synset, its probability of being a hypernym and the maximum probability of its hyponyms to be a hypernym to the target word were calculated. When calculating the probability of synset, the probabilities of all synset entries were calculated, and then averaged. The resulting BERT probability is:

$$bert_prob = 0.6 \cdot syn_bert_prob + 0.4 \cdot max_hyp_syn_bert_prob \quad (3)$$

And then, the final modified formula for ranking is as follows:

$$synset_score = upd_pattern_score \cdot (1 + bert_prob) \quad (4)$$

For example, for the word *agnostic* there is a hypernym synset with name *follower*. In this case, BERT probability will be 0.99, model predicts such a probability directly for a given synset.

5. Results

In this competition, for each target word, it was necessary to provide a list of 10 candidate hypernym synsets, ranked from more probable to less probable. To assess the quality of the models, the mean average precision (MAP) and mean reciprocal rank (MRR) metrics were used.

Table 2: Nouns track results

	public		private	
	MAP	MRR	MAP	MRR
base	0.446	0.478	0.440	0.474
base-ne	0.449	0.482	0.444	0.478
base-ne, co-hyponym patterns	0.473	0.509	0.467	0.502
base-ne, hypernym patterns	0.467	0.505	0.456	0.494
base-ne, both patterns	0.482	0.522	0.481	0.520
base-ne, bert	0.471	0.507	0.455	0.490
base-ne, bert, co-hyponym patterns	0.494	0.533	0.475	0.509
base-ne, bert, hypernym patterns	0.484	0.524	0.470	0.508
base-ne, bert, both patterns	0.509	0.550	0.493	0.531

Table 3: Verbs track results

	public		private	
	MAP	MRR	MAP	MRR
base	0.254	0.295	0.255	0.291
base, co-hyponym patterns	0.285	0.327	0.246	0.286
base, hypernym patterns	0.260	0.303	0.254	0.290
base, both patterns	0.288	0.332	0.244	0.284
base, bert	0.277	0.322	0.265	0.305
base, bert, co-hyponym patterns	0.314	0.358	0.259	0.302
base, bert, hypernym patterns	0.288	0.334	0.264	0.304
base, bert, both patterns	0.324	0.371	0.254	0.296

For verbs and nouns, slightly different parameters were applied. For nouns, embeddings with only window 1 were considered; for verbs, embeddings with windows 1, 3, and 5 were combined. In case when named entity contexts were excluded, the results of the models with and without this preprocessing were also combined by averaging of similarities. The configuration called **base** refers to the basic approach described in [section 4.1](#), and the **base-ne** configuration refers to **base** configuration with names excluding, described in [section 4.2](#). [Table 2](#) describes the results of the proposed methods for nouns, and [Table 3](#) contains the results for verbs.

The following conclusions can be drawn from the results:

- Exclusion of named entity contexts slightly improves the results;
- In all cases, except for private verbs, matching patterns significantly improves the results. The problem with private verbs will be investigated later;
- Using BERT in the described way always improves the results.

The described approach obtained the fourth result on the noun private set without any additional vocabulary definitions, used by the first two approaches (according to the provided descriptions) and is very close to the results of the third place

approach. In practice, in specific domains external vocabularies can be irrelevant to the domain and provided text collection. The use of the combined method improved the hypernym predictions for nouns by more than 10% in the public and private sets if compared to the basic model.

For verbs, the improvement on the public set was also significant (more than 25%) but in the private dataset the improvement was not reproduced.

6. Conclusion

The paper describes a combined approach to the hypernym detection task. The approach combines the following techniques: distributional semantics, rule-based patterns, and modern neural. Hypernyms are extracted only from the text collection provided by the organizers. This reflects the real situation when it is necessary to expand the existing thesaurus to a new specific domain.

The described approach obtained the fourth result on the private nouns track, with 0.493 MAP and 0.531 MRR. It was found that the use of the described patterns can significantly improve the results. Also, using the BERT model as an additional factor always helps to improve the results. The exclusion of name contexts for target words also helps to improve the results of the system, by improving the results on specific words.

As further research, it is necessary to more deeply explore the possibilities of using BERT for this task. In addition, it is planned to investigate the unstable behavior of patterns on verbs. Also, the behavior of the system strongly depends on the algorithm for obtaining candidates, so it is planned to explore various ways of doing this.

7. Acknowledgements

The research is carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University. The participation of M. Tikhomirov in the reported study was funded by RFBR, project number 19-37-90119. The work of N. Loukachevitch concerning formulation of main principles of work with a text collection, accounting named entities, and patterns is funded by Russian Science Foundation, research project № 19-71-10056, financed through Kazan Federal University.

References

1. *Baroni M., Lenci A.* (2011), How we BLESSed distributional semantic evaluation, Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, Association for Computational Linguistics, Edinburgh, pp. 1–10.
2. *Bernier-Colborne G., Barriere C.* (2018), CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery, Proceedings of The 12th International Workshop on Semantic Evaluation, New Orleans, pp. 725–731.

3. *Bordea G., Lefever E., Buitelaar P.* (2016) Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, pp. 1081–1091.
4. *Camacho-Collados J., Delli Bovi C., Espinosa-Anke L., Oramas S., Pasini T., Santus E., ... & Saggion H.* (2018). SemEval-2018 task 9: Hypernym discovery, Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, pp. 712–724.
5. *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.* (2018), Bert: Pre-training of deep bidirectional transformers for language understanding, available at <https://arxiv.org/abs/1810.04805>.
6. *Fu, R., Guo, J., Qin, B., Che, W., Wang, H., & Liu, T.* (2014), Learning semantic hierarchies via word embeddings, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1199–1209.
7. *Hearst M.* (1992), Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the 14th conference on Computational linguistics-Volume 2, pp. 539–545.
8. *Held W., Habash N.* (2019), The Effectiveness of Simple Hybrid Systems for Hypernym Discovery, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3362–3367.
9. *Levy, O., Remus, S., Biemann, C., & Dagan, I.* (2015), Do supervised distributional methods really learn lexical inference relations?, Proceedings of NAACL-HLT 2015, pp. 970–976.
10. *Levy, O., Goldberg, Y., Dagan, I.* (2015), Improving distributional similarity with lessons learned from word embeddings, Transactions of the Association for Computational Linguistics, Vol. 3, pp. 211–225.
11. *Loukachevitch, N. V., Dobrov, B. V.* (2014), RuThes linguistic ontology vs. Russian wordnets, Proceedings of the Seventh Global Wordnet Conference, pp. 154–162.
12. *Loukachevitch N. V.* (2019), Corpus-based Check-up for Thesaurus, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5773–5779.
13. *Miller, G.* (1998), Nouns in WordNet, WordNet – An Electronic Lexical Database, Fellbaum, C (ed), The MIT Press, pp. 23–47
14. *Navigli R., Ponzetto S. P.* (2010), BabelNet: Building a very large multilingual semantic network, Proceedings of the 48th annual meeting of the association for computational linguistics, pp. 216–225.
15. *Nikishina I., Logacheva V., Panchenko A., Loukachevitch N.* (2020), RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian language, Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2020.
16. *Roller S., Erk K., Boleda G.* (2014), Inclusive yet selective: Supervised distributional hypernymy detection, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1025–1036.

17. *Sabirova K., Lukanin A.* (2014), Automatic Extraction of Hypernyms and Hyponyms from Russian Texts, Supplementary Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST 2014), pp. 35–40.
18. *Sang E. T. K., Hofmann K.* (2009), Lexical patterns or dependency patterns: which is better for hypernym extraction? ,Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pp. 174–182.
19. *Sarkar, R., McCrae, J.P., Buitelaar, P.* (2018), A supervised approach to taxonomy extraction using word embeddings, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)
20. *Shwartz V., Dagan I.* (2016), Path-based vs. distributional information in recognizing lexical semantic relations, COLING, pp. 24–29.
21. *Snow R., Jurafsky D., Ng A. Y.* (2004), Learning syntactic patterns for automatic hypernym discovery, Proceedings of Advanced in Neural Information Processing systems (NIPS), pp. 1297–1304.
22. *Snow R., Jurafsky D., Ng A. Y.* (2006), Semantic taxonomy induction from heterogeneous evidence, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 801–808.
23. *Ustalov, D., Arefyev, N., Biemann, C., & Panchenko, A.* (2017), Negative Sampling Improves Hypernymy Extraction Based on Projection Learning, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, pp. 543–550.
24. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I.* (2017), Attention is all you need, Advances in neural information processing systems, pp. 5998–6008.
25. *Yamane, J., Takatani, T., Yamada, H., Miwa, M., & Sasaki, Y.* (2016), Distributional hypernym generation by jointly learning clusters and projections, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1871–1879.
26. *Yang H., Callan J.* (2009), A metric-based framework for automatic taxonomy induction, 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: Volume 1-Volume 1, Association for Computational Linguistics, pp. 271–279.
27. *Yao L., Mao C., Luo Y.* (2019), KG-BERT: BERT for Knowledge Graph Completion, available at <https://arxiv.org/pdf/1909.03193.pdf>.

DISCOURSE FEATURES OF BLOGS IN SUBCORPUS OF RUSSIAN RU-RSTREEBANK¹

Toldova S. (toldova@yandex.ru)

NRU Higher School of Economics, Moscow, Russia;

Davydova T. (tdadidik@gmail.com),

Kobozeva M. (kobozeva@isa.ru),

Pisarevskaya D. (dinabpr@gmail.com)

FRC CSC RAS, Moscow, Russia

The paper presents a corpus study of the discourse features in the corpus of blogs. It is based on the data of Ru-RSTreebank annotated within the framework of the Rhetorical Structure theory [Mann, Thompson 1988]. The Ru-RSTreebank represents genres of news and popular science, scientific papers, and blogs texts. Blog subcorpus contains such topics as travelling, cosmetics, sports and health, psychology, IT and tech and some others. Blogs texts constitute a specific genre as they combine properties of written and spoken discourse. The purpose of the paper is to investigate discourse features of blogs in comparison with other genres. We analyze the variation in rhetoric relations distribution among genres, and single out the differences in discourse connectives usage. Furthermore, we check the distribution of other discourse features reported in different studies for spoken discourse and for social media in the Ru-RSTreebank blogs subcorpus. The general frequency analysis and the experiments on RandomForest classifier application to genre recognition have shown that the most important rhetoric relations specific to blogs are Evaluation and Contrast, that there is a tendency to use shorter discourse units and not to express the discourse relations overtly via subordinative conjunctions.

Keywords: discourse analysis, rhetorical structure theory, blogs annotation, corpus linguistics, corpus annotation

DOI: 10.28995/2075-7182-2020-19-747-761

ДИСКУРСИВНЫЕ ОСОБЕННОСТИ БЛОГОВ НА МАТЕРИАЛЕ РУССКОЯЗЫЧНОГО ДИСКУРСИВНОГО КОРПУСА RU-RSTREEBANK

¹ The study was funded by Russian Foundation for Basic Research according to the research project № 17-29-07033.

1. Introduction

The research on discourse coherence and on how this coherence is achieved has experienced a revival in the last few years. The core questions are how the discourse relations are established and what are signals of these relations. In order to answer these questions, corpora supplied with discourse annotation have been constructed for English and many other languages [Matthiessen, Teruya 2015]; [da Cunha 2016]; [Iruskieta et al. 2015]; [Zeldes 2017]. Besides serving as a source for theoretical studies, these corpora are used as datasets for building discourse parsers.

The structure of texts from different registers (e.g. written vs. spoken texts) and genres varies in a number of parameters, such as typical discourse relations, sentence length, verb forms, conjunctions usage etc. [Chafe 1982]. This variation should be considered in discourse research and in building the applications. In recent decades, new genres have emerged, pertaining to online social media and blogs communication. Its discourse characteristics are widely discussed (e.g. [Simaki et al. 2018]; [Berger, Hennig 2015]; [Germasheva 2010] etc.). It is generally assumed that it combines properties of written and spoken modes and, besides, manifests its own features. Thus, the investigation of the discourse parameters of social media as compared to other genres and registers is of great importance.

The present work deals with the analysis of discourse features of blogs. The data is taken from Ru-RSTreebank corpora [Pisarevskaya et al. 2017]; [Chistova et al. 2019] annotated within the RST framework (Rhetorical Structure theory) [Mann,Thompson 1988]. The new release of this corpus in 2019 includes 104 texts of blogs. The main research questions are: (a) whether there is a significant difference in rhetoric relations distribution among genres; (b) what other characteristics of discourse units distinguish blogs from other genres; (c) whether there is a difference in discourse connectives usage among the three genres represented in Ru-RSTreebank; (d) what other discourse-oriented devices let blogs differ from written texts. Our aim is to check the claim, articulated in the literature (e.g. [Simaki et al. 2018]), that blogs carry some of the features of spoken discourse. One can find the correlates of specific types of features pertaining to spoken discourse, such as short subclausal discourse units (elaborations, parcellations etc.), markers of interaction or regulatory markers and some others. We enhanced the corpus with additional annotation for these features. In this work, we provide the corpus-based analysis of blogs with respect to spoken discourse features.

The paper is structured as follows. We start with a brief description of theoretical assumptions (2.1) and short overview of discussions, devoted to the written vs. spoken discourse opposition (2.2), and blogs discourse (2.3). Next, we describe the corpus data and its preparation for the analysis (3.1–3.3). After that, we turn to quantitative (4.1–4.3) and qualitative analysis of various discourse features (5.1–5.2).

2. Background

2.1. Discourse structure

Our study is based on the corpus analysis of Ru-RSTreebank (<https://rstreebank.ru/dataset>), the first discourse-annotated corpus for Russian [Pisarevskaya et al. 2017]. According to RST, the discourse is a hierarchical structure. A text can be successively split into spans (discourse units, DUs, up to elementary discourse units, EDUs) based on the rhetorical relations between them (e.g. Cause-Effect, Concession, Joint etc.). The relations between DUs are somehow parallel to the relations between clauses in complex sentences. They can be asymmetrical (cf. subordinate vs. main clause) or symmetrical ones. The ‘canonical’ EDUs usually describe events or states and, hence, syntactically, the typical EDUs are simple clauses [Kibrik, Podlesskaya eds. 2009]. However, there are different types of EDUs smaller than a clause (subclausal) in spoken discourse (ibid.).

2.2. Difference in register for spoken and written discourse

There is a claim in the literature that microblogs and blogs are similar to spoken conversations (e.g. [Scheffler et al. 2019], [Akhapkina 2014] etc.). According to [Chafe 1982] (see also [Kibrik, Podlesskaya eds. 2009]), spoken vs. written discourse are opposed with respect to two basic dimensions. The first one is ‘**integration/fragmentation**’. Written texts contain more complex sentences: nonfinite clauses (nominalizations, relative clauses, infinitives). In speaking, units are shorter; the relations between them are often expressed covertly without special conjunctions. Consequently, sentences and clauses are longer in a written discourse. The second opposition is ‘**detachment/involvement**’. Passive voice is more common for written texts. Spoken discourse contains frequent reference to the speaker, more indexicals (*you, me, here, now*) and more particles.

According to [Castellà 2004], spoken conversation has a verbal style (more verbs and verbal complements); more reduced phrases and clauses, flexible structures and more repetition as well as profuse use of discourse markers. Spoken genres use a reduced variety of connectives, they are polyfunctional.

Other important features are **fragmentation and special discourse particles**. Thus, speech is discontinuous, its production is a sequence of segments, and the standard concept of a sentence is inapplicable to it (cf. [Lapteva 1976]).

In [Wang et al. 2019: 156] additional relations for RST-style annotation of spontaneous speech are suggested, i.e. unfinished utterance relations, discourse particle relations (*as you know, or right, Okey*, which are satellites of adjacent spans). There are spoken discourse corpora annotated according to PDTB (Penn Discourse Treebank), [Prasad et al. 2008] standards, the annotation was discussed in [Tonelli et al. 2010]; [Rehbein et al. 2016]. According to the study [Crible, Cuenca 2017], based on this type of annotation, the structures containing discourse markers in unplanned speech are often truncated (the second argument is missing/not complete). The clusters of discourse markers are also quite frequent in spoken discourse (cf. *and so, because if*).

To sum up, according to the literature, clauses in spoken discourse are simpler and shorter as compared to the written one. Moreover, there are a number of incomplete subclausal units in the former. The spoken discourse is characterized by high frequency of particles. A high percentage of complex sentences with non-finite forms is a written text feature.

2.3. Spoken discourse features in blogs

There is a considerable number of blog corpora ([Macdonald, Ounis 2006], [Burton et al. 2009], [Mishne et al. 2005], [Quan, Ren 2009], [Santos et al. 2018] etc.), primarily designed to model topic classification or opinion mining tasks. The annotation of blog subcorpus of Ru-RSTreebank focuses on discourse structure.

Many studies deal with the features that blogs share with spoken conversations ([Simaki et al. 2017: 14], [Scheffler et al. 2019] etc.). Fragmented or incomplete clauses, dialogic interaction expressions are among them. According to [Berger, Hennig 2015], blog texts look more personal and diary-like than regular news. There are in-depth studies of blogs in Russian (e.g. [Germasheva 2010], [Kuznetsova 2008], [Novikova 2005]). The researches point out the high frequency of ‘dialogical’ features in blogs, including interaction markers, questioning particles (*da?* ‘yes’, *a?*, *pravilno?* ‘right?’), the high frequency of second person verb forms, imperatives, politeness formulas, questions, and others. Another bulk of features mentioned in studies deals with intentional simulating spoken discourse, its spontaneity and disfluency: ellipsis, a ‘but’ for a topic shift and regulatory particle *nu* for summarization; attitude expressions; ellipses marks that simulate slowed-down speech under the effect of emotional states etc. Other spoken discourse features are nouns in nominative case, incomplete phrases, self-corrections, high frequency of simple sentences, parcellations, colloquialisms and usage of highly expressive lexis.

Therefore, many features differentiating the spoken and written discourse are relevant to blogs characterization. We focus on the ones that are related to discourse structure and discourse properties. These are segmentation into DUs and their structural properties (e.g. part of speech proportions), distribution of rhetoric relations and some properties of discourse markers (e.g. interaction and regulatory markers).

3. Data

3.1. The corpus

The data for our analysis consists of three subcorpora of Ru-RSTreebank:

- (1) news and popular science texts (129 texts);
- (2) scientific papers (100 texts);
- (3) blog subcorpus.

(1) and (2) have 279,426 tokens in total. (3) contains 104 blog texts, 128,917 tokens. Their topics are traveling, cosmetics, sports and health, everyday life, psychology, IT and tech, politics, social aspects (13 texts per a topic). Three main types of blogs—news, commentary, journal (diaries) are presented.

3.2. Blogs segmentation

Though typical elementary discourse units (EDUs) are clauses, according to Ru-RSTreebank annotation rules, certain types of subclausal EDUs are possible in all genres and registers. These are prepositional phrases for cause, purpose etc.: e.g. *из-за переездов* ‘due to relocation’.

Besides, blogs authors sometimes separate text fragments lacking overt predicates by sentence punctuation marks (e.g. *Зря*. ‘In vain’ or *Конечно*. ‘Of course’). Some of these fragments are similar to subclausal EDUs, used in spoken discourse, e.g. parcellation or increments. These segments are considered as EDUs according to annotation rules. Moreover, there are many borderline cases when an EDU consists of a noun phrase in Nominative. These EDUs are treated as full clausal units with a zero copula: *Остановка автобуса. 5 утра. Ни одного человека на улице*. ‘A bus stop. 5 o’clock in the morning. Nobody on the street.’

For experiments, we excluded headlines, bibliography and other metainformation from texts, as well as vocatives and politeness formulas (e.g. ‘Thank you all! Good night’, ‘And how are you doing?’). We also excluded markers for images (IMG) in blogs. This is a controversial issue since some pictures in blogs can be considered as EDUs participating in meaningful discourse relations (e.g. Evidence). IMG can be a part of an EDU, where gestures or deictic indication would be used in speaking, so eliminating IMGs in this case would result in the incoherent text. We keep this type of references to images as part of EDUs.

3.3. Data preprocessing

For our research, we enhance the annotation of Ru-RSTreebank by adding syntactic and morphological layers (Universal dependencies standards). We used DeepPavlov library (<http://deeppavlov.ai/>) for this task (a pretrained model for Russian—ru_syntagrus_joint_parsing). General statistics, involving morphological and syntactic properties of discourse units, is based on the automatic annotation (accuracy reported for DeepPavlov morphological tagging 96.23). We assume that parsing errors do not affect the differences in relative frequencies of parameters among genres.

Besides, we added four other discourse-related layers:

- 1) **morphological type of EDU** (finite or non-finite EDUs dependent on grammatical form of the predicate);
- 2) **syntactic type of EDU**: prepositional phrase, subordinate clause, coordinate clause, etc.;
- 3) **the type of subclausal EDUs** (parcellation, external topics, splits etc. according to [Kibrik, Podlesskaya eds. 2009] classification);
- 4) **parenthetical and other types of intervening constructions**.

4. EDU structural properties

4.1. Rhetoric relations in blogs as compared to other genres

We examine the frequency distribution of different rhetoric relations among genres, taking into account the relations between all level discourse units (not only EDUs). The Elaboration relation is the most frequent one for all three genres—scientific articles, news and blogs (more than 30% of all the segments). General statistics for relations is given in Fig. 1.

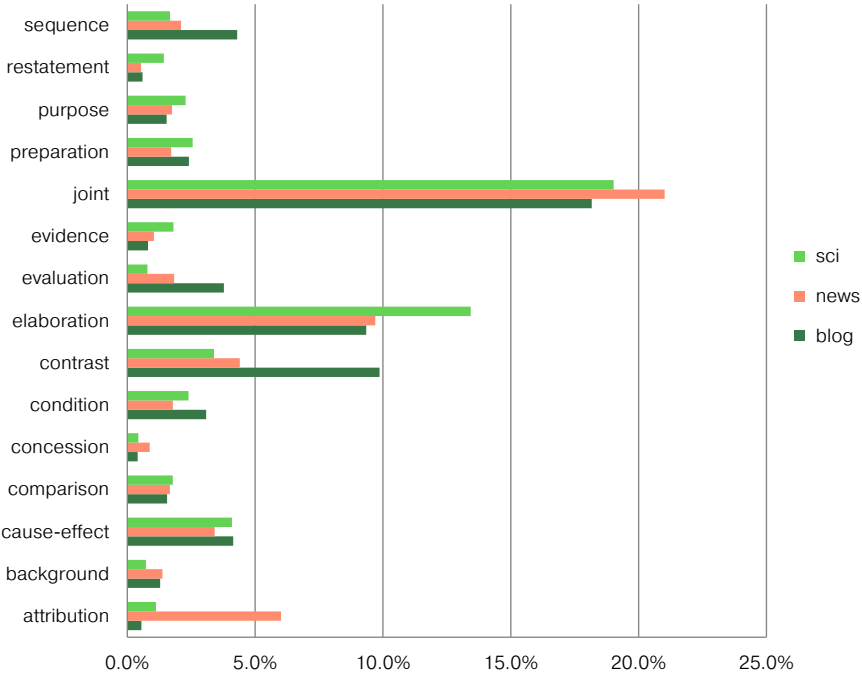


Fig. 1. Rhetoric relations relative frequency for different genres

Fig. 1 shows that for all genres the Joint and Elaboration relations are the most frequent ones. While Attribution is a specific relation for news, and Elaboration is more frequent in scientific texts, blogs are characterized by higher proportion of Sequence, Evaluation and Contrast as compared to other genres. In order to find out the impact of particular relations for genre differentiation, we built a RandomForest classifier with rhetorical relations as features (f1—82%). The relations with the highest impact are Contrast (0.15), Attribution (0.14), Evaluation (0.12) and Elaboration (0.8).

The high proportion of relations with no overt marking (Contrast and Sequence) reflects the tendency for fragmentation. The large number of Evaluation is an evidence of high degree of ‘involvement’. Thus, both tendencies for spoken discourse, named by W. Chafe ([Chafe 1982], see also 2.3) are supported by the rhetoric relations distribution in blogs.

4.2. EDU length

Taking into consideration the tendency for fragmentation in spoken discourse, one would expect a smaller number of complex sentences in blogs, fewer EDUs per sentence. However, the data in **Table 1** shows that there is no statistically significant difference in the number of EDUs per sentence among genres.

Table 1. The average number of EDUs per sentence

Text Genre	Average of sentence segments number	StdDev of sentence segments number
blog	2.12	0.47
news	2.19	0.47
science	2.18	0.44
Grand Total	2.16	0.46

The difference lies in the length of the EDUs (**Table 2**):

Table 2. The average number of tokens per EDU

Text Genre	Average of tokens/segment	StdDev of tokens/segment
blog	8.59	1.50
news	11.08	1.93
science	14.75	2.77
Grand Total	11.02	3.03

While the longest EDUs are in scientific texts, the EDUs in blogs are significantly shorter.

4.3. Morphological and syntactic features of EDUs

Other features referring to the ‘fragmentation/integration’ opposition are sentence complexity, POS (part-of-speech) distribution and subordinate vs. coordinate conjunctions usage.

4.3.1. Verb and noun forms distribution

Table 3 presents the distribution of finite/non-finite verb forms in EDUs among genres (verbless EDUs are not included into the table):

Table 3. The proportion of different verb forms per EDU

Text Genre	Finite	Participle	Infinitive	Converb
blog	63%	4%	22%	2%
news	67%	7%	22%	2%
science	58%	12%	19%	2%

In general, the distribution of verb forms looks similar. For all text genres, the most widely used grammatical type is a finite verb EDU, while converbs are rarely used in all genres. There is a slight difference in EDUs headed by participles. They are more frequent in scientific texts.

As for nouns, the distribution of nouns per EDU is shown in **Table 4**:

Table 4. Nouns frequency

Text Genre	Nouns average per EDU
blog	1.92
news	2.90
science	4.13

This data agrees with the expectations that written discourse is characterized by a high proportion of nouns. The rate of nouns in blogs is lower than in news or scientific texts (cf. [Chafe 1982] concerning the opposition of written vs. spoken discourse).

4.3.2. Verbless EDUs

According to 4.1, there is no big difference in verb forms distribution across genres. The main difference is in EDUs without verbs, see **Table 5** (we do not include sentences, erroneously parsed as rootless, though many of them are also clauses with no overt copula):

Table 5. The average number of root sentences without verbs per EDU

Text Genre	Average of root sentences without verbs per EDU
blog	18.01%
news	8.96%
science	10.98%
Grand Total	12.55%

There are verbless EDUs in news and science subcorpora. These EDUs are often prepositional phrases signaling purpose or cause relations. Another type of verbless EDUs are clauses with no overt copula. In scientific texts, these are primarily definitions or characterizations:

- (1) *Аргументация это универсальный феномен.*
Argumentation is a universal phenomenon.

However, the highest proportion of verbless sentences is in blogs. Indeed, there are a lot of sentences with zero copula (see also 5.2):

- (2) *Кухня одна на всех.*
lit. Kitchen one for all.
- (3) *Медынь, автостанция.*
lit. Medyn, a bus-station.

The high proportion of EDUs consisting of noun phrases only is also an evidence for high degree of fragmentation in blogs.

4.3.3. Subjectless EDUs

Based on general assumption that scientific writers often try to avoid the use of personal expressions and to demote human subjects, we expected the high proportion of subjectless clauses in the scientific subcorpus. However, our research has shown that in the blogs subcorpus the proportion of subjectless clauses is the highest one (**Table 6**):

Table 6. Subjectless main clauses

Text Genre	No subject + root
blog	17.12%
news	7.27%
science	9.92%
Grand Total	4341

While the subjectless clauses in scientific texts or news are often impersonal clauses (e.g. *было показано* ‘(it) was shown’), in blogs the majority of the clauses of this type are clauses with an anaphoric or a personal zero pronoun (pro) as Subject (or with pronominal ellipsis), as in *Встала рано. Вышла ровно в 6* ‘(I/she) got up early in the morning. (I/she) went out at 6 a.m.’

4.3.4. Impact of POS and dependency relations distribution on genre differentiation

In order to check features impact on genre differentiation, we built a Random-Forest classifier based only on POS and some syntactic relations features (Accuracy: 0.84 (\pm 0.02)). The top 11 features (out of 25) are presented in **Table 7**.

Table 7. POS and syntactic relation features impact

Feature	Impact
number of nouns per EDU	0.14
number of tokens per EDU	0.10
relative frequency of adverbs	0.07
number of parataxis relations per EDU	0.07
relative frequency of pronouns	0.07
relative frequency of verbless EDU	0.05
relative frequency of particles	0.05
relative frequency of EDUs started with a coordinative conjunction	0.05
relative frequency of prepositional phrases (CASE syntactic relation)	0.04
relative frequency of subjectless clauses	0.04
relative frequency of nouns	0.04

Table 7 shows that ranking of features goes hand in hand with the analysis of features suggested above. The most important features are number of tokens per EDU, number of nouns per EDU (these are highly correlated features), the proportion of verbless and subjectless per text.

Another significant difference between blogs as opposed to news and scientific texts is the distribution of subordinate vs. coordinate conjunctions. While there is no significant difference in subordinate conjunctions distribution, the number of EDUs started with coordinate conjunctions is a quite important feature (20% EDUs in blogs, 12% in news and 8% in scientific texts), as well as number of particles per token (4% of tokens in blogs, 2% in news, and only 1% in scientific texts). The high proportion of coordinative conjunctions and particles in blogs agrees with the high number of Contrast relations in blogs.

5. Parallels between blog and spoken EDUs features

5.1. Non-canonical EDUs in blogs

5.1.1. Speech disfluency and prosodic features

Some of the prosodic features of spoken discourse can be mapped into special cases of punctuation in blogs. While incomplete EDU boundaries in speech are usually detected via pauses and prosodic contour, the speaker in blogs often uses punctuation marks for separating the parts of a clause (a phenomenon similar to parcellation or incrementation):

- (4) *Вот, встретили по пути борзую. В махровых тапочках.*
Here, we met a greyhound along the way. **In terry slippers.**

Besides, speakers use ellipses marks in blogs as an alternative of hesitation markers.

5.1.2. “Quasi-echo” EDUs

One of the constructions imitating spoken phenomena in blogs is a sequence of incomplete clauses (with an ellipses mark) with the repetition of the same idea:

- (5) a) *Знаете, такое ощущение...*
b) *вот мне почему-то кажется, что это...*
c) *нет, ну я могу ошибаться... А что если это связано с новым президентом?*
a) You know, I **have such a feeling**...
b) for some reason, **it seems to me that**...
c) no, **well, I could be wrong**... What if this could be related to the new president?

This construction can imitate false start (cf. (a) and (b) in (5)).

The EDUs in this construction can contain regulatory markers and markers of interaction (e.g. *вот* (interaction marker), *ну* ‘well’, *просто* ‘just’, *знаете* ‘you know’, etc.). Their function is to express uncertainty or to focus the attention on a particular DU.

5.1.3. Focus/topic extraction

An isolated noun phrase can precede a clause where the same noun phrase is an argument (10 examples in blog subcorpus). It can be either in focus or in topic in this clause. Author uses topic repetition to highlight the main topic.

- (6) **Факты.** *Всегда смотрите на факты.*
Facts. Always look at the facts.

Another case is topic extraction with corresponding noun phrase ellipsis:

- (7) **Внутренние монголы.** *Путешествуют всегда организованными группами.*
Inner Mongols. Travel always in organized groups.

5.1.4. Retrospective subclausal EDUs

Blog-writers use retrospective subclausal EDUs mostly as “adjuncts or attributes that semantically belong to a clause but constitute a separate short EDU” [Kibrik 2015: 229]. There are 101 examples of retrospective EDUs in blog subcorpus (cf. 12 cases in news, five are in reported speech):

- (8) *И каждый менеджер уникален. Как снежинки. Уникальные люди-снежинки.*
 And every manager is unique. **Like snowflakes.** Unique snowflake people.

5.1.5. Noun phrase chaining

Another construction, simulating spoken discourse features, is nominal clause chaining or chaining of noun phrases with no overt copula (their exact syntactic status is often unclear):

- (9) *Отель-чемодан-метро. Тайм-чек.*
 Hotel-suitcase-metro. Time check.
- (10) *Моя кожа: 29 лет, комбинированный тип, чувствительная, акне в ремиссии, пост-акне, быстро забиваются поры.*
 lit. My skin: 29 years old, mixed type, sensitive, acne in remission, post-acne, pores are quickly logged.

5.1.6. Mixture of predicate types in coordinate constructions

One more construction of special interest is a multinuclear relation where syntactically heterogeneous phrases (noun phrases, finite clauses, etc.) form a coordinating construction.

- (11) *Электричка "Стандарт-плюс". Цивильная и комфортная, с мягким ходом, есть вай-фай. Два часа блаженства и резкий контраст с тем, что будет дальше.*
 Electric train Standard plus. Civil and comfortable, with a smooth ride, there is Wi-Fi. Two hours of bliss and a sharp contrast with what will happen next.

5.1.7. Splits

There are special types of splits in blogs that are marked with three dots or exclamation marks, interjections or emoticons (77 cases)

- (12) *А ты мама... вот тогда злилась на меня, когда мы опаздывали на электричку!*
And you mom... then got mad at me when we were late for the train!

- (13) **Ой:)** *сейчас же в моде такие словечки...*
Oh:) these words are now in fashion

5.2. Embedded parataxis constructions

The parenthetical phrases intervening into EDUs occur in all three genres (blogs—528 (3.7% of EDUs), news 482 (3.9%), scientific texts—592 (6.8%)). Table 8 presents the distribution of factual vs. evaluative EDUs occurring in parenthesis:

Table 8. The distribution of different types of parenthetical constructions among genres

Parameter / text genre	blogs	news	scientific texts
factual information: elaboration / disclosure	266	513	580
evaluation / interpretation	189	43	12
Total	455	546	592

In news, the function of phrases in parenthesis is to enforce author’s interpretation of facts. Among the three genres, the parenthetical constructions most often occur in scientific texts. Interactive communication with the audience in scientific texts is represented only by reduced scientific abbreviation-clichés: *ср.* ‘cf’, *см.* ‘see’, etc.

In blogs, parenthetical constructions are used as blog-writer “protocols”. They often include evaluative expressions:

- (14) *Сладости я оставила только без сахара и органический шоколад в очень умеренных дозах (чтоб не так хотелось убивать людей))*.

I kept only sweets without sugar and organic chocolate in very moderate doses (not to want to kill people)).

Many of them contain irony:

- (15) *Газовщица заполнила договор (один экземпляр), теща подписала и спрашивает (все-таки бывший директор школы), а ей экземпляр?*

The gas worker filled out the contract (one copy), the mother-in-law signed and asked (after all, the former principal), but what about her copy?

Moreover, sometimes they are markers of interaction:

- (16) *Сегодня будем размышлять над следующим изречением (тем, кто не знает иностранных языков — Google в помощь).*

Today we will reflect on the following saying (for those who do not know foreign languages—Google can help (you)).

6. Conclusion

To sum up, our research shows the difference in discourse properties of the three subcorpora in Ru-RSTreebank. The distribution of rhetoric relations differs among genres.

Our corpus study provides an additional evidence for the claim that blogs have spoken discourse features. The rhetoric relations distribution, the occurrence of sub-clausal EDUs of certain types, the part of speech distribution differ blogs from the other two genres, science articles and news. There are special devices that are used by blog authors to imitate the spontaneous speech.

References

Examples of sources for blog texts:

- <http://www.livejournal.com>
 - <https://medium.com>
 - <http://www.blogspot.com>
 - <https://kosmetista.ru>
 - <https://habr.com>
 - <https://airfit.ru/blog>
1. *Akhapkina Ya. E.* (2014), About the grammar of oral-written speech [O grammatike ustno-pis'mennogo viskazivaniya], Contemporary Russian language in Internet, pp. 181–194.
 2. *Berger P., Hennig P., Schoenberg V., and Meinel C.* (2015), Blog, forum or newspaper? Web genre detection using SVMs, 2015 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 3, pp. 64–68.
 3. *Burton K., Java A., and Soboroff I.* (2009), The icwsm2009 spinn3r dataset, Third Annual Conference on Weblogs and Social Media (ICWSM 2009).
 4. *Castellà J. M.* *Oralitat i Escriptura.* (2004), Dues Cares de la Complexitat del Llenguatge, Publicacions de l'Abadia de Montserrat, Barcelona.
 5. *Chafe, Wallace* (1982), Integration and involvement in speaking, writing, and oral literature, Spoken and written language: Exploring orality and literacy, ed. D. Tannen, Norwood: Ablex, pp. 35–54.
 6. *Chistova, E., Shelmanov, A., Kobozeva, M., Pisarevskaya, D., Smirnov, I., and Toldova, S.* (2019), Classification models for RST discourse parsing of texts in Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”, pp. 163–176.
 7. *Cribble L., Cuenca M. J.* (2017), Discourse markers in speech: characteristics and challenges for corpus annotation, Dialogue and Discourse, Vol. 8(2), pp. 149–166.
 8. *Da Cunha, I.* (2016), Towards discourse parsing in Spanish, TextLink–Structuring Discourse, Multilingual Europe Second Action Conference Károli Gáspár University of the Reformed Church in Hungary, Budapest, 11–14 April, p. 40.
 9. *Germasheva T. M.* (2010), Studies of linguistic and paralinguistic features of blogs discourse [Issledovanie lingvisticheskikh i paralingvisticheskikh kharakteristik blog-discursiva], Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta im. A. I. Gertsena, 126, pp. 150–155.
 10. *Iruskieta, M., Da Cunha, I., and Taboada, M.* (2015), A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora, Language resources and evaluation, 49(2), pp. 263–309.

11. *Kibrik A. A.* (2009), Modus, genre and other parameters of discourses classification [Modus, ganr i drugie parametri klassifikatsii discursov], *Voprosi yasikoznaniya*, 2 (3).
12. *Kibrik A. A.* (2015), The problem of non-discreteness and spoken discourse structure, *Computational linguistics and intellectual technologies*, 14 (21), vol. 1, pp. 225–233.
13. *Lapteva O. A.* (1976), Russian spoken syntax [Russkiy razgovorniy sintaksis], Moscow, Nauka.
14. *Macdonald C., Ounis I.* (2006), The trec blogs06 collection: Creating and analysing a blog test collection. Department of Computer Science, University of GlasgowTech Report TR-2006–224, 1:3–1.
15. *Mann W. C., Thompson S. A.* (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, *Text* 8, 3, pp. 243–281.
16. *Matthiessen C. M., Teruya K.* (2015), Grammatical realizations of rhetorical relations in different registers, *Word*, 61(3), pp. 232–281.
17. *Mishne G. et al.* (2005), Experiments with mood classification in blog posts, *Proceedings of ACM SIGIR 2005workshop on stylistic analysis of text for information access*, Vol. 19, pp. 321–327.
18. *Novikova E. G.* (2005), Language features of text structuring in classic and online diaries [Yazikovie osobennosti organizatsii tekstov klassicheskogo i setevogo dnevnikov], AutoAbstract of Ph. D. Thesis, Stavropol.
19. *Pisarevskaya D. et al.* (2017), Towards building a discourse-annotated corpus of Russian, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017”*, pp. 194–204.
20. *Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., and Webber, B.* (2008), The penn discourse treebank 2.0, *Proceedings of the 6 International Conference on Language Resources and Evaluation, LREC’08*, pp. 2961–2968.
21. *Quan C., Ren F.* (2009), Construction of a blog emotion corpus for chinese emotional expression analysis, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Vol. 3, pp. 1446–1454.
22. *Rehbein I., Scholman M., and Demberg V.* (2016), Annotating Discourse Relations in Spoken Language: A Comparison of thePDTB and CCR Frameworks, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1039–1046.
23. *Santos H., Woloszyn V., Vieira R.* (2018), BlogSet-BR: A Brazilian Portuguese Blog Corpus, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 661–664.
24. *Scheffler T., Aktaş B., Das D., and Stede M.* (2019), Annotating Shallow Discourse Relations in Twitter Conversations, *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pp. 50–55.
25. *Simaki V., Skeppstedt M., Paradis C., Kerren A., and Sahlgren M.* (2017), Annotating Speaker Stance in Discourse: The Brexit Blog Corpus, *Corpus Linguistics and Linguistic Theory*, 34 pages. Available at https://portal.research.lu.se/ws/files/34256897/Corpus_Linguistics_and_Linguistic_Theory_Annotating_Speaker_Stance_in_Discourse_The_Brexit_Blog_Corpus.pdf.

26. *Simaki V., Paradis C., and Kerren A.* (2018), Evaluating stance-annotated sentences from the Brexit Blog Corpus: A quantitative linguistic analysis, *ICAME Journal*, Vol. 42, pp. 133–165.
27. *Night dream stories: Russian Corpus Study of Oral Discourse* (2009) [Rasskazy o snovideniyah: korpusnoe issledovanie ustnogo russkogo diskursa], Ed. Kibrik A. A., Podlesskaya V. I.
28. *Tonelli S., Riccardi G., Prasad R., and Joshi A.* (2010), Annotation of discourse relations for conversational spoken dialogs, *Proceedings of the 7 International Conference on Language Resources and Evaluation, LREC'10*, pp. 2084–2090.
29. *Wang X., Gyawali B., Bruno J. V., Molloy H. R., Evanini K., and Zechner K.* (2019), Using Rhetorical Structure Theory to Assess Discourse Coherence for Non-native Spontaneous Speech, *Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019)*, pp. 153–162.
30. *Zeldes A.* (2017), The GUM corpus: Creating multilayer resources in the classroom, *Language Resources and Evaluation*, 51(3), pp. 581–612.

DISTRIBUTIONAL MODELS AND AUXILIARY METHODS FOR DETERMINING THE HYPERNYMS OF WORDS IN RUSSIAN

Yadrintsev V. V. (vvyadrincev@gmail.com)

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; Peoples Friendship University of Russia (RUDN University), Moscow, Russia

Ryzhova A. A. (Anastasiia.Ryzhova@skoltech.ru)

Skolkovo Institute of Science and Technology, Moscow, Russia

Sochenkov I. V. (sochenkov@isa.ru)

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

This paper describes our participation in the first shared task on Automatic Taxonomy Construction for the Russian language RUSSE'2020. The goal of this task is the following: input words (neologisms that are not yet included in the taxonomy) need to be associated with the appropriate hypernyms from an existing taxonomy. For example, for the input word "duck", it is expected that participants will provide a list of its ten hypernyms-synsets to which the word can most likely be attributed, such as "animal," "bird" and so on. An input word can refer to one, two, or more "parents" at the same time.

In this article we are trying to answer the following question: what results can be achieved using only "raw" vectors from distributional models without additional training? The article presents the results for several pre-trained models that are based on fastText, Elmo, and BERT algorithms. Also, an out-of-vocabulary analysis was performed for the models under consideration. Taking into account all public scores from the leaderboards, we showed the results corresponding to the following places in the ranking: the 3rd place on public nouns, the 2nd on private nouns, the 4th on public verbs, and the 4th on private verbs.

Keywords: vector models, hypernym discovery, fastText, Elmo, BERT, rus-vectores, RuWordNet

DOI: 10.28995/2075-7182-2020-19-762-772

ДИСТРИБУТИВНЫЕ МОДЕЛИ И ВСПОМОГАТЕЛЬНЫЕ МЕТОДЫ ДЛЯ ОПРЕДЕЛЕНИЯ ГИПЕРОНИМОВ СЛОВ РУССКОГО ЯЗЫКА

Ядринцев В. В. (vvyadrincev@gmail.com)

ФИЦ ИУ РАН; РУДН, Москва

Рыжова А. А. (Anastasiia.Ryzhova@skoltech.ru)

Сколковский институт науки и технологий, Москва

Соченков И. В. (sochenkov@isa.ru)

ФИЦ ИУ РАН, Москва

В работе описывается наше участие в первой задаче по автоматическому построению таксономии для русского языка RUSSE'2020. Цель этой задачи заключается в следующем: входным неизвестным словам (неологизмам, которых ещё нет в таксономии) нужно сопоставить гиперонимы из существующей таксономии. Например, ожидается, что для слова «утка» участники предоставят список десяти его наиболее вероятных синсетов-гиперонимов («животное», «птица» и т. д.). Входное слово может одновременно относиться к одному, двум или более «родителям». В этой статье мы показываем, каких результатов можно достичь, используя только «сырые» векторы из дистрибутивных моделей без какого-либо дополнительного обучения. В работе представлены результаты для нескольких предобученных моделей, которые основаны на алгоритмах fastText, Elmo и BERT. Кроме того, для рассматриваемых моделей был проведён анализ полноты словарей. Принимая во внимание все опубликованные результаты рейтингов участников, мы показали результаты, соответствующие следующим местам: третьему на «общедоступных» существительных, второму на «конфиденциальных» существительных, четвертому на «общедоступных» глаголах и четвертому на «конфиденциальных» глаголах.

Ключевые слова: векторные модели, определение гиперонима слова, fastText, BERT, rusvectors, RuWordNet

1. Introduction

A hypernym—hyponym relation is a word/phrase pair (x, y) such that x is a hyponym of y , the “is-a” relationship, for example, “a dog is an animal.” Here “dog” is a hyponym for “animal”, and “animal” is a hypernym for the word “dog”.

Identifying hypernymic relations has a lot of applications in Natural Language Processing, especially in semantically intensive tasks, such as Question Answering, Textual Entailment, and semantic search systems. These relations play a crucial role in thesauri construction, but it is challenging and not effective to extract them manually.

We participated in the shared task on Automatic Taxonomy Construction for the Russian language (RUSSE'2020¹). The goal of this task is the following: neologisms need to be associated with the appropriate hypernyms from an existing taxonomy. As a taxonomy the RuWordNet (Russian WordNet) is used, the format of which is similar to the English WordNet format. The task consists of two subtasks:

- nouns (two test sets: public and private)
- verbs (two test sets: public and private).

The organizers provided a baseline that leverages pre-trained models to obtain word vectors. Our method is an improvement on the baseline. We intentionally employed a simple approach to identifying a hypernym of a word, which we describe below. The reason for this was that we were interested in whether the Russian taxonomy construction task can be solved using already available algorithms and pre-trained models without additional training. Even using the simple approach, we showed results that were not lower than the 4th place (from more than 13 participants) on each of the test sets.

The rest of the paper is organized as follows. **Section 2** briefly outlines the previous work related to our task. In **Section 3** we present the datasets offered by the shared task organizers and used pre-trained models. **Section 4** provides the details of the employed approach. In **Section 5** we describe the results, and in Section 6 we conclude.

2. Related work

Many automatic methods for identifying hypernymic relations have been explored in the last years. There are two popular ways of extracting such relations, a pattern-based one and a distributional one. The pattern-based approach uses the joint co-occurrence of the word and its hypernyms in texts [1], [11], while the distributional approach exploits distributional representations of words [3], [16]. Marti Hearst first introduced the now widely used pattern-based method for the English language in 1992 [5], [13]. She manually designed the patterns for hypernym—hyponym extraction from texts. For example, the pattern “such NP as NP” helps to extract such pair as “author, Shakespeare” from the sentence “such authors as Shakespeare.” Shared tasks are described in the paper of the Organizers [12].

For the Russian language, this problem is not so highly investigated. In [14] Sabirova et al. propose a rule-based method for hypernym—hyponym extraction from Russian texts. They created six patterns, e.g., “*Y—вид/тип/форма/разновидность/сорт X (Y is a kind/type/form/sort of X)*”, and then applied finite-state transducers to extract the patterns from texts. In [6] the researchers clustered the definitions from the large dataset (using [7] as a starting point) and then extracted hypernym candidates using patterns for verbose candidates. As a complementary method they trained the SVM classifier to obtain the best candidates.

¹ <https://russe.nlpub.org/2020/isa/>

3. Data overview

RuWordNet thesaurus and train data are described in the paper of the Organizers [12]. In the present work we use the following pre-trained models:

1. `ft_cc_ru_300`²,
2. RuBERT³,
3. `ruscorpora_none_fasttextskipgram_300_2_2019`,
4. `tayga_none_fasttextcbow_300_10_2019`,
5. `araneum_none_fasttextcbow_300_5_2018`,
6. `tayga_lemmas_elmo_2048_2019`.

The first one, `ft_cc_ru_300`, includes pre-trained word vectors for Russian from *Facebook* [4]. The second one, **RuBERT**, is an adopted BERT for Russian [10]. Models 3–6 contain pre-trained word vectors for Russian from **rusvectors**⁴ [9].

Please note that in the RuBERT model we only consider the hidden layer with dimension 3,072, using it as word vectors. This idea is taken from the baseline provided by the organizers of the competition. Accordingly, vector dimensions of models 1, 3–5 are 300, model 2—3,072, and model 6—1,024.

It is most likely that the largest text corpus was used for `ft_cc_ru_300`, which includes *Wikipedia* and *Common Crawl*⁵ (we do not know the exact volume of crawl data for Russian, but roughly 24 terabytes of plain text were used for 157 languages [4]). **RuBERT** was trained on the *Wikipedia* and *news data*, **ruscorpora_none_fasttextskipgram_300_2_2019**—on *Russian National Corpus*⁶. **Tayga_none_fasttextcbow_300_10_2019** and **tayga_lemmas_elmo_2048_2019** were trained on the *TAIGA*⁷ corpus [15]. Finally, **araneum_none_fasttextcbow_300_5_2018** was obtained by training on the *Araneum Russicum Maximum* [2].

4. Our approach

The first subsection briefly describes the baseline. The following subsections describe additional steps taken to improve the baseline. Proposed improvements significantly increased the results on the test samples.

4.1. Baseline

This subsection briefly describes the baseline provided by the competition organizers. The common-crawl fasttext (300-d) model is used to obtain synset vectors

² <https://fasttext.cc/docs/en/crawl-vectors.html>

³ <http://docs.deeppavlov.ai/en/master/features/models/bert.html>

⁴ <https://rusvectors.org/ru/models/>

⁵ <https://commoncrawl.org/>

⁶ <http://ruscorpora.ru/>

⁷ https://tatianashavrina.github.io/taiga_site/

and unknown word vectors. The synset vector is the average word vector of all synset senses. Variables *nouns_cnt* and *verbs_cnt* denote the number of synsets-nouns and synsets-verbs respectively. As noted earlier, the total number of nouns is ~29,300, of verbs—~7,500. For the existing taxonomy, separate vector matrices are created for nouns and verbs of sizes *nouns_cnt* × 300 and *verbs_cnt* × 300 respectively. For each unknown word, the closest synsets are searched by cosine measure, and, depending on the approach, they are considered as synonyms or hypernyms.

4.2. Proposed improvements

To achieve better results, we proposed the following improvements:

1. Addition of ranking at the final stage: sorting synsets based on the recalculated rate for each *synset_id*. It gave the most significant improvement in results (the MAP was increased by 5–6%) and will be described separately in [section 4.2.1](#).

2. Extension of the string representation of the synset. The following fields were considered as parameters: **ruthes_name**, **definition**, **sense_name**, **sense_lemma**, and **sense_main_word**. We have discovered that for nouns a combination of two fields (**ruthes_name**, **sense_name**) is better, while for verbs all fields combined work the best. The above combinations were applied for all models except RuBERT. For RuBERT we leveraged a standard string representation, consisting of the **sense_names** of the senses. The usage of the non-standard combinations improved the results only slightly (the MAP increased by 1–3%). Here is an example of a synset: **synset_id**="109649-N" **ruthes_name**="ДЗЮДО" ("judo") **definition**="японская борьба, произошедшая из джиу-джитсу, олимпийский вид спорта" ("Japanese wrestling that took place from Jiu-Jitsu, an Olympic sport"). Here are the senses of the synset 109649-N:

- **sense_id**="109649-N-181880" **sense_name**="БОРЬБА ДЗЮДО"
sense_lemma="БОРЬБА ДЗЮДО" **sense_main_word**="БОРЬБА";
- **sense_id**="109649-N-136843" **sense_name**="ДЗЮДО"
sense_lemma="ДЗЮДО" **sense_main_word**="".

Thus, for the synset 109649-N the following line will be initial: "ДЗЮДО<sep>БОРЬБА ДЗЮДО<sep>ДЗЮДО" (in case the fields **ruthes_name** and **name** are used). Space plays the role of the separator <sep>.

3. Addition of other relationships between synsets. We tried adding the "domain" relation. For example, word "judo"⁸ is a part of "sport" (*спорт*) and "amateur wrestling" (*спортивная борьба*) domains, and "judo" has hypernyms "Martial Arts" (*боевые искусства*) and "east Martial Arts" (*восточные единоборства*). However, it worsened the results slightly.

4. Usage of train data to get "parents." It influenced minimal deterioration.

5. Normalization of the words of the string representation of synsets. It improved the results (the MAP was increased by 1–3%) and will be described separately in [section 4.2.2](#).

6. Lemmatization of all words from a string representation of a synset. The results have changed slightly.

⁸ <http://www.ruwordnet.ru/ru/search/%D0%94%D0%97%D0%AE%D0%94%D0%9E>

4.2.1. Ranking

This improvement consists of adding parameters to the original algorithm. The ranking algorithm uses the following parameters:

- The number of synsets-associates— k .
- The number of final synsets-hypernyms— n .
- The probability that the synset-associate is a hypernym of the input word— p_1 .
- The probability that the hypernyms of the synset-associate are the input word hypernyms— p_2 .
- The probability that the hypernyms of the hypernyms of the synset-associate are the input word hypernyms— p_3 .

For the synsets a matrix of vectors M is formed. Vector V is assigned an input word. The number of rows in the matrix M is the same for all models: it is equal to the number of synsets-nouns or synsets-verbs. The number of columns, as well as the dimension of the vector V , depends on the model. It is mentioned in the corresponding [section 3](#). The relevance R is calculated using an unnormalized measure. In the beginning, each synset from the thesaurus is associated with $R = 0$. At the first step of the algorithm, a search is performed (by cosine measure) for the k closest synsets-associates. Technically, we look for vectors that are close to V in the matrix M . Assume r is a cosine measure for a synset-associate. There is a simple recalculation of R , consisting of three steps:

- R of the synset-associate increases by $r \cdot p_1$;
- R of hypernyms of the synset-associate increases by $r \cdot p_2$;
- R of hypernyms of synsets from previous step increases by $r \cdot p_3$.

Hypernyms in the second and third steps are taken from the thesaurus using the “*hypernym*” relation. At the end of the algorithm, the top n (by R) synsets-hypernyms are selected for the answer.

4.2.2. Normalization

- Firstly, all words are converted to lowercase.
- Secondly, all punctuation except for a hyphen (“-”) is replaced by a space. The list of punctuation symbols is as follows: ‘\$’, ‘!’, ‘.’, ‘?’, ‘+’, ‘[’, ‘\xa0’, ‘%’, ‘\“\”, \u00bb’, ‘*’, ‘;’, ‘:’, ‘)’, ‘@’, ‘/’, ‘\u00a7’, ‘”’, ‘_’, ‘\u00b7’, ‘;’, ‘#’, ‘\u0213’, ‘\|\|’, ‘:’, ‘\xad’, ‘{’, ‘\u0214’, ‘>’, ‘|’, ‘\u00ab’, ‘}’, ‘}’, ‘\“’, ‘&’, ‘=’, ‘^’, ‘<’, ‘(’, ‘~’, ‘\u00b0’. Note that non-standard characters from the RuWordNet words are also included in this list.
- Then, using the pymorphy2⁹ [8] morphological analyzer, functional words (prepositions, conjunctions, etc) are removed. We restricted the tags *NPRO*, *PRED*, *PREP*, *CONJ*, *PRCL*, *INTJ*.
- If “Geox” is present in the word tag list, the first letter is replaced with a large one. If parameter *lowercase* == *true*, then this change does not work.

⁹ <https://pymorphy2.readthedocs.io/>

4.3. Out-of-vocabulary analysis

Table 3 presents the out-of-vocabulary analysis for all models (except **RuBERT**) on public, private, and RuWordNet words. RuWordNet words are normalized in the same way as in evaluation. The first line in **Table 3** shows the number of unique words separately for nouns and verbs. It should be noted that the string representation of the synset can include nouns, verbs, and other parts of speech, regardless of the synset part of the speech. Thus, the number of words for N (53,082) in the latest column does not mean that all 53 thousand words are nouns.

It was interesting for us to see how well the words are presented in the vocabularies of models. The observations from **Table 3** are the following:

- **ft_cc_ru_300** best represents the words of RuWordNet (coverage is 86.8% for Nouns and 89.2% for Verbs).
- **araneum_none_fasttextcbow_300_5_2018** best represents the test nouns (coverage is 97.1% for Public Nouns and 96.9% for Private Nouns).
- **tayga_lemmas_elmo_2048_2019** best represents the test verbs (coverage is 89.1% for Public Verbs and 88.8% for Private Verbs).

Table 3. Out-of-vocabulary analysis

Model	public N = 762 public V = 175 in vocab (rate) PoS	private N = 1525 private V = 350 in vocab (rate) PoS	RuWordNet synsets. <i>normalized=True,</i> <i>lemmatized=False.</i> N = 53,082; V = 27427
ft_cc_ru_300	722 (0.947) N 140 (0.8) V	1,443 (0.946) N 279 (0.797) V	46,079 (0.868) N 24,470 (0.892) V
ruscorpora_none_ fasttextskipgram_ 300_2_2019	548 (0.719) N 145 (0.828) V	1,094 (0.717) N 281 (0.802) V	30,625 (0.576) N 17,659 (0.643) V
tayga_none_fasttextcbow _300_10_2019	550 (0.721) N 153 (0.874) V	1,100 (0.721) N 302 (0.862) V	31,089 (0.585) N 17,975 (0.655) V
araneum_none_ fasttextcbow_300_5_2018	740 (0.971) N 100 (0.571) V	1,479 (0.969) N 208 (0.594) V	31,341 (0.590) N 13,827 (0.504) V
tayga_lemmas_ elmo_2048_2019	592 (0.776) N 156 (0.891) V	1,209 (0.792) N 311 (0.888) V	32,563 (0.613) N 18,640 (0.679) V

5. Results

The results are presented in **Table 4**. Note that we used RuBERT in an uncommon way. Also, we would like to highlight that in this case the set of fields for the string representation of the synset is different from other models.

Here we list the same parameters for all models in **Table 4**:

- The ranking algorithm is used with the parameters $p1 = 0.1$, $p2 = 1.0$, $p3 = 1.0$, $k = 10$ and $n = 10$. These parameters were obtained with the grid search. The following values were considered: for $p1$, $p2$, $p3$ —0.1, 0.5, 1.0, 1.5; for k —3, 5, 7, 10, 20, 50, 100; for p —3, 5, 7, 10.
- Neologisms (input words) are lowercase.

- The comparison indicator is the MAP¹⁰ provided by the organizers of the competition.

String representations of the synsets are different for **RuBERT**: all models except RuBERT used **ruthes_name** and **sense_name** for **Nouns** and all possible descriptions for **Verbs**. RuBERT used just **sense_name** for both **Nouns** and **Verbs**.

Next, we describe the names of the columns and rows of the tables. The first column is the name of the model. The second and the next columns are results for a **Public** or **Private** test set. “Lemmas” means that morphological analysis and lemmatization by pymorphy2 are performed. The main cells show the result, the letter after the MAP denotes part of speech (N—nouns, V—verbs).

Table 4. Results by models

Model	Public			Private		
	lowercase MAP PoS	lemmas MAP PoS	lemmas lowercase MAP PoS	lowercase MAP PoS	lemmas MAP PoS	lemmas lowercase MAP PoS
ft_cc_ru_300	0.511 N 0.291 V	0.512 N 0.287 V	0.512 N 0.286 V	0.512 N 0.359 V	0.516 N 0.345 V	0.515 N 0.346 V
tayga_none_fasttext cbow_300_10_2019	0.250 N 0.210 V	0.249 N 0.220 V	0.248 N 0.219 V	0.254 N 0.253 V	0.254 N 0.253 V	0.255 N 0.253 V
araneum_none_fasttext cbow_300_5_2018	0.345 N 0.188 V	0.350 N 0.209 V	0.350 N 0.208 V	0.365 N 0.235 V	0.371 N 0.229 V	0.372 N 0.229 V
tayga_lemmas_ elmo_2048_2019	0.360 N 0.334 V	0.365 N 0.314 V	0.367 N 0.307 V	0.410 N 0.387 V	0.405 N 0.379 V	0.405 N 0.370 V
RuBERT	0.329 N 0.183 V	—	—	0.318 N 0.190 V	—	—

Here are some observations from **Table 4**:

- Lemmatization (of synset representations) did not significantly affect the results. Some models showed a slightly better result, and some a little worse.
- **ft_cc_ru_300** performed the best results on nouns.
- **tayga_lemmas_elmo_2048_2019** performed the best results on verbs.
- On Private Verbs models show the results which are 4–6% better than on Public Verbs. However, we do not observe this on Nouns, except the **tayga_lemmas_elmo_2048_2019** model.
- The application of the model “**RuBERT**” in this way did not show high results.

Finally, **Table 5** shows our best-submitted results compared to the baseline and the best results in the competition. As one can observe, the results we have obtained are competitive.

¹⁰ [https://en.wikipedia.org/wiki/Evaluation_measures_\(information_retrieval\)#Mean_average_precision](https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#Mean_average_precision)

Table 5. The best results of our experiments, which we submitted to the RUSSE shared task. **Our** team submitted results through the participant **vyadrincev**

Dataset	Model, method	Test MAP (public)	Rank (public)	Test MAP (private)	Rank (private)
Nouns	Unknown, best in the competition	0.5590	1 of 14 ¹¹	0.5522	1 of 17 ¹²
Nouns	ft_cc_ru_300, our	0.5115	3 of 14	0.5163	2 of 17
Nouns	ft_cc_ru_300, baseline	0.4348	9 of 14	0.4210	9 of 17
Verbs	Unknown, best in the competition	0.4033	1 of 14 ¹³	0.4483	1 of 14 ¹⁴
Verbs	tayga_lemmas_ elmo_2048_2019, our	0.3342	4 of 14	0.3874	4 of 14
Verbs	ft_cc_ru_300, baseline	0.2759	8 of 14	0.3335	6 of 14

6. Discussion and conclusion

This article is a description of our participation in the joint task RUSSE’2020 on automatic taxonomy construction for the Russian language. We intended to create a simple method based on the baseline, using pre-trained models.

Using BERT as a distribution model for obtaining vectors, we were not able to achieve high results. Therefore, as future work, we want to train RuBERT for classifying strings like “<WORD> is a <PARENT SYNSET>”. However, we can face some challenges. Firstly, the string representation of synsets is often quite long. Secondly, the difficulties may arise in constructing high-quality training data, since the RuWordNet thesaurus, in our opinion, the latter is far from complete.

The following is the contribution we made:

- It is tested how the use of various fields from the RuWordNet affects the result. For nouns it has been shown that adding **ruthes_name** to the string representation of synsets leads to better results, while adding **definition**, **lemma**, and **main_word** does not improve the performance. For verbs it has been shown that adding all possible fields is the best solution.
- The ranking is added to the baseline and synsets-synonyms, and their “parents” and “grandparents” are taken into account. This improvement is beneficial since we got a list of synsets-candidates sorted by relevance.
- It is shown that even without additional training competitive results can be achieved. That is, using only pre-trained distributive models and adding a few steps to the baseline, you can get competitive results.

¹¹ Table “Practice (NOUNS)” is [taken into account](#).

¹² Tables “Evaluation (NOUNS)” and “Post-Evaluation (NOUNS)” [are taken into account](#).

¹³ Table “Practice (VERBS)” is [taken into account](#).

¹⁴ Tables “Evaluation (VERBS)” and “Post-Evaluation (VERBS)” [are taken into account](#).

- We showed that `ft_cc_ru_300` achieves the best result on nouns (compared to other models from our work) and `tayga_lemmas_elmo_2048_2019`—on verbs.
- Python source code is available online¹⁵.

7. Acknowledgments

We express our gratitude to the RUSSE'2020 organizers for the chance to participate in an exciting shared task. We are grateful to the reviewers for careful reading of the manuscript and helpful remarks. The reported study was funded by RFBR according to the research projects №18-29-03187 and №18-37-20017 and with the support of the “RUDN University Program 5-100”.

References

1. *Alain Auger and Caroline Barrière*. Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, 14:1–19, 01 2008.
2. *Benko V., Zakharov V.* (2016). Very Large Russian Corpora: New Opportunities and New Challenges. In *Computational Linguistics and Intellectual Technologies*, pp. 79–93.
3. *Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-Chieh Shan*. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 23–32, Avignon, France, April 2012. Association for Computational Linguistics.
4. *Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov*. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
5. *Marti A. Hearst*. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*, 1992.
6. *Maria Karyaveva, Pavel Braslavski, and Yury Kiselev*. Extraction of hypernyms from dictionaries with a little help from word embeddings. In *Alexander Panchenko, Wil M. van der Aalst, Michael Khachay, Panos M. Pardalos, Vladimir Batagelj, Natalia Loukachevitch, Goran Glavač, Dmitry I. Ignatov, Sergei O. Kuznetsov, Olesia Koltsova, Irina A. Lomazova, Andrey V. Savchenko, Amedeo Napoli, and Marcello Pelillo*, editors, *Analysis of Images, Social Networks and Texts—7th International Conference, AIST 2018, Revised Selected Papers, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 76–87, Germany, 1 2018. Springer Verlag.
7. *Kiselev Y., Porshnev S., Mukhin M.* Method of extracting hyponym-hypernym relationships for nouns from definitions of explanatory dictionaries [metod izvlecheniya rodovidovykh otnosheniy mezhdu sushchestvitel'nymi iz opredeleniy

¹⁵ <https://github.com/vvyadrincev/taxonomy-enrichment>

- tolkovykh slovarey]. Software Engineering [Programmnaya inzheneriya]. 10, pp. 38–48, 2012.
8. *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts, pp. 320–332, 2015.
 9. *Andrey Kutuzov and Elizaveta Kuzmenko.* WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models, pp. 155–161. Springer International Publishing, Cham, 2017.
 10. *Kuratov, Y., Arkhipov, M.* (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213.
 11. *Roberto Navigli and Paola Velardi.* Learning word-class lattices for definition and hypernym extraction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1318–1327, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
 12. *Irina Nikishina, Varvara Logacheva, Alexander Panchenko, and Natalia Loukachevitch.* RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian Language. In Proceedings of the 26th International Conference on Computational Linguistics and Intellectual Technologies (Dialogue-2020). Moscow, Russia, 2020.
 13. *Stephen Roller, Douwe Kiela, and Maximilian Nickel.* Hearst patterns revisited: Automatic hypernym detection from large text corpora. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 358–363, Melbourne, Australia, July 2018. Association for Computational Linguistics.
 14. *K. Sabirova and A. Lukanin.* Automatic extraction of hypernyms and hyponyms from Russian texts. CEUR Workshop Proceedings, 1197:35–40, 01 2014.
 15. *Shavrina T., Shapovalova O.* To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. In Proceedings of “CORPORA2017”, Saint Petersburg, pp. 78–84, 2017.
 16. *Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller.* Learning to distinguish hypernyms and co-hyponyms. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 2249–2259, Dublin, Ireland, August 2014. Dublin City University and the Association for Computational Linguistics.

НАРЕЧИЕ ДАВНО ПО ДАННЫМ ЗВУЧАЩЕГО КОРПУСА¹

Янко Т. Е. (tanya_yanko@list.ru)

Институт языкознания РАН; Институт русского языка им. А. С. Пушкина, Москва, Россия

DOI: 10.28995/2075-7182-2020-19-773-783

RUSSIAN ADVERB *DAVNO* ,LONG AGO, FOR A LONG TIME' REVISITED FROM A CORPUS PERSPECTIVE

Yanko T. E. (tanya_yanko@list.ru)

Institute of Linguistics; Pushkin State Russian Language Institute, Moscow, Russia

During the last twenty years, the Russian adverb *davno* 'long ago, for a long time' was widely discussed in literature. It was recognized that the unique parameter of *davno* is its inability to be the theme of a sentence. Moreover, if *davno* functions in the context of aspectual forms relating to the past it can only be the rheme. In the context of the aspectual verbal forms relating to the past but preserving the connection with the moment of speech, *davno* can be either the rheme proper, or a component of the rheme. A classic example of an aspectual verb form referring to the past is the general factual meaning of the imperfective aspect. At present, the spoken data corpora can shed light on the communicative structure analysis, since the prosodic structure of the sound speech provides a straightforward access to the communicative structure. Novel parameters of *davno* are as follows. 1) Whereas *davno* is traditionally recognized as a word of rhematic polarity it can nevertheless function as a component of the theme in the context of attributive clauses and constructions (*Davno soglasovannyj visit dolzhen byl sostojatsja v aprele* 'A visit planned long ago would take place in April'). 2) The general factual meaning of the imperfective aspect, contrary to what was assumed before, is not an absolute prerequisite for *davno* to function as the rheme. The spoken corpus showed that in the context of negation and in the context of the verbs of speech, the general factual allows for *davno* to function as a component of the rheme but not the rheme proper (*Ja davno*

¹ Работа поддержана РФ (проект 18-18-00462 «Коммуникативно-синтаксический интерфейс: типология и грамматика», реализуемый в Государственном институте русского языка им. А. С. Пушкина)

tebja ne videl 'I have not been seeing you for a long time'; *My davno govoriš, čto naša zadacha — eto borba s terrorismom* 'We have been insisting for a long time that our main goal is the struggle against terrorism'). 3) A specific type of questions with the initial *davno* (as well as with other adverbs with the meaning of a considerable quantity like *často* 'often', *mnogo* 'much', and *daleko* 'far away') is singled out. Such questions cannot be unambiguously classified either as *yes-no*-questions or as *wh*-questions (*I davno vy zdesj stoite?* 'And how long are you staying here?'). A description of unique prosody of such questions is given. 4) In the context of discourse continuity, *davno* acquires the rising prosody which is in fact uncharacteristic of a word, which is unable to be the theme (*Xotel eto sdelat' davno, no teperj sdelaju točno* 'I wished to do it long ago, but now I will do it for sure'). The rising tone is accounted for by the meaning of continuity, which has the same prosody as the theme. 5) In constructions *kogda-to davno* 'once upon a time', *ochenj davno* 'very long ago', *davno-davno* 'very long ago', *davnym-davno* 'very long ago', *dovoljno davno* 'quite long ago', *ne tak davno* 'not so long ago' *davno* loses its rhematic polarity.

The parameters of *davno* are exemplified by spoken fragments taken from the Multimodal corpus of the Russian National corpus, and the minor working collection of the Russian speech recordings specifically set up for this investigation. The software program Praat was used in the process of analyzing the sound data.

Key-words: Russian adverb *davno* 'long ago, for a long time', theme, rheme, prosody, sound corpora, spoken speech, rhematic polarity, general factual, aspect, Russian

Наречие *давно* в течение последних более, чем двадцати лет, неоднократно служило предметом внимания лингвистов [Apresjan 1988], [Paducheva 1997], [Yanko 2001: 255–269]. Основным параметром, характерным для *давно*, была признана неспособность *давно* служить темой предложения. В работе [Paducheva 1997] было указано, что *давно* служит собственно ремой или компонентом ремы предложения, а в контексте общефактического значения несовершенного вида *давно* служит ремой: *Папа покупал эти часы давно*. Собственно рема — ударное слово ремы — здесь и далее выделяется полужирным шрифтом. В дальнейшем было высказано предположение, что *давно* служит ремой не только в контексте общефактического, но и в других видовременных контекстах, где связь прошлого события с моментом речи утрачена [Yanko 2001: 255–269]: *Случилось это давно; Сам я в Крыму давно был, ничего не могу советовать* [НКРЯ]. Идея о том, что новое, странное, далекое от говорящего и неожиданное склонно к роли ремы, и двадцать-тридцать лет назад не была новой. Словам с фиксированной коммуникативной функцией в предложении посвящена большая литература, см., например, [Ducrot 1973] [Apresjan 1988], [Bulygina, Shmelev 1990], [Partee 1994], [Dryer 1996], [Boguslavsky 1996: 63]. Тогда же были выделены и универсальные «возмущающие» контексты² — контраст и коммуникативная нерасчлененность — которые снимают коммуникативные и просодические запреты,

² О «возмущающих контекстах» см. [Yanko: 2001: 236–270, 323, 338].

накладываемые на *давно* (а также и на другие слова т.н. рематической полярности, например, *мало* и *редко*): *не моя Аннет, давно уж не моя/, а косноязычного этого Засядского* [НКРЯ] (контраст *моя* vs. *Засядского*); — *Полу вас какой-то щербатый*. — *Красили давно*. В последнем примере предложение *красили давно* служит цельным нерасчлененным обоснованием того, почему в доме неровный пол. В нерасчлененных ремах в соответствии с выделенными принципами выбора акцентоносителя [Yanko 2001: 71–73] акцентоносителем служит не *давно*, ср. также: *Огурцы в этом году у нас горькие: дождь\ был давно*.

В связи с используемым здесь понятием рематической полярности поясним, что под ним понимается свойство *давно* функционировать преимущественно в роли ремы предложения, но не во всех контекстах, а с известными — объяснимыми — отклонениями. Соответственно, в задачи этой работы входит обоснование свойства *давно* служить ремой, а также фиксация и обсуждение контекстов, говорящих о том, что это свойство не является абсолютным. Анализ показывает, что с расширением эмпирической базы исследования диапазон таких контекстов расширяется.

В настоящее время в связи с разработкой звучащих корпусов появилась возможность проверки на практике, какую в действительности коммуникативную роль играет в предложении *давно*, ибо просодия, которая служит основным средством выражения коммуникативной структуры, дает к коммуникативной структуре прямой доступ.

Забегая вперед, можно утверждать, что в целом гипотеза относительно рематической полярности *давно* при корпусном анализе звучащей речи подтверждается, но вместе с тем в описании, предложенное ранее, становится возможным внести существенные уточнения.

В результате корпусного анализа обнаружены следующие не известные ранее свойства *давно*. 1) *Давно*, традиционно считавшееся словом рематической полярности, способно быть компонентом темы в структуре атрибутивных конструкций (*Давно согласованный визит должен был состояться в апреле*). 2) Общефактическое значение несовершенного вида, вопреки тому, что считалось ранее, не служит абсолютным условием ремастичности *давно*. Анализ корпуса показал, что при отрицании и в контексте глаголов речи с сентенциальным дополнением общефактическое значение допускает использование *давно* в качестве компонента ремы, но не собственно ремы (*Я давно тебя не видел\; Мы давно говорили/, что наша задача — это борьба с терроризмом*). 3) Выделяется класс вопросов с начальным *давно* (а также других наречий со значением большого количества *часто, много, далеко, долго*), которые нельзя однозначно отнести ни к классу *да-нет*-вопросов, ни к классу вопросов с вопросительным словом (ср. *И давно вы здесь стоите?*). Предложено описание уникальной просодии и прагматики вопросов подобного типа. 4) В контексте дискурсивной незавершенности *давно* приобретает нехарактерную для него как для слова ремастической полярности восходящую просодию (*Хотел это сделать давно/, но теперь сделаю точно*). 5) В составе конструкций *когда-то давно, очень давно, не так давно, давно-давно, довольно давно* ремастическая полярность *давно* утрачивается.

Для анализа разработан малый рабочий корпус записей на основе Мультимедийного подкорпуса НКРЯ общим звучанием 60 минут. В работе использована машинная система анализа устной речи Praat [Boersma, Weenink 2020]. Записи снабжены транскриптами с просодической разметкой. Для разметки используется нотация, которая восходит к традиции С. В. Кодзасова, см., например, [Kodzsov, Bonch-Osmolovskaja, Zaharov, Kobozeva] и, одновременно, имеет соответствия с классификацией интонационных конструкций ИК Е. А. Брызгуновой [Bryzgunova 1982a]. Это следующие обозначения.

- I. \ — падение частоты тона типа ИК-1 с понижением на ударном слоге словоформы-акцентоносителя и дальнейшим понижением или ровным низким тоном на заударных слогах, если они есть.
- II. / — подъем частоты типа ИК-3 на ударном слоге акцентоносителя и падение на заударных, если они есть.
- III. \ / — падение или ровный низкий тон на ударном слоге акцентоносителя плюс подъем на заударных, если они есть, типа ИК-4. Если заударных нет, единое нисходяще-восходящее движение тона фиксируется на ударном слоге носителя.
- IV. /- — подъем частоты на ударном слоге плюс ровные или слабо нисходящие заударные типа ИК-6.

В примерах ниже знак акцента устанавливается после словоформы-акцентоносителя: *давно/-*.

1. *Давно* в структуре темы³

Корпусный анализ показал, что контекст препозитивных причастных и атрибутивных оборотов, а также придаточных предложений позволяет *давно* выступать в роли компонента темы:

- (1) *Давно забытые межнациональные противоречия/ вновь начали выходить на авансцену европейской **политики*** \ [НКРЯ].

В примере (1) *давно* артикулируется безакцентно, т. е. не несет на себе никаких коммуникативно релевантных акцентов, и входит в состав темы *Давно забытые межнациональные противоречия* с акцентоносителем темы — словоформой *противоречия*. Рему предложения составляет глагольная группа *вновь начали выходить на авансцену европейской **политики*** с акцентоносителем ремы — словоформой *политики*. Пример (1) и другие говорят о том, что *давно* способно быть компонентом темы (но все-таки не собственно темой).

³ В предложении *Давно/- не записывал я ничего в эту книжонку* \ [НКРЯ] наречие *давно* служит не темой, как это могло бы показаться, а компонентом ремы. То, что *давно* — не тема, доказывается тем, что *давно* здесь не может нести прототипического акцента темы ИК-3. На *давно* реализуется акцент ИК-6, который свидетельствует о том, что *давно* — это компонент ремы, но не тема.

2. *Давно* в контексте общефактического значения несовершенного вида

Контексты *давно* плюс общефактическое в рабочем массиве представлены следующими основными классами: 1) повествовательные предложения без отрицания с глаголами разнообразной семантики, кроме глаголов речи; 2) *да-нет*-вопросы; 3) предложения с отрицанием; 4) предложения с глаголами речи и сентенциальным дополнением.

Пример (2) и другие аналогичные примеры, которые иллюстрируют контекст общефактического значения с глаголом без отрицания и в отсутствие сентенциального дополнения, полностью соответствуют положению о рематичности *давно*:

- (2) <— А это дяди Шурика сапожки! — Какого Шурика? — Какого-какого.>
Какой у нас *давно* \ ночевал [НКРЯ].

Далее. Обратимся к *да-нет*-вопросам с *давно*. Неизвестное *да-нет*-вопроса мы с некоторой долей условности понимаем как ремоподобный компонент коммуникативной структуры предложения. Отличие между ремой, формирующей иллокутивную силу сообщения, и неизвестным вопросом, формирующего иллокутивную силу вопроса, безусловно, есть, но в данном случае мы не считаем это отличие принципиальным: и то, и другое служит акцентоносителем собственного иллокутивного компонента речевого акта определенного типа. Соответственно, *давно* в роли неизвестного *да-нет*-вопроса — это также релевантный контекст для настоящего анализа. В *да-нет*-вопросах представлены глаголы разнообразной семантики:

- (3) — Вы в белье *давно* / заглядывали?; Ты ей памперсы *давно* / менял?;
<Посмотри, во что ты превратилась!> Взвешивалась *давно*? [НКРЯ].

Да-нет-вопросы также поддерживают гипотезу о рематичности *давно* в контексте общефактического.

В предложениях с отрицанием также представлены разнообразные глаголы:

- (4) <Так мы ж договорились ушами... Форму потерял.> *Давно* не *шевелил*;
Давно мы не сталкивались с гвардейцами *кардинала*! [НКРЯ].

Примеры (4), в противовес тому, что предполагалось для *давно* в контексте общефактического ранее, говорят о том, что «общефактическое плюс отрицание» не ведет к обязательной рематичности *давно*. Это наблюдение требует объяснения. Действительно, если исходить из того, что к рематичности *давно* ведет не собственно общефактическое, а семантика удаленности релевантного положения дел от точки отсчета и отсутствие актуальных результатов в настоящем, то это условие в контекстах с отрицанием может не выполняться. При отрицании ситуация концептуализуется не как давность произошедшего события, а как то, что в течение длительного времени подобное событие не имело места:

- (5) *Сынок, чего-то я давно не видел твой дневник* [НКРЯ].

Отец долго не видел дневника сына и, возможно, не видит его и в момент речи.

Существенно, что интерпретация давности события как отсутствие его новых инстанций (просмотра дневника сына, шевеления ушами и стычек с гвардейцами кардинала) возможна только для тех ситуаций, которые понимаются как способные к повторению, дискретные (по М. Я. Гловинской [1982: 134]). Так, предложение *?Давно папа не покупал эти часы* (с конкретно референтными часами) может осмысленно интерпретироваться только в семантически и прагматически нестандартном контексте, когда папа, скажем, будучи ювелиром, регулярно скупающим и продающим старые часы, получает удивительную возможность получать во владение одни и те же часы в результате различных закупок. Последний пример показывает, что не все онтологические ситуации способны к естественному повторению и замене сообщений со значением 'Р было давно' на 'Р давно не было'.

Таким образом, русским языком для указания на временную удаленность события разработана замещающая концептуальная модель, которая может быть реализована не во всех контекстах, но там, где она возможна, она, вместо того, чтобы сигнализировать о давнем событии, указывает на то, что в течение длительного времени подобного события не было. Между тем, для коммуникативной структуры замена концептуальной модели предложения не проходит бесследно: отсутствие события понимается как особое положение дел, которое в отличие от абсолютной давности события, способно *восстанавливать* связь положения дел (теперь уже не самого события, а его отсутствия) с точкой отсчета (моментом речи). Соответственно, рематичность *давно* в таком контексте может нивелироваться. Как показывают данные корпуса, модель с отрицанием в повествовательном предложении с *давно* более частотна, чем модель без отрицания.

Перейдем к повествовательным предложениям с глаголами речи в общефактическом значении (*говорить, предлагать, объяснять*). Во всех доступных нам корпусных примерах *давно* воплощено в виде ремы:

- (6) *Я же **давно** говорил, что... возникнет и виртуальный брак и виртуальные дети; Она мне **давно** про тебя разъясняла, что ты в Аглаю Епанчину как кошка влюблён; Я ей **давно** говорила: не бросишь его, проклянц; Я **давно** говорил: попугаи у нас не приживутся* [НКРЯ].

Между тем экспериментальный перенос рематического акцента с *давно* на другую словоформу оказывается возможным. Сравним: *Я **давно** говорил, что попугаи у нас не приживутся* vs. *Я давно говорил, что попугаи у нас не **приживутся***. Следовательно, рематичность *давно* в предложениях глаголами передачи информации и сентенциальным дополнением тоже не абсолютна.

Это свойство предложений с глаголами передачи информации мы объясняем следующим образом. Предложения с глаголами передачи информации имеют импликацию сохранения информации в сознании получателя после того, как акт передачи состоялся. Если А сказал В, что Р, то с известной вероятностью В будет знать, что Р. Сохранение информации в сознании получателя не препятствует тому, чтобы А говорил В о Р вновь и вновь: *много раз говорил, что Р, не раз говорил, всегда говорил*. Однако, итерация акта говорения связана не с сообщением действительно новой информации, а с повторением того, что уже известно, скажем, с целью напомнить, заставить лишний раз задуматься.

Таким образом, то, что контекст глаголов речи в общефактическом значении благоприятствует рематичности *давно*, мы объясняем тем, что, если факт говорения имел место по крайней мере один раз и состоялся давно, это может приводить к рематичности *давно*. Здесь рематичность связана с ассертивной частью значения, а именно, с локутивным актом. Между тем рематичность *давно* здесь не абсолютна, и при трансформационном анализе большинство имеющихся примеров допускает перенос акцента с *давно* на глагольную группу. Возможную утрату рематичности *давно* мы объясняем имплицатурой, состоящей в том, что локутивный акт достиг своей цели: слушающему уже известно, что Р. Это создает эффект сохранения положения дел в точке отсчета и отменяет обязательную рематичность *давно*.

Приведем пример переноса рематического акцента с *давно* на другую словоформу в предложении с глаголом речи. Так, в предложении (7) из речи министра иностранных дел *давно* играет роль ремы:

- (7) ...мы **давно** \ говорили, что искоренение терроризма... является ... более важной задачей, чем смена режима [НКРЯ].

В записи звучащего примера нисходящий рематический акцент приходится на *давно*. Сообщение понимается так, что идея о борьбе с терроризмом высказана уже давно и существенна давность ее возникновения. Но, поскольку эта идея актуальна до момента речи, *давно* может уступать роль ремы глагольной группе. Акцент ремы может покинуть *давно* и переместиться на дополнение. Акцентирующей ремы становится словоформа *режима*: ...мы *давно* говорили/, что искоренение терроризма... является ... более важной задачей, чем смена **режима** \.

Итак, анализ корпуса показал, что контекст общефактического не служит абсолютным условием рематичности *давно*. Общефактическое значение глаголов с отрицанием и глаголов передачи информации допускает утрату *давно* рематической функции. Это объясняется семантикой сохранения некоторого положения дел, начиная от удаленного момента в прошлом вплоть до точки отсчета. Таким образом, условие рематичности *давно* следует формулировать не в терминах общефактического значения, а в терминах временной удаленности события и его результатов от точки отсчета.

3. Вопросы с начальным *давно*

Обращает на себя внимание особый тип вопросительных предложений с *давно* и с другими параметрическими наречиями, обозначающими большое значение параметра: *И давно ты на фронте?*; *И много он с вас берёт?*; *Часто выигрывали?* С формальной и функциональной точки зрения вопросы с *давно*, *много* и *часто* могут представлять собой и 1) *да-нет*-вопросы, которые маркируются акцентом типа ИК-3 на акцентоносителе вопроса, как (8), и быть подобны 2) вопросам с вопросительным словом с акцентами ИК-4 или ИК-6, которые необязательно фиксируются на *давно*, как (9):

- (8) — **Давно**/ вы знаете Горбункова С. С.? [НКРЯ]),

(9) *И давно вдоветь/- изволите?* [НКРЯ] (\approx И сколько времени вы вдовеете/-?).

Вопросы типа (9) с *давно* в роли квазивопросительного слова, подобного вопросительному слову, например, *сколько*, ранее, насколько нам известно, специальному рассмотрению не подвергались.

Начнем обсуждение со способов маркирования вопроса в русском языке. Вопросы без вопросительного слова (*да-нет-вопросы*), маркируются с помощью вопросительной просодии, а именно: интонационной конструкцией ИК-3, которая фиксируется на акцентоносителе вопроса, как в примере (8) из «Бриллиантовой руки». Это стандартный *да-нет-вопрос*.

Вопросы же с вопросительным словом, или т. н. частные вопросы, маркируются с помощью вопросительных слов *что?*, *где?*, *когда?*, *какой?*. Имеется также тип общих вопросов, в которых иллюкутивная сила маркируется с помощью вопросительной частицы *ли*. В вопросах с вопросительными словами и вопросительной частицей просодия не служит средством выражения иллюкутивной силы, а играет лишь формирующую роль. Просодические средства оформления вопросов с вопросительным словом вариативны. В них могут использоваться интонационные конструкции ИК-1, ИК-2⁴, ИК-4 и ИК-6. Из найденных вопросов с начальными *давно*, *много*, *часто* (всего их около 50) три четверти примеров приходится на ИК-4 и ИК-6, распределенных примерно поровну. ИК-1 и ИК-2 используются периферийно, и обсуждение этих типов, равно, как и проблему выбора словоформы — акцентоносителя вопроса мы здесь оставляем в стороне. Это предмет для отдельного обсуждения и он не влияет на способность параметрических наречий *давно*, *много*, *часто*, *быстро* и *дорого* функционировать в роли квазивопросительных слов в вопросах типа вопросов с вопросительным словом. Обратимся к вопросу с ИК-4:

(10) *И давно вы здесь стоите\/?* [НКРЯ].

В онтологической структуре ситуаций, где реализуются вопросы с *давно*, *долго* и *много*, первоначально представлено некое исходное положение дел, о параметрах которого (когда, сколько и за какие деньги) задается вопрос. Союзы *И* и *А* (*И давно?*, *И много?*, *А дорого?*) в большинстве примеров связывают исходную ситуацию и текущий вопрос. Конструкция ИК-4, характерная, вообще говоря, для контрастных ситуаций сравнения и сопоставления (ср. *А ваш\/? где билет?*) также поддерживает прагматику положения дел, предполагающего не один, а два шага: 1) исходная ситуация и 2) вопрос о ее параметрах, ср. другие примеры:

(11) *И давно он объявился\/?; Много вас таких пахарей\/? было?; И давно вы тут сидите\/?; И давно ты у Клавы на посылках\/?* [НКРЯ].

Возникает вопрос, используется ли ИК-4 в аналогичной ситуации в канонических вопросах с вопросительным словом. Корпусный анализ говорит о том,

⁴ ИК-2 — падение частоты основного тона с рельефным падением частоты на ударном слоге словоформы-акцентоносителя, причем падению частоты, как правило, предшествует небольшой подъем тона в начале ударного слога [Bryzgunova 1982b: 97–122].

что в вопросах со стандартными вопросительными словами ИК-4 используется в тех же прагматических ситуациях, что и в вопросах с *давно* и *много*, а именно, когда исходное положение дел так же вырастает из предтекста или экстралингвистической ситуации. Между тем ИК-4 — это не дефолтная просодия, которая приписывается вопросу с вопросительным словом в грамматиках. Основным просодическим средством формирования вопроса с вопросительным словом считается ИК-2 [Bryzgunova 1982b: 397–399]. ИК-4 же используется в вопросах с явно выраженным переходом от исходной ситуации к результирующей, что — факультативно — подкрепляется союзами *И* и *А*:

(12) *И сколько вы за нее просите\/?* [НКРЯ]

Вопрос (12) задается в ситуации, когда предмет торга налицо и остается только узнать его цену. Таким образом, просодическое оформление в вопросах с *давно* и *сколько* в определенном контексте совпадает.

Обратимся к вопросам с ИК-6. В них наблюдается картина, аналогичная той, которая зафиксирована для ИК-4. В примерах (13) (ср. пример (9)) также имеется исходная ситуация, о параметрах которой говорящий задает вопрос:

(13) *И много денег/- должен?; И часто выигрывали/-?* [НКРЯ]

В канонических вопросах с вопросительным словом ИК-6 также используется:

(14) *Ну и сколько он так под водой/- провисел?* [НКРЯ].

Между ИК-4 и ИК-6 в рассмотренных типах вопросов имеется тонкое различие. Оно состоит в следующем. Вопросы с ИК-4 либо маркированно вежливы (как (12)), либо, наоборот, звучат с вызовом, как (10) и (11): вызов содержит вопрос или любезность — выражается тембром голоса. Вопросы же с ИК-6 говорят о немаркированном отношении говорящего к собеседнику.

Рассмотренные здесь вопросы с *давно* имеют сходство с вопросами с вопросительными словами *как* и *сколь* плюс *давно* (*как давно?* и *сколь долго?*). Анализ сходств и различий этих типов — с *давно?* и с *как давно?* — это тоже актуальная задача, но здесь мы ее оставляем в стороне, так как сопоставление просодии в вопросах с *давно* и *как давно* представляет собой предмет отдельного исследования.

Итак, наречия большой степени признака *много*, *далеко*, *быстро*, *долго*, *дорого* и другие образуют особый лексико-семантический класс, члены которого способны играть роль вопросительного слова в вопросах, подобных вопросам с вопросительным словом (ср. *далеко? ≈ где?*; *давно? ≈ сколько времени?*; *дорого? ≈ сколько стоит?*). Эти наречия формируют вопросы о параметрах ситуации, известной из контекста. Соответственно, логико-семантическая структура контекста, включающего вопрос с наречием, быть может, экстралингвистического, состоит из а) блока, который соответствует раскрытию ситуации, и б) вопроса о ее параметре. Структура контекста объясняет присутствие союзов *И* и *А*, которые, как правило, присоединяются к вопросу (*И давно?*; *И дорого?*; *А долго?*). Вопросы с наречиями могут а) формировать стандартный *да-нет-вопрос* с акцентом ИК-3 на наречии; б) формировать *вопрос типа вопроса*

с вопросительным словом с акцентами ИК-4 или ИК-6 на словоформе — акцентоносители вопроса. Отличие между вопросами с ИК-4 и ИК-6 состоит в том, что вопросы с ИК-4 выражают бóльшую степень вежливости или, наоборот, вызова, и заинтересованности говорящего в слушающем, чем вопрос с ИК-6.

4. Давно в контексте дискурсивной незавершенности

Анализ корпусных данных говорит о том, что в структуре дискурса *давно* может вступать в отношения незавершенности, основное средство выражения которой — восходящий акцент ИК-3 (а также в случаях, заслуживающих отдельного обсуждения, которое мы здесь за недостатком места опускаем, ИК-4 и ИК-6). Если говорящий выражает свою мысль как имеющую продолжение, наречие *давно* может получать нехарактерную для него как для слова рематической полярности восходящую просодию:

(15) ... я и постарше Вас, и плюс всё-таки я уже **давно**/ во власти, я помню девяностые годы; Это было уже известно **давно**/, что количество солнечных пятен... меняется от года к году; Но это было **давно**/, года четыре-пять-шесть назад; ...я Вас знаю не близко, но знаю **давно**/ и не я один считаю Вас очень умным человеком [НКРЯ].

Примеры (15) демонстрируют, что в контексте незавершенности восходящий акцент на *давно* говорит не об утрате рематической полярности, а о вхождении составляющей с *давно* в фигуру незавершенности: просодические средства выражения темы и незавершенности совпадают.

5. Давно в составе конструкций

В конструкциях и словах, производных от *давно*, рематическая полярность, может сниматься: *когда-то давно, как-то давно, очень давно, давным-давно, давно-давно, уже давно, не так давно*. Из производных рематичность сохраняется только у *давненько* и *давнехонько*. На то, что *давным-давно*, в отличие от *давно*, может быть не только ремой, но и темой, впервые обратил внимание Ю. Д. Апресян еще в 1988 году [Аpresjan 1988]. Утрата производными от *давно* рематичности может быть связана с отрицанием (*не так давно, не очень давно*), а также с семантическим фактором, отменяющим рематичность *давно*. Обоснование коммуникативных различий между *давно* и *когда-то давно* (*давным-давно, очень давно*) состоит в том, что производные от *давно* вносят в семантику предложения значение существования: 'Было так, что Р. Р было давно'. Квантификация существования нейтрализует обязательную рематичность *давно* (ср. пару *редко* vs. *изредка* с тем же соотношением коммуникативных ролей и структурой толкования значений) [Yanko 2001: 265].

Корпусный анализ звучащего массива, включающего примеры с наречием *давно*, подтверждает гипотезу о рематической полярности *давно* и, одновременно, обосновывает не учтенные ранее в описаниях *давно* контексты, в которых *давно* либо не служит ремой, либо не несет характерного для ремы нисходящего акцента. Кроме того, анализ показывает, что *давно* входит в особый лексико-семантический класс наречий (*много, далеко, быстро, долго, дорого*), которые способны образовывать вопросы, подобные вопросам с вопросительным словом. Предложено описание прагматических, структурно-дискурсивных и просодических параметров таких вопросов.

Литература

1. *Apresjan Ju. D.* [Types of communicative information for explanatory dictionary] // *Jazyk: sistema i funkcionirovanie*. M., 1988. С. 10–22.
2. *Boguslavsky I. M.* Sfera dejstvija leksicheskikh edinic [The scope of lexical units]. M.: Jazyki russkoj kul'tury, 1996.
3. *Bulygina T. V., Shmelev A. D.* Quantifiers in the dictionary // Proceedings of the IV International Congress (EURALEX'90). Malaga. 1990.
4. *Boersma P., Weenink D.* (2020). Praat: Doing phonetics by computer. Version 6.1.08. Online: <http://www.praat.org/>.
5. *Bryzgunova E. A.* (1982a) Intonation [Intonatsiya], Russian Grammar [Russkaya grammatika]. Vol. 1, Nauka, Moscow, pp. 98–118.
6. *Bryzgunova E. A.* (1982b) Designating the unknown in questions (interaction of lexical, contextual, and grammatical means) [Sredstva vyrazhenija neizvestnogo v voprose (vzaimodejstvie leksiki, konteksta i grammatiki)], Russian Grammar [Russkaya grammatika]. Vol. 2, Nauka, Moscow, pp. 397–399.
7. *Dryer M. S.* Focus, pragmatic presupposition, and activated propositions // *Journal of Pragmatics* 26. 1996.
8. *Ducrot O.* French *peu* and *un peu*: a semantic study // *Generative Grammar in Europe*. Dordrecht. 1973.
9. *Glovinskaya M. Ja.* Semanticheskie tipy vidovogo protivopostavlenija russkogo glagola [Semantic types of aspectual correlation for the Russian verbs]. M., 1982.
10. *Kodzasov S. V., Bonch-Osmolovskaja A. A., Zaharov L. M., Kobozeva I. M., Krivnova O. F.* (2005) Data Base 'Intonation of Russian Dialogue: Interrogative Phrases' [Baza dannyh «Intonacija russkogo dialoga»: voprositel'nye repliki] // Proceedings of the International Conference "Dialog 2005". P. 245–247.
11. *Paducheva E. V.* Davno 'long ago' and dolgo 'for a long time' // *Logicheskij analiz jazyka: Jazyk i vremja*. M.: Indrik, 1997.
12. *Partee B. H.* Focus, Quantification, and Semantics-Pragmatics Issues, Preliminary Version // *Focus and Natural Language Processing*. Vol. 2: Semantics. IBM Heidelberg. 1994.
13. *Yanko T. E.* Kommunikativnye strategii russkoj rechi [The communicative strategies of the Russian speech]. Moscow, Yazyki slavyanskoi kul'tury Publ., 2001.

РУССКОЕ КАК БЫ: СЕМАНТИКА, ПРАГМАТИКА, ДИАХРОНИЯ¹

Зализняк Анна А. (anna.zalizniak@gmail.com)

Институт языкознания РАН, Институт проблем информатики ФИЦ ИУ РАН, Москва

DOI: 10.28995/2075-7182-2020-19-784-794

THE RUSSIAN *KAK BY*: SEMANTICS, PRAGMATICS, AND DIACHRONY

Zalizniak Anna A. (anna.zalizniak@gmail.com)

Institute of Linguistics of the RAS; Institute of Informatics Problems of the FRC CSC RAS, Moscow, Russia

The article considers the semantics of the Russian word *kak by*. It demonstrates that there are three main types of use of this word that are relevant for the modern Russian language: 1) as an approximation indicator, i. e. the marker of an approximative, indirect or metaphorical use of the linguistic unit it introduces (cf. *léd na reke sluzhil kak by mostom* ‘ice on the river served as a kind of bridge’; *on kak by veduschij specialist v dannoj oblasti* ‘he is sort of leading specialist in this field’); 2) as an indicator of epistemic indefiniteness (cf. *infljatsii kak by net* ‘there is <*kak by*> no inflation’); 3) as an illocutionary operator (“illocutionary mitigator”), mitigating the illocutionary force of the assertive speech act (cf. *Ja kak by ispolnitel’nyj direktor kompanii* ‘I am <*kak by*> the chief executive officer of the company’, uttered by the actual CEO of the company). We suggest that the initial meaning of *kak by* is that of a marker of descriptive indefiniteness (in an outdated use after the verbs of fuzzy perception), which has served as a source for both the approximation meaning, which is the main function of this word in contemporary Russian and that of epistemic indefiniteness. In its function as an “illocutionary mitigator” that emerged at the very end of the 20th century in the course of pragmaticalisation, the word *kak by* belongs to the class of discourse markers that ensure the success of a communicative act.

The study was based on the Russian National Corpus (www.ruscorpora.ru), including its oral and parallel subcorpora.

Key words: Russian language, semantics, metaphor, approximation, discourse markers, pragmaticalisation, illocutionary operator

¹ Статья написана при частичной финансовой поддержке РФФИ, грант №19-012-00505.

Русское слово *как бы*² заслуживает внимания сразу в нескольких отношениях. Прежде всего, оно заключает в себе специфическую концептуальную конфигурацию, которая уже анализировалась исследователями (ср. [Арутюнова 1996], [1997], [Разлогова 2004], [Летучий 2008], [Урысон 2009], [Benigni 2014]), однако, как представляется, вопрос нельзя считать решенным. В частности, слово *как бы* является в высокой степени лингвоспецифичным: в отличие от близкого по значению *как будто*, оно не имеет точного аналога ни в одном из основных европейских языков³. Анализ текстов параллельных корпусов НКРЯ показал, что в качестве «стимулов» и «моделей» перевода чаще всего используются выражения со значением 'как', 'в качестве', 'как если бы', 'почти', 'нечто вроде', 'так сказать', а также глагол 'казаться'; при этом приблизительно в половине случаев русское *как бы* остается без перевода или появляется «как бы ниоткуда» в переводе на русский язык. Отдельный интерес представляет факт необычайного распространения этого слова в 90-е — 2000-е годы, которое сопровождалось мощной волной метаязыковой рефлексии, охватившей широкие слои говорящих на русском языке и оценивающих данное явление как свидетельство падения норм культуры речи, ср. в частности, [Васильев 2000], [Лаптева 2004], [Разлогова 2004: 141–145], [Урысон 2009]. Наиболее подробный анализ этого слова предложен в статье [Летучий 2008], которая содержит ряд точных наблюдений, однако не со всеми выводами автора можно согласиться.

Задача данной статьи состоит в том, чтобы предложить анализ актуальной и диахронической семантики слова *как бы* и, в частности, высказать аргументы «в защиту» того типа употребления, который характеризуется в [Лаптева 2004: 26] как «дань манере речевой уклончивости и жеманства». Исследование проводилось на материале основного корпуса НКРЯ, а также его параллельного и устного подкорпусов.

1. Аппроксиматор

Принято считать, что слово *как бы* в своем центральном значении выражает сравнение (образует сравнительную конструкцию), ср. [Пешковский 2001], [Черемисина 1976, РГ-80]; [Летучий 2008], [Урысон 2009]; В [МАС] это сравнение названо «условно-предположительным», в [РГ-80, II: 493] — «недостоверным». (При этом в В РГ-80 имеется в виду значение, иллюстрируемое примерами типа *Целую ночь она протрадала и проплакала, и проснувшись была*

² Имеется в виду цельная языковая единица, в отличие от сочетания местоименного наречия *как* с частицей *бы*, ср.: *Интересно, как бы он поступил на моем месте; Как бы нам не опоздать; Как бы он нас не заметил, Как бы мы ни старались, ничего не получится* и т.п.

³ Наибольшую близость к русскому *как бы* обнаруживают англ. *kind of* и *sort of*, а также разговорное нем. *quasi*, ср. *Um so mehr, wo Sie sich jetzt quasi als ein anderer fühlen, müssen Sie alles aufbieten, dem Richterspruch zu entgehen!* [Gustav Meyrink. Der Golem (1914)] — Тем более, что вы чувствуете себя *как бы* иным человеком, вы должны все сделать для того, чтобы вас не осудили! [Густав Майринк. Голем (Д. Выгодский, 1922)].

так худа и бледна, как бы (= как если бы была) после тяжкой болезни (Писем.), справедливо характеризуемыми как «устаревшие».) В статье [Лаптева 2004] различаются два значения — «литературное» как бы 1, «регламентированное соотношением его значения ‘как будто’ и семантики высказывания» (*весь день стоит как бы хрустальный; вдруг что-то как бы подхватило его и как бы бросило к ее ногам*) и «нелитературное», «нерегулируемое» как бы 2, парадоксальность которого, по мнению автора, определяется сочетаемостью со словами «четкого конкретного значения, которое нельзя ослабить и представить как приблизительное и которое не подвергается обозначению через сравнение», ср.: *Я сегодня как бы ходила целый день; Эта авиация не является как бы военной в прямом смысле, а она предназначена как бы для обороны; Он как бы инвалид, шесть лет в коляске* и т. п. В статье [Легучий 2008] как бы и как будто рассматриваются как образующие две различные сравнительные конструкции; конструкции типа *он как бы спал* рассматриваются как сравнительные, поскольку в них усматривается сравнение реальной ситуации с неким эталоном. Однако, как представляется, сравнение с неким эталоном происходит при любом присваивании имени фрагменту внеязыковой действительности, просто в одних случаях совпадение с эталоном бывает более, а в других — менее полным. Как бы маркирует неполноту этого совпадения.

Н. Д. Арутюнова помещает слово как бы в ряд «показателей кажимости» как *будто, будто, словно, точно*, но отмечает, что в этом ряду как бы занимает особое место — сближаясь с «показателями неопределенности» *как-то* и *какой-то*: «как бы в приведенных контекстах⁴ — знак неполного соответствия предиката обозначаемому им явлению. Такой вид варьирования помещает как бы на границе между модальной неопределенностью, предполагающей истинностную (вероятностную) оценку, и семантической (концептуальной) приблизительностью» [Арутюнова 1996: 68].

Так или иначе, значение слова как бы не исчерпывается сравнением, включая в свою семантику также идею тождества. Вообще-то тождество и подобие несовместимы: X может быть *похож* на Y только если X не является Y-ом⁵. Однако слово как бы именно совмещает эти две идеи (что и обеспечивает его лингвоспецифичность). Так, в примере (1) как бы означает не просто, что поведение молодого человека было *похоже* на приглашение, но что оно одновременно являлось приглашением⁶.

- (1) Поймав взгляд Иуды, молодой человек приветливо улыбнулся и призывно замахал руками, **как бы приглашая** бездомного к очагу. [Давид Маркиш. Статья Лютовым. Вольные фантазии из жизни писателя Исаака Бабеля // «Октябрь», 2001]

⁴ Имеются в виду контексты типа *как бы с трудом ворочал языком; в лице его выразилось как бы нечто болезненное*.

⁵ Так, один мой знакомый попал в весьма неловкое положение, сказав своим польским друзьям «Львов — очень красивый город, *похож на Польшу*».

⁶ Примеры со ссылкой в квадратных скобках взяты из НКРЯ (www.ruscorpora.ru).

Совмещение в языковом значении логически несовместимых идей тождества и подобия — это свойство метафоры, ср. одно из ее определений, предложенных Н. Д. Арутюновой [1983: 6]: «Метафора создается тем, что подобию придается вид тождества»⁷. Близость семантики *как бы* к когнитивному механизму метафоры обнаруживает себя также в тяготении этого слова к контекстам метафорического описания ситуации: слово *как бы* служит сигналом того, что вводимая им языковая единица употреблена в непрямом, в том числе — метафорическом, значении, ср.:

- (2) То ли потому, что у нас было слишком много общего в характерах, и с её уходом я **как бы** потерял часть себя? [Илья Петрусенко. Я вдыхаю ветер воли... // «Народное творчество», 2003]
- (3) — У нас только на валюту, — прохрипел он, раздражённо глядя из-под лохматых, **как бы** молью изъеденных, сивых бровей. [М. А. Булгаков. Мастер и Маргарита, часть 2 (1929–1940)]

Для *как бы*, вводящего метафорическую номинацию, характерен контекст употребления перед существительным в твор. падеже, ср.:

- (4) Не потому, наверное, начал я всерьез писать после переезда в эту квартиру, что не мог писать в другом каком-нибудь месте, а потому, еще раз, что этот переезд был **как бы** чертою, отделившей меня от уже кончавшейся юности [Алексей Макушинский. Город в долине (2012)]
- (5) Итак, символ выступает **как бы** конденсатором всех принципов знаковости и одновременно выводит за пределы знаковости. [Ю. М. Лотман. Символ в системе культуры (1982–1992)]

Использование предикатного слова в переносном значении часто служит «стимулирующим контекстом»⁸ для появления *как бы* при переводе на русский язык. Ср. следующие примеры из параллельных корпусов НКРЯ, где глагол в метафорическом переносном значении (выделено курсивом) передается в русском переводе конструкцией с *как бы*: (англ.) The face staring back at him was thin and washed out. — На него смотрело худое лицо с невыразительными, **как бы** стертými чертами; (фр.) Mais j'ai voulu qu'un moment encore ma vie reste en suspens à la terrasse de ce café — Но мне хотелось, чтобы моя жизнь еще на минуточку **как бы** остановилась в нерешительности на террасе этого кафе; (исп.) Para comprobarlo basta acercarse cualquier mañana hasta la Plaza del Mercado <...> у perderse por las callejuelas aldeañas, cuyos evocadores nombres y tiendas nos remiten a la vieja Valencia de los mercaderes. — <...> или затеряться в лабиринте улочек, названия которых, наполненные ароматом истории, и уютные магазинчики

⁷ Отметим, что в русском языке имеется еще одна близкая по семантике к *как бы* конструкция, также характеризующаяся «гибридизацией тождества и подобия» [Арутюнова 1983: 6], — это творительный сравнения, или «творительный метафорический»: летел стрелой, упал камнем, скакал козлом (ср. [Зализняк 2013: 452]).

⁸ О понятии «стимулирующего контекста» см. [Зализняк, Падучева 2019], [Микаэлян, Зализняк 2019].

как бы погружают нас в старинную, купеческую Валенсию; (нем.) <...> durch einen Sonnenuntergang *kosmisch beleuchtet* — озаренный **как бы** космическими лучами заходящего солнца.

Помимо метафоры, **как бы** может указывать на приблизительность, неточность названной характеристики или на то, что объект обладает данным признаком (относится к данной категории) не в полной мере; ср. характерное для конструкции с **как бы** последующее уточнение номинации в (6):

- (6) Любовь Пушкина к жене была **как бы** *довершением или, точнее, жизненным осуществлением* того взгляда на красоту, который проходит через всю его поэзию. [И. Ф. Анненский. Вторая книга отражений (1909)]
- (7) Зимой губа Калгалакша и другие губы замерзают, и во время отлива лёд становится **как бы** *мостом*, консольно висящим в воздухе над водой. [Николай Иванов. По Белому морю и таёжным озёрам на байдарках // «Наука и жизнь», 2009]

В этом случае **как бы** оказывается близко по значению к слову *своего рода*, ср. (8), (9), где можно было бы также употребить слово **как бы**; и наоборот, в примерах (4), (5), можно было бы употребить выражение *своего рода*.

- (8) А раз так, рождение ребёнка для многих *своего рода вопрос престижа* — у нас «всё как у людей», благополучнее некуда. [Мария Давыдова. Кто в доме хозяин? (2003) // «100% здоровья», 2003.01.15]
- (9) Это, правда, создаёт некоторые ограничения свободы исследовательской мысли — оказывается *своего рода «смирительной рубашкой»* для фантазии. [Е. А. Климов. Психология в XXI веке // «Вопросы психологии», 2003]

Заметим, что аналогичная функция имеется у кавычек (ср. [Зализняк 2013: 499]): употребляя слово в неконвенциональном переносном значении, мы можем в речи предварить его словом **как бы**, а на письме с той же целью поставить в кавычки, ср.: *Метафорический контекст как бы стимулирует* (или «стимулирует» в кавычках) *появление в переводе слова «как бы»*. Иными словами, **как бы** в устной речи может выступать в качестве эквивалента для знака кавычек (а также сопровождаться жестом, обозначающим кавычки).

Итак, при употреблении в контексте характеризующих предикатов, иллюстрируемом примерами (1)–(7), **как бы** выступает в роли **аппроксиматора**: говорящий при помощи **как бы** маркирует то обстоятельство, что выбранная им номинация не совсем точна; в том числе, потому что она использована в переносном смысле. В функции аппроксиматора **как бы** может относиться к слову любого грамматического разряда и семантического класса; особенно характерно употребление **как бы** с глаголами внутреннего состояния (точная идентификация которого в принципе затруднительна), при этом часто — в составе деепричастного оборота, ср.:

- (10) Пока я говорил, она подымалась с места и все более и более краснела; но вдруг **как бы испугалась** чего-то, какой-то черты, которую не надо бы пересказывать, и быстро перебила меня [Ф. М. Достоевский. Подросток (1875)]

- (11) — Приезжайте обедать ко мне, — решительно сказала Анна, **как бы** рассердившись на себя за свое смущение [Л. Н. Толстой. Анна Каренина (1878)]

Отметим, что конструкция с *как бы*-аппроксиматором появляется не позднее конца XVIII в., ср.:

- (12) <...>но положение сестры моей тягостнее было вдвое, потому что она не имела еще столь обширного круга знакомых и приятелей и оставалась **как бы брошенной на чужой стороне**. [И. М. Долгоруков. <...> (1788–1822)]
- (13) Словом, публика вся **как бы проснулась**, даже и дамы стали вмешиваться в судебные диспуты, рассуждать о законах, бредить о конституциях. [И. М. Долгоруков. <...> (1788–1822)]
- (14) Есть морское небольшое животное, называемое Nautilus, или корабле-образец, который <...> вытягивает из своей спины некоторый род природного паруса и по ветру **как бы едет на воде**. [С. С. Бобров. Содержание и примечания к «Херсониде» (1798–1804)]

2. Показатель эпистемической неопределенности

До сих пор речь шла о контекстах характеризующих предикатов, описывающих признаки, которые допускают градацию (неполноту реализации) или же могут быть поняты в переносном смысле. В контексте предикатов, не допускающих аппроксимацию или метафору *как бы* выражает эпистемическую неопределенность: то ли было, то ли не было. Более того, *как бы* может выступать в роли своего рода миропорождающего оператора, создавая особую альтернативную реальность; это значение появляется в контексте отрицания имевших место фактов, ср.⁹:

- (15) Она раскаивалась утром в том, что она сказала мужу, и желала только одного, чтоб эти слова были **как бы не сказаны**. [Л. Н. Толстой. Анна Каренина (1878)]
- (16) Про инфляцию правительство запретило говорить, чтобы её **как бы не было**. [Борис Немцов, Елена Трегубова. «Мы и правда не из разведки!!» // «Коммерсантъ-Власть», 2002]
- (17) Знаю такие моменты в своей жизни, *которых как бы не было*. [Светлана Алексиевич. Время second-hand // «Дружба народов», 2013]

Эпистемическая неопределенность может поляризоваться таким образом, что *как бы* становится «оператором фикциональности», указывая на то, что следующий за ним предикат описывает положение вещей, не соответствующее реальности; так, *как бы невзначай*, *как бы случайно* означает 'намеренно', 'не случайно'; ср. также:

⁹ Заметим, что сходную функцию имеет глагол *считать* [Зализняк 2013: 177], ср. пример (15) и *Считай что я это не говорил*; ср. также *Это не считается*.

(18) —А! — равнодушно сказала Анна, **как бы** мало интересуясь этим, — и продолжала улыбаясь: — Как может ваше общество компрометировать кого-нибудь? [Л. Н. Толстой. Анна Каренина (1878)]

Здесь *как бы* указывает на то, что Анна только делала вид, что мало интересуется этим, ср. франц. перевод: Ah! fit Anna jouant l'indifférence¹⁰ — букв. 'изображая равнодушие, притворяясь равнодушной'. Ср. также:

(19) За первым же углом Эраст Петрович отпустил извозчика и медленно, **как бы** прогуливаясь, зашагал в обратном направлении. [Борис Акунин. Азазель (1998)]

(20) Но зачем, скоро понял: потому что возили его **как бы** на расстрел. «Ну все, тебе конец», — приговаривали охранники, передергивая затворы. [Анна Политковская. Путинская Россия (2004)]

В итал. переводе примера (19) *как бы* прогуливаясь передано как *facendo vista di fare due passi* [Boris Akunin. La regina d'inverno (Pia Pera)] (т. е. 'делая вид, что прогуливается'). В примере (20): на самом деле, как выяснилось, расстреливать не собирались.

Вплоть до XIX в. *как бы* употреблялось также в функции подчинительного союза при глаголе восприятия, вводящем обозначение воображаемой ситуации при помощи самостоятельной предикации (т. е. содержащей собственное подлежащее и сказуемое), ср. (21)–(23); в современном языке здесь следовало бы употребить *будто бы* или *как будто*.

(21) А которые люди во дворах слышали, то ино слышелось им **как бы** Симановских колоколов звон звучит. [Вологодско-Пермская летопись (852–1538 гг.) (1550–1590)]

(22) На берегу по четырём сторонам озера стояли неописанной величины четыре истукана, которые как будто бы под тяжким бременем нагнулись и имели чрез спины на плечах железные цепи, и, казалось, **как бы** они тащили что-нибудь из воды. [М. Д. Чулков. Пересмешник, или Славенские сказки (1766–1768)]

(23) Казалось, **как бы** вместе с нею влетел солнечный луч в комнату, озаривши вдруг потолок, карниз и темные углы ее. [Н. В. Гоголь. Мертвые души (1842)]

При отсутствии во вводимом *как бы* фрагменте текста предикативного центра *как бы* теряет функцию союза, встраиваясь в бывшее главное предложение и превращаясь в аппроксиматор — в первую очередь, при глаголах нечеткого восприятия, ср.:

(24) Лицо его было бледно и сухо и при лунном свете казалось **как бы** мертвым; мутные, бродящие глаза его показывали, что голова его не в самом здоровом состоянии. [О. М. Сомов. Юродивый (1827)]

¹⁰ Henri Mongault. Anna Karenine. Paris: [Gallimard, 1994].

- (25) Руки же у него и лицо были все окровавлены, сам же казался **как бы помешанным**. [Ф. М. Достоевский. Братья Карамазовы (1880)]

Постепенность этого сдвига ясно видна из примера (26), где в одном и том же значении (а именно, для описания результата обработки перцептивных данных) во фрагменте (в) использована конструкция с **как бы** в функции подчинительного союза при глаголе *казаться*, в (а) — бессоюзная подчинительная конструкция с тем же глаголом, а в (б) — конструкция с **как бы** в функции аппроксиматора.

- (26) Когда она говорила, у ней, (а) *казалось, все стремилось* вослед за мыслью: выраженье лица, выраженье разговора, движенье рук, самые складки платья (б) **как бы стремились** в ту же сторону, и (в) *казалось, как бы она сама вот улетит* вослед за собственными ее словами. Ничего не было в ней утаенного. [Н. В. Гоголь. Мертвые души (1842)]

3. Иллокутивный оператор

Вернемся теперь к тому — наиболее интересному и специфичному — «нелитературному» типу употребления **как бы**, который был упомянут в начале; ср. (примеры из статьи [Урысон 2009]):

- (27) [Студентка обращается к преподавателю с просьбой перенести защиту курсовой работы. Мотивируя свою просьбу, она произносит]:
Я **как бы** беременна.
- (28) [аспирант, мнс, обращается к руководителю сектора] Юрий Трофимович, я тут **как бы** смету составил по гранту, Вы не посмотрите?
- (29) Я тут **как бы** на операцию ложусь. Ты не могла бы гулять с моей собакой на следующей неделе?
Ср. также примеры из статьи [Лаптева 2004]:
- (30) Я сегодня являюсь **как бы** президентом компании
[из интервью президента компании Славнефть]
- (31) Подлежащее и сказуемое — это **как бы** главные члены предложения
[из ответа студента на экзамене]

Такое употребление в 90-е годы было относительно новым¹¹; в частности, в своей статье 1996 г. Н. Д. Арутюнова пишет (с.69): «Нельзя **как бы написать письмо, как бы поступить в институт, как бы родить ребенка**». Т.е. в 96-м году так сказать было еще нельзя (точнее, сама Нина Давидовна так не говорила), сейчас — можно; ср. (32) и (33):

¹¹ См. [Левонтина 2016: 347–351] об эволюции социолингвистического статуса этого типа употребления **как бы** от «интеллигентского отказа от категоричности» до «простого жеманства».

(32) Он **как бы** ведущий специалист в данной области.

(33) Он **как бы** ведущий научный сотрудник нашего института.

В (32) употреблено «литературное» *как бы* — аппроксиматор, который вводит оценочное суждение; используя слово *как бы* говорящий вносит некоторую неопределенность в используемую характеристику. В (33) речь идет о факте, не подверженном аппроксимации: *ведущий научный сотрудник* — это название должности. *Как бы* здесь является дискурсивным словом, выполняющим чисто прагматическую функцию «hedge» (по [Lakoff 1973]), или **иллокутивного оператора**, а именно маркера «смягчения» иллокутивной силы утверждения¹². Русский эквивалент термина «hedge» — «перегородка» — представляется не совсем удачным, лучше подошло бы слово «смягчитель», или, если воспользоваться уже практикуемой в данной сфере физической метафорой, «амортизатор» иллокутивной силы. Когда я говорю *Я как бы ложусь в больницу, не можешь ли ты погулять с моей собакой?* (ср. пример (29)) дискурсивное слово *как бы* служит средством «не грузить» слушающего всеми обстоятельствами, вытекающими из сообщаемой информации (необходимостью проявлять участие, выражать сочувствие и т. д.), а сигналом, что слушающий должен принять эту информацию только в обозначенном дальше отношении: может ли он во время моего отсутствия погулять с моей собакой? То же самое — в примере (27): студентка не имеет в виду рассказывать преподавателю о своих личных обстоятельствах, ей нужно лишь перенести экзамен. Как справедливо замечено в [Урысон 2009], такое *как бы* сопровождается специфической просодией — говорящий *как бы* будто немножко извиняется.

Функция «амортизации» иллокутивной силы высказывания может совмещаться с функцией аппроксимации. Так, в примерах (34), (35) из устного подкорпуса НКРЯ употребление *как бы* служит своего рода извинением за слишком «пафосное» заявление, одновременно внося приближительность в используемую номинацию.

(34) И вот эта вот редакторская стезя / она **как бы** через всю жизнь у меня.

[На краю... (Владивосток). Д/ф из цикла «Письма из провинции» (ТК «Культура») // Т/к «Культура». Д/ф из цикла «Письма из провинции», 2009]

(35) Сын / это вот ну **как бы** плоть и кровь / правая рука моя. [Псковское братство. Д/ф из цикла «Письма из провинции» (ТК «Культура») // Т/к «Культура», 2009]

¹² Ср. понятие “conversational mitigation” в [Fraser 1980], где оно определяется как “reduction of certain unwelcome effects which a speech acts has on the hearer”.

4. Заключение

Итак, было выделено три основных типа употребления *как бы*:

- **показатель аппроксимации**, т.е. приблизительности номинации или употребления слова в каком-то непрямом смысле, в том числе метафорическом, значении (*он как бы ведущий специалист в данной области; лед на реке служил как бы мостом; я как бы потерял часть себя*);
- **показатель эпистемической неопределенности** (*инфляции как бы не существует; шел как бы прогуливаясь*);
- **иллокутивный оператор** («иллокутивный амортизатор»), смягчающий иллокутивную силу речевого акта утверждения (*Он как бы ведущий научный сотрудник нашего института*).

С точки зрения семантической деривации исходным для *как бы* является употребление в функции подчинительного союза при глаголах нечеткого восприятия, и, далее, в роли самостоятельного показателя не полностью отчетливого впечатления. Этот тип употребления послужил источником для значения эпистемической неопределенности (в контексте сообщения о факте), а также для значения аппроксимации (в контексте характеризующих предикатов), которое является основным для слова *как бы* в современном русском языке.

Функция «иллокутивного амортизатора» появилась у слова *как бы* в самом конце XX в. — в результате переноса признака неопределенности из дескриптивной области в коммуникативную (что вообще характерно для процессов прагматикализации); при этом оно подверглось определенному «семантическому выходу» (термин предложен в [Шмелев 2018]). Традиционно такие единицы называются «словами-паразитами», однако, как показано в [Шмелев 2018], «во всех таких случаях семантическая пустота языкового выражения оказывается в некотором смысле мнимой, поскольку оно наполняется богатым прагматическим содержанием»; такие слова являются неотъемлемой принадлежностью устной речи, обеспечивая успешность коммуникации (ср. также [Levontina, Shmelev 2007]).

Литература

1. Арутюнова Н. Д. (1983), Тожество или подобие? // Проблемы структурной лингвистики 1981. М., 1983.
2. Арутюнова Н. Д. (1996), Стиль Достоевского в рамке русской картины мира // Поэтика и стилистика. Язык и культура: Памяти Т. Г. Винокур. М., 1996. С. 61–89.
3. Арутюнова Н. Д. (1997), Модальные и семантические операторы // Облик слова. Сборник статей памяти Д. Н. Шмелева. М., 1997, с. 22–40.
4. Васильев А. Д. (2000), Слово в телеэфире. Очерки новейшего словоупотребления в российском телевидении. Красноярск, 2000.
5. Зализняк Анна А. (2013), Русская семантика в типологической перспективе. М.: Языки славянских культур, 2013.

6. *Зализняк Анна А., Падучева Е. В.* (2019), Русское что-то как дискурсивное слово. // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2019. М., 2019. С. 755–770.
7. *Лаптева О. А.* (2003), Самоорганизация движения языка: внутренние источники преобразований (статья первая) // ВЯ, 2003, №6.
8. *Лаптева О. А.* (2004), Самоорганизация движения языка: внутренние источники преобразований (статья вторая) // ВЯ, 2004, №5.
9. *Левонтина И. Б.* (2016), Русский со словарем. 2-е изд. М., 2016
10. *Летучий А. Б.* (2008), Конструкции сравнения ситуаций с показателями как бы и как будто. // НТИ, сер. 2, 2008, № 2. С. 19–27.
11. *Микаэлян И. Л., Зализняк Анна А.* (2019), Производные значения русского неопределенного наречия как-то: опыт корпусного анализа. // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2019. М., 2019. С. 448–461.
12. *Пешковский А. М.* (2001), Русский синтаксис в научном освещении. М., 2001.
13. *Разлогова Е. Э.* (2004), Логико-когнитивные и стилистические аспекты семантики модальных слов. М., 2004.
14. *Шмелев А. Д.* (2018), Вехи прагматикализации. (https://www.slavistik.uni-muenchen.de/forschung/tagungen/tagungsarchiv/konferenz-pragmatik/abstract_aleksej-shmelev.pdf)
15. *Черемисина М. В.* (1976), Сравнительные конструкции в русском языке. Новосибирск, 1976.
16. *Benigni V.* (2014), Strategie di approssimazione lessicale in russo e in italiano // L'architettura del testo. Studi contrastivi slavo-romanzi. — Alessandria: Edizioni Orso, 2014. — P. 203–224.
17. *Fraser B.* (1980), Conversational mitigation. // «Journal of pragmatics» 4, pp. 341–350.
18. *Lakoff G.* (1973), Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. // Journal of Philosophical Logic, Vol. 2, № 4, pp. 458–508.
19. *Levontina I. B., Shmelev A. D.* (2007), False Emptiness: Are So-called “Parasitical Words” Really Semantically Void? // Kim Gerdes, Tilmann Reuther, Leo Wanner (eds.). Meaning — Text Theory 2007. Proceedings of the 3rd International Conference on Meaning-Text Theory. Klagenfurt, May 20–24, 2007. Wiener Slavistischer Almanach, Sonderband 69, München — Wien 2007.

ZERO FORMS IN MORPHOLOGICAL PARADIGMS: THE VERB “BE” IN RUSSIAN¹

Zimmerling A. V. (fagraey64@hotmail.com)

Pushkin state Russian language institute; Institute of Linguistics, Russian Academy of science; Moscow pedagogical state university, Moscow, Russia

This paper offers a corpus analysis of the Russian verb *быть* ‘be’ which has an abnormal present tense paradigm including a zero form $\emptyset^{\text{BE.PRES}}$ and overt forms *есть*^{BE.PRES} and *суть*^{BE.PRES} which do not discriminate person and number and are distributed syntactically. I discuss different approaches to the grammar of *быть* and argue that Apresjan’s model which recognizes $\emptyset^{\text{BE.PRES}}$, *есть*^{BE.PRES} and *суть*^{BE.PRES} as parts of one and the same lemma is superior to alternative models splitting *быть* into two lemmas representing copula vs content verb ‘be’. The peripheral status of overt present BE-forms compared with $\emptyset^{\text{BE.PRES}}$ in the Russian National Corpus is confirmed by three measures: 1) dispersion of texts where a BE-form occurs; 2) uneven coverage in different persons and numbers; 3) ratio of copular uses vs content verb uses. 1–2 person present tense BE-forms attested in RNC are internal borrowings from Old Russian and Old Church Slavonic, while *есть*^{BE.PRES} and *суть*^{BE.PRES} are inherited 3rd person elements which take over 1–2 person uses. The historical 3PI *суть* is redundant in a system, where a more frequent 3rd person form *есть* is licensed in the plural: it survives by a minority of speakers either as an optional 3PI copula in formal discourse or as an emphatic copula in oral discourse. The form *есть*^{BE.PRES} occurs in all persons and numbers both as content verb and as copula but is underrepresented as 3PI copula: this gap is filled by $\emptyset^{\text{BE.PRES}}$. The frequency of the zero copula $\emptyset^{\text{BE.PRES}}$ can be measured in corpora without syntactic annotation on the basis of systemic proportion between present vs past tense uses of *быть* and on the basis of approximation samples for contexts where overt copulas alternate with $\emptyset^{\text{BE.PRES}}$.

Keywords: corpus linguistics, Russian, parametric grammar, morphology, agreement, lemmatization, copula, zero syntactic elements

DOI: 10.28995/2075-7182-2020-19-795-810

¹ This paper is written with financial support from the Russian science foundation, project RSCI 18-18-00462. I am grateful to the anonymous reviewers for the valuable comments.

НУЛЕВЫЕ ФОРМЫ В МОРФОЛОГИЧЕСКИХ ПАРАДИГМАХ: ГЛАГОЛ «БЫТЬ» В РУССКОМ ЯЗЫКЕ

Циммерлинг А. В. (fagraey64@hotmail.com)

Государственный институт русского языка им. А. С. Пушкина; Институт языкознания РАН; Московский педагогический государственный университет, Москва, Россия

1. The verb *быть* in Russian: grammar, lexicography and frequency

The Russian verb *быть* ‘be’ has an abnormal present tense paradigm consisting of 3 elements not distributed according to the principle of person-and-number agreement. A salient part of its uses is realized by the zero copula $\emptyset^{\text{BE.PRES}}$, which reduces the frequency of the overt present forms *есть* и *суть*. The lemma *быть* has a lower frequency than comparable lexemes in Standard Average European (SAE) languages. *Есть* and *суть* are historically linked with 3Sg and 3Pl respectively but their usage in Modern Russian does not follow these tags. It is generally acknowledged that *есть* spread over all persons and numbers. Basing on corpus data, I argue that *суть* underwent a similar development. A number of authors [Ščerba 1928]; [Jevgenjeva 1999] suggest that the distribution of $\emptyset^{\text{BE.PRES}} \sim \text{есть}$ follows the distinction of the copular ‘be’ vs content verb ‘be’. However, the theory that copular *быть* and content word *быть* are different lemmas must be rejected, since both $\emptyset^{\text{BE.PRES}}$ and *есть* are used both as copula and content verb. I measure the ratio of copular vs content verb for each person and number form and argue that the ratio of overt copular sentences gives a key to the part covered by $\emptyset^{\text{BE.PRES}}$. This study is based on Russian National Corpus (RNC). The method requires partial or complete syntactic analysis of contexts involving the present tense forms of *быть* in order to identify them as part of the existential, copular or perfect construction. Direct measurement is possible only for forms with the lowest frequency, in other cases I implement approximation samples based on the next-neighbor method: the adjacent elements often diagnose the type of *быть* construction without look-up of the entire syntactic structure. An advantage of the chosen approach is that it minimizes the role of the text meta-data in a balanced corpus.

1.1. БЫТЬ vs ЕСТЬ in Russian lexicography

Vladimir Dal’s (1880) dictionary claims that *есть* is the 3Sg form of the verb ‘be’, which is “dropped where other languages use it” [Dal 1880 I: 523]. The 3Pl *суть* is not mentioned. Dmitry Ušakov’s dictionary (1935) has two entries— *БЫТЬ* and *ЕСТЬ* in the first volume. The first one claims that *быть* “lacks present tense except

for the 3Sg *есть* and outdated 3Pl *суть* in some meanings” [Ušakov 1935 I: 214]. The second entry tells that *есть* is used in all persons “due to the loss of the old forms of the present tense of *быть*” [ibid., 838]. The fourth volume adds *СУТЬ* introduced as a “bookish and outdated 3Pl of *быть*, primarily used in contexts of enumeration” [Ušakov 1940 IV: 599]. This description hints that *суть* is optional but does not specify, whether it is a variant of *есть*.

Ušakov’s description is influenced by Lev Ščerba’s theory that copular *быть* and content verb *быть* are different lexemes [Ščerba 1928]. He starts listing the uses of *БЫТЬ* from contexts where “the copula is dropped in the present tense” as in the “position between a subject and a nominal predicate” and in the participle passive [Ušakov 1935 I: 218]. Ščerba’s program is implemented in the Minor Academic Dictionary (1957–1961) edited by Anastasia Jevgenjeva. Her description is close to Ušakov, but the entry *БЫТЬ* starts from contexts for the content verb. She claims that *БЫТЬ* lacks present forms “except for the 3Sg *есть* and the outdated 3Pl *суть*” [Jevgenjeva 1999 I: 130–131]. The entry *ЕСТЬ* however admits that *есть* is used in all persons and numbers “due to the loss of the present forms of *быть*” [ibid., 468]. The fourth volume has a short entry *СУТЬ*² defined as a bookish 3Pl form occasionally used in 3Sg [Jevgenjeva 1999 IV, 310]: this statement is based on examples like *Сие_{sg} не *суть*^{BE.PRES} угроза_{sg}* “This is not a threat” (M. Gorki, 1912), which lack an agreement controller in the plural form.

Andrej Zaliznjak’s Grammatical Dictionary tells that *есть* stands for all persons and numbers of *БЫТЬ*, while *суть* is a 3Pl form rooted in scientific or archaic discourse [Zaliznjak 1977: 133]. The loss of the present tense forms is not mentioned. Sergej Ožegov’s dictionary revised by Natalia Švedova is close to Ušakov but less consistent. The entry *БЫТЬ* states that this verb lacks present tense “except for 3Sg *есть* and outdated bookish 3Pl *суть*” [Ožegov, Švedova 1992: 64]. The entry *ЕСТЬ* tells that this form spread over all persons and numbers “due to the loss of the old present tense forms” [ibid., 191]. The entry *БЫТЬ* starts from content verb contexts, while the entry *ЕСТЬ* starts from copular contexts. The entry *СУТЬ* claims that this bookish form of 3Pl is now primarily used as a copula, if both arguments are expressed by substantives [ibid., 808].

The author of the most detailed lexicographical description of *БЫТЬ*, [Jurij Apresjan 1996] rejects Ščerba’s and Jevgenjeva’s theory on two separate BE-lemmas and reinstates one paradigm consisting of 3 present forms: $\emptyset^{\text{BE.PRES}}$, *есть* and *суть*. $\emptyset^{\text{BE.PRES}}$ and *есть* lack person-and-number features, while *суть* is an optional variant of the copular BE but not the content verb BE in 3Pl [Apresjan 1996: 518, 528]. Apresjan shows that both $\emptyset^{\text{BE.PRES}}$ and *есть* have parallel uses as a copula and as a content verb, so that the identification of the copular BE with the hypothetical lexeme selecting $\emptyset^{\text{BE.PRES}}$ in the present tense and the content BE with a different lexeme selecting *есть* is impossible. This description has three advantages: 1) it recognizes $\emptyset^{\text{BE.PRES}}$ as part of the paradigm; 2) it does not stick to historical notions; 3) it does not identify *есть* as a content verb in all its uses. I adopt Apresjan’s approach, but argue that his tag for *суть* must be fixed.

1.2. The verb ‘be’ in SAE languages and in Russian: frequency and grammar

The verb ‘be’ is a high frequent word in SAE languages, with a rank comparable to the ranks of the definite article and the conjunction ‘and’. The high rank of the SAE ‘BE’ in the top 5–10 lemmas is due to the fact that it is widely used in three types of contexts:

- Type I contexts: ‘BE’ as a content verb expressing a variety of existential, locative and possessive meanings;
- Type II contexts: ‘BE’ as a copula with different types of nominal predicates;
- Type III contexts: ‘BE’ as an auxiliary element in analytical verb forms and constructions like Germanic, Romance or Slavic BE-perfect or BE-progressive in English etc.²

In Slovenian [Gigafida], the lemma ‘BE’ heads the list of the most frequent lemmas. In both British English [BNC] and American English [COCA] the lemma ‘BE’ holds the 2nd rank after the definite article. In German it holds the 3rd rank. In Russian [RNC], the lemma ‘BE’ is only the 6th from above, behind *и* ‘and’, *не* ‘not’, *в* ‘in’, *на* ‘on’ and *я* ‘1Sg’ [Lyaševskaja, Šarov 2009]. This results from two deviant features of *быть*. For the first, overt present forms of *быть* lack person-and-number specification which is unusual for SAE languages: English retains full-fledged person-and-number agreement exactly where Russian gives it up— the present tense of *be*. For the second, the most frequent present form of *быть* is $\emptyset^{\text{BE.PRES}}$. The status of $\emptyset^{\text{BE.PRES}}$ as part of the *быть* paradigm in Russian is acknowledged in linguistic typology [Stassen 1994]; [Pustet 2003]. Frequency lists normally ignore zero forms, since taking them into account would require processing uniform syntactic annotations for a family of corpora. Overt present forms of *быть* do not match the frequencies of the non-present forms. *Есть* (393,200 raw hits in RNC) is almost 6 times less frequent than the past tense forms *был*, *была*, *было*, *были* (2,267,476 raw hits). This is predictable since the past tense forms of *быть* correspond both to \emptyset in Type II contexts and to *есть* in Type I contexts.

Tab. 1: Present vs non-present forms of *быть* in Russian

Type of context	Present tense of <i>быть</i>	Past tense of <i>быть</i>
Type I. The overt present form <i>есть</i> is optional or obligatory: <i>есть/был, -а, -о, -и</i> .	(1) а. У Ивана есть машина _{SG.F} . ‘John has a car.’	(1) б. У Ивана была _{SG.F} машина _{SG.F} . ‘John had a car.’
	(2) а. У Ивана есть книги _{PL} . ‘John has books.’	(2) б. У Ивана были _{PL} книги _{PL} . ‘John had books.’
	(3) а. Ты и есть доктор. ‘You _{2SG} are indeed a doctor.’	(3) б. Ты и был _{SG.M} доктором. ‘You _{2SG} were indeed a doctor.’

² Type III contexts must be kept apart from Type II contexts, since Type II sentences always refer to present events, while Type III sentences with present tense BE-auxiliaries in such complex verbal forms as perfect and plusperfect refer to past events.

Type of context	Present tense of <i>быть</i>	Past tense of <i>быть</i>
Type II. Overt present forms are excluded, the silent form \emptyset^{BE} . \emptyset^{BE} is obligatory: $\emptyset^{BE.PRES}$ /был, -а, -о, -и.	(4) а. Иван $\emptyset^{BE.PRES}$ умен _{SG.M} . 'John is intelligent.' (5) а. Ты $\emptyset^{BE.PRES}$ умен _{SG.M} . 'You _{2SG} are intelligent.' (6) а. Иван и Марья $\emptyset^{BE.PRES}$ умны _{PL} . 'John and Mary are intelligent.'	(4) б. Иван был _{SG.M} умен _{SG.M} . 'John was intelligent.' (5) б. Ты был _{SG.M} умен _{SG.M} . 'You _{2SG} were intelligent.' (6) б. Иван и Марья были _{PL} умны _{PL} . 'John and Mary were intelligent.'

Neither $\emptyset^{BE.PRES}$ nor *есть* discriminate number and gender, while past tense forms do. $\emptyset^{BE.PRES}$ and *есть* do not discriminate person either, cf. (3a) and (5a)³. A general prediction for SAE languages is that present forms of BE are more frequent than the non-present ones given that corpora display the same proportion of present and non-present events. This holds both for languages with person-and-number agreement (English, German, Slovenian⁴) and for languages with a single present form (Danish, Swedish). In Danish, the only present tense BE-form, Da. *er* heads the frequency list for all word forms, while in Swedish, the only present tense BE-form, Sw. *är* holds the third rank. Ru. *есть* with its 66th rank in the list of frequent word forms is far behind, which is due to the fact that overt present forms of BE are excluded from all Type II and Type III contexts:

- (7) а. Он $\emptyset^{BE.PRES}$ болен.
'He is ill.'
б. *Он **есть**^{BE.PRES} болен.
- (8) а. Он $\emptyset^{BE.PRES}$ арестован.
'He is arrested.'
б. *Он **есть**^{BE.PRES} арестован.

³ This feature however does not make a contrast with the past tense.

⁴ In Slovenian [Gigafida] the lemma *biti* 'be' heads the list of most frequent lemmas with 91,522,113 uses (https://www.clarin.si/noske/run.cgi/view?corpname=gfida20_dedu_p:usesubcorp=:q=q%5Blemma%3D%3D%22biti%22%5D), whereby 66,247,726 of *biti* sentences (72.38%) have present tense BE-forms, incl. perfect and plusperfect auxiliaries in Type III contexts. Slovenian lacks simple past forms: therefore, the lexical form *bil, bila, bilo, bili* total only 7,750,160 forms (8.46%).

2. The present tense BE-paradigm in Modern Russian: grammar and corpus tags

2.1. Parametric grammar and present tense agreement

The definitional feature of the Russian present tense BE-paradigm is the absence of person agreement⁵. The key for what is recognized as the Modern Russian period is furnished by the extinct Old Russian construction of the *л*-perfect, which required person agreement and overt BE-auxiliaries in the 1–2 p.: *пришел*_{PART.SG.M} *есмь*_{1SG} ‘I (male) came’, *пришла*_{PART.SG.F} *есмь*_{1SG} ‘I (female) came’, *пришел*_{PART.SG.M} *еси*_{2SG} ‘you (sg, male) came’, *пришла*_{PART.SG.F} *еси* ‘you (sg, female) came’, *пришли*_{PART.PL} *есмѣ*_{1PL} ‘we came’, *пришли*_{PART.PL} *есте*_{1PL} ‘you (pl) came’, *пришла*_{PART.DU} *есвѣ*_{1PL} ‘we two came’, *пришла*_{PART.DU} *еста*_{2DU} ‘you two came’. This construction is incompatible with Russian grammar, since *л*-participles changed their morphological status from nominal to purely verbal forms which do not combine with BE-auxiliaries. Consequently, phrases like *пришел есмь* diagnose borrowed grammar in a language, where *л*-forms are verbal. The *л*-perfect is a Type III structure i.e. an analytical verb form with an auxiliary. For Type II structures with nominal predicates and copular BE the diagnostics is less clear, since the corresponding contexts survive in contemporary Russian.

2.2. Borrowed agreement in the Russian National Corpus: the 1–2 p. of *быть*

The occurrences of historical present 1–2 p. BE-forms in the main corpus of RNC must be explained as borrowings either from Old Russian or Old Church Slavonic. The form *суть* despite the tag ‘archaic’ assigned by Russian lexicographers is an inherited part of the paradigm. The main corpus of RNC (ca. 1700–) includes some historical present forms of *быти*, which come from dated texts, quotations, parodies or philological commentary. This is confirmed by the limited number of texts where these forms occur: the search for 1Sg *есмь* returns 442 documents and 951 hits, for 2Sg *еси*—538 documents and 1645 hits, for 1Pl *есмы*—85 documents and 129 hits, for 2Pl *есте*—92 documents and 180 hits, for 1Du *есвѣ* and *есва*—just 2 hits in 1 document, for 2Du *еста*—6 hits in 6 documents. These figures are low compared to 3rd p. forms: 3Pl *суть* occurs in 6,329 documents and 3Sg *есть*—in 41,160 documents. In this period, the *л*-perfect is extinct. Russian authors which tried to emulate the Church Slavonic usage occasionally attached agreement markers not to the *л*-participle, which is not specified for person but to the verbal forms that already had inflexional person markers, e.g. to present tense as in (9) or the aorist⁶ in (10). Such fail-

⁵ The identification of Russian and Hungarian as ‘languages with a zero copula’ in [Benvenist 1960] does not capture an essential difference between their BE-paradigms. Hungarian just as Old Russian has zero copula in the 3rd p., but overt copulas in 1–2 p. [Bánhidi, Jókay, Szabó 1965: 67–69], while Modern Russian has a 1–3 p. zero present BE-form both in the contexts for a copula and for a content verb [Apresjan 1996: 528; Testeleets 2008: 784].

⁶ The traditional estimate for the elimination of the aorist in Russian is late XV century [Borkovskij, Kuznetsov 1963: 279].

ures prove that the *л*-перфект did not correspond to the speakers' own idiom. They treat *л*-forms as verbs and combine the dated agreement forms of the auxiliary with verbs on the basis of a wrong analogy: *наказал*_{PART.SG.M} *еси*_{2SG} *мя*_{ACC.SG} 'you punished me' → **наказуешь*_{PRES.2SG} *еси* *мя* 'you punish me'.

- (9) Сосет под ложечкой неимоверно. Господи, за что *наказуешь*_{PRES.2SG} *еси*_{2SG} *мя*? [Влад. Азов. Маленький фельетон. Из дневника дипломата Уступчивого (1908.10.17) // «Русское слово», 1908].
- (10) да будут Очи Твои отверсты на Дом сей день и ночь, на Место сие, о нем же *глаголах*_{AOR.1.SG} *еси*_{2SG}, будет Имя твое тамо, еже услышати молитву [А. И. Богданов. Описание Санкт-Петербурга (1751)].

Unequal distribution of the *л*-perfect confirms that this construction is a borrowing. I checked all forms of 1–2 p. including the dual, which died out in Old Russian ca. 1600. Sequences like *был*_{PART.SG.M} *еси*_{2SG} i.e. combinations of a present tense auxiliary with a lexical form of *быть* were excluded.

Tab. 2. The *л*-перфект with 1–2 p. BE-forms in the main corpus of RNC from 1700 A.D.

	1700–1799		1800–1899		1900–1999		2000–...	
	All	<i>л</i> -perfect	All	<i>л</i> -perfect	All	<i>л</i> -perfect	All	<i>л</i> -perfect
1Sg: <i>есмь</i>	69	0	202	5 (2.5%)	336	102 (97%)	56	3 (5.35%)
2Sg: <i>еси</i>	291	169 (58%)	252	103 (40.9%)	487	201 (41.3%)	98	44 (44.9%)
1Pl: <i>есмы</i>	57	2 (0.35%)	20	2 (10%)	37	2 (0.54%)	3	0
2Pl: <i>есте</i>	57	1 (0.18%)	30	7 (23.3%)	20	1 (5%)	1	0
1Du: <i>есвь, -а</i>	2	0	0	0	0	0	0	0
2Du: <i>еста</i>	1	0	1	0	1	0	1	1

Tab. 2 shows that the *л*-perfect is more or less regularly reproduced in 2Sg, where it makes up 47.3% of the sample. Other combinations are sporadic: 32 hits from total 995 uses (3.21%). There is no substantial increase or decline of frequency in the use of 1–2 p. forms. I conclude that they are lexical borrowings that do not revive the lost mechanism of the person-and-number agreement. The variety of vernacular Old Russian described in [Zaliznjak 2008: 236] lacked overt 3rd p. auxiliaries in the *л*-перфект. Phrases like *пришел*_{PART.SG.M} *есть*_{3SG}, *пришли*_{PART.PL} *суть*_{3PL} must be extremely rare in Modern Russian, since the speakers lack inherited grammar for such combinations. This prediction is born out: we found just 4 examples with *л*-perfect in the sample of total 5,040 uses from 1700 A.D. on.

Tab. 3. The *л*-перфект with 3PI *суть* in the main corpus of RNC from 1700 A.D.

	1700–1799		1800–1899		1900–1999		2000–...	
	All	<i>л</i> -perfect	All	<i>л</i> -perfect	All	<i>л</i> -perfect	All	<i>л</i> -perfect
3PI: <i>суть</i>	1.433	2	1.889	0	1.519	2	199	0

The negligible percentage of the *л*-перфект (0.08%) is expected if *суть* and *есть* are part of the BE-paradigm both in the source language(s) and in the target language, but the *л*-перфект is lacking from the target language. The survived uses of *есть* and *суть* correspond not to the *л*-perfect but to Type II contexts (copular BE) and Type I contexts (BE as a content verb).

Tab. 4. The *л*-perfect in Old Russian vs past tense in Modern Russian

	Source languages		Target language
	Old Russian	Old Church Slavonic	Modern Russian
<i>л</i> -forms as past tense markers	part of the analytical construction		single word form
BE-auxiliary in the past tense construction with <i>л</i> -forms	agreement marker		absent
Combination of an <i>л</i> -form with a 1–2 p. BE-auxiliary, type <i>пришел есмь</i>	required		ungrammatical
Combination of an <i>л</i> -form with a 3 rd p. BE-auxiliary, type <i>пришел есть, пришли суть</i>	optional	required	ungrammatical

2.3. Modern Russian *суть*: residual agreement or a redundant present marker?

The form *суть* is more than 20 times less frequent (16,088 raw hits in the main corpus of RNC) than *есть* (393,200 raw hits⁷). The verb form *суть*₂ must be distinguished from the homonymic noun *суть*₂ ‘essence’ and from the collocation *не суть важн-о, -ое, -а, -ы, -ые* ‘does not matter’. Preliminary observations show that *суть*₁ and *суть*₂ have comparable frequency, but the frequency of *суть*₁ increases towards the end of the period, while *суть*₂ displays the opposite tendency. As stated above,

⁷ The vast majority of the occurrences feature the presence BE-form *есть*₂ and not the imperfective infinitive *есть*₁ ‘to eat’. The search for the parallel perfective infinitive *съесть* ‘to eat up’ returns only 2,268 hits. There is also a third candidate for the disambiguation—the military response *есть*₃! ‘I obey’, which is an infrequent word.

Russian lexicographers link *суть*₂ to scientific or archaizing discourse. This practice is confirmed by the stats: *суть* occurs in the main corpus of RNC only in 6,329 documents vs 41,160 documents for *есть*. The search gives back both *суть*₁ ‘essence’ and *суть*₂: texts containing *суть*₁ can lack *суть*₂ and vice versa. The majority of hits for the search < *суть*₁ ∨ *суть*₂ > come from non-fiction texts (5,024 documents, 12,703 hits), most of them are from the groups ‘journalism’ (3,346 documents, 6,454 hits) and ‘academic/pedagogical texts’ (718 documents, 4,522 hits). Meanwhile, the group ‘church and theology’ adds only 217 documents and 709 hits.

[Jevgenjeva 1999: IV: 305] treats *суть*₂ as an optional form of the 3rd p. primarily used in 3Pl, while [Apresjan 1996] disapproves *суть*₂ in 3Sg and treats it as an optional variant of the copular BE in 3Pl [Apresjan 1996: 518, 528]. This model is rendered in **Tab. 5**:

Tab. 5. The present tense paradigm of *быть* in Russian, after [Apresjan 1996]

	∅ _{BE.PRES}		ЕСТЬ		СУТЬ	
	Content verb	Copula	Content verb	Copula	Content verb	Copula
1Sg.	+	+	+	+	*	*
2Sg.	+	+	+	+	*	*
3Sg.	+	+	+	+	*	*
1Pl.	+	+	+	+	*	*
2Pl.	+	+	+	+	*	*
3Pl.	+	+	+	+	*	+

I checked the main corpus of RNC for contact sequences of the type subject pronoun + суть in the window <-1; 1>. The search was limited by the period 1800–2015 in order to exclude doubts about the grammar of the XVIII century texts. The sample for *суть*₂ totals 239 sentences. 3Pl prevail (89.1%), but all other combinations are attested. In the second group, the most frequent combination is 1Pl мы ‘we’ + суть (12 examples). In the first group, 26 sentences (12.2% from all 3Pl uses) show *суть* as a content verb, therefore, Apresjan’s statement that this option is out must be softened.

Tab. 6. The distribution of *суть*₂. The figures show the number of contact sequences with subject pronouns in the main corpus of RNC from 1800 A.D.

	SG		PL	
	Content verb	Copula	Content verb	Copula
1p.	0	3 (1.25%)	0	12 (5%)
2p.	1 (0.42%)	2 (0.83%)	0	4 (1.67%)
3p.	0	4 (1.67%)	26 (10.87%)	187 (78.24%)

The examination of the RNC examples with *суть*₂ in 1–3Sg and 1–2Pl shows that such uses are rooted in the Russian language of 1800–1950. The list of authors includes Ivan Turgenev, Maxim Gorki, Sergei Bulgakov, Ivan Šmelev, Alexander Kuprin, Vyačeslav Šiškov, Konstantin Fedin. In 1950–2000, the list of authors who license *суть*₂ in 1–2 p. and in 3Sg includes Nina Berberova, Vladimir Makanin, cf. (11), Strugacki brothers and Iosif Brodskij, cf. (12). This prompts a hypothesis that for a group of speakers *суть*₂ survived as part of oral discourse, where it loses the person-and-number specification and assumes the status of an emphatic copula in the meaning ‘X is in essence Y’.

- (11) Вроде как **все**_{PL} **мы**_{1PL} **суть**^{BE.PRES} брежневские инвалиды [Владимир Маканин. Андеграунд, или герой нашего времени (1996–1997)]
 ‘It looks like **all of us** are in essence invalids from Brežnev’s time.’
- (12) Ибо **война**_{SG} **суть**^{BE.PRES} **эхо**_{SG} кочевого инстинкта.
 [И. А. Бродский. Путешествие в Стамбул // «Континент», 1985].
 ‘Since **war** is in essence **an echo** of the nomadic instinct.’

The spreading of the more frequent form *есть*₂ over the plural makes a special form of the 3Pl redundant. That the latter survived is due to the tendency towards using *есть*₂ and *суть*₂ in different contexts. This tendency is captured by Апре-*jan’s* model, but the distribution in Tab. 6 has never been achieved because of the opposite tendency towards expanding the coverage of *суть*. This begs an alternative model outlined in Tab. 7 below.

Tab. 7. The present tense paradigm of *быть* in Russian: a corpus alternative

	∅ ^{BE.PRES}		ЕСТЬ		СУТЬ	
	Content verb	Copula	Content verb	Copula	Content verb	Copula
1Sg.						
2Sg.						
3Sg.					(*)	
1Pl.	+	+	+	+		
2Pl.						
3Pl.					(+)	↑ +

2.4. ∅^{BE.PRES} vs *есть*: syntax and semantics

There is a consensus that ∅^{BE.PRES} is a separate element in syntax but not an elided form of *есть* [Peškovskij 1928: 303]; [Testelets 2008]; [Letučij 2018]. The correlations between the distribution of ∅^{BE.PRES} vs *есть* and the taxonomic semantic type (existence, possession, location, characterization, identification etc.) are shown in [Arutyunova, Širyaev 1983]. I adopt this analysis with the additions proposed in [Yanko 2000]; [Dymarskij 2018]. [Letučij 2018] argues that ∅^{BE.PRES} and *есть* are always non-synonymic, so that (13a) presumably means ‘John’s flat is big’, while (13b) means ‘John has a big flat’. However, the shift from alienable possession

to characterization is not induced by $\emptyset^{\text{BE.PRES}}$, cf. the conjoined structure ‘X has Y and Z’ in (14), where the possessive reading is required.

- (13) a. У Ивана $\emptyset^{\text{BE.PRES}}$ большая квартира.
 ‘John has a big flat.’ [alienable possession].
 Or: ‘John’s flat is big’ [characterization]
 b. У Ивана **есть** большая квартира.
 ‘John **has** big flat.’ [alienable possession], # ‘John’s flat is big’
- (14) У Ивана $\emptyset^{\text{BE.PRES}}$ большая квартира в городе и уютный дом в деревне.
 ‘John **has** a big flat in the downtown and a nice house in the village.’
 #‘John’s flat in the downtown is big and his house in the village is nice.’

3. The silent head: measuring the impact of the zero present form

The distribution of the *есть*₂ and *суть*₂ is constrained by the expansion of $\emptyset^{\text{BE.PRES}}$. It ousted the overt forms from a number of contexts and made them optional elsewhere. The ratio of the $\emptyset^{\text{BE.PRES}}$ vs *есть*₂ uses cannot be measured directly in corpora without syntactic annotation, but there are indirect estimates. I measure the distribution of *есть*₂ for different persons and numbers in the same context as with *суть*₂. The search was reduced to the sequences of the type subject pronoun + *есть*₂ in the window <-1; 1>. The uses of the content verb *есть*₂ vs copula *есть*₂ were measured on a separate basis. The default hypothesis is that unequal distribution of *есть*₂ reflects the impact of $\emptyset^{\text{BE.PRES}}$ which fills in the gap in certain persons and numbers.

3.1. The proportion of *есть* & *суть* vs $\emptyset^{\text{BE.PRES}}$

The sample for *есть*₂ with a contact subject pronoun totals 7,458 sentences. This is ca. 30 times larger than the sample for *суть*₂ in the same context (239). Tab. 8 shows the ratio of copular uses in each combination subject pronoun + *есть*₂. A separate line shows how this ratio changes if measured for the pair *есть*₂ & *суть*₂.

Tab. 8 shows a big increase (>1%) with the adding of *суть*₂ only in 3Pl and in 1Pl. The percentage of copular *есть*₂ is abnormally low in 3Pl (4.24%), therefore adding 187 sentences with *суть*₂ is relevant. The combined ratio for 3Pl (16.01%) is nevertheless low compared to other persons and numbers. This confirms that *суть*₂ retains a systemically important status mainly as a 3Pl copula, where *есть*₂ is underrepresented. Since *суть*₂ is a low frequent word, it does not fully compensate this gap which must be filled by $\emptyset^{\text{BE.PRES}}$. The expectancy of an overt copula is higher in 1–3 Sg. (combined ratio 32.83%) than in the 1–3 Pl. (combined ratio 20.68%). This indicates that copular $\emptyset^{\text{BE.PRES}}$ is especially salient in the plural. The positions of *есть*₂ as a content verb are stable both in Sg and in Pl. The share of all uses in the 1–2 p. (both content verb and copula) is ca. 5 times less compared to the 3rd p.⁸: the figures are almost identical for Sg (21.03%) and Pl (20.64%). The ratio of the copular uses in 1Sg and

⁸ In a sample including non-pronominal subjects, the contrast is even sharper.

2Sg is nevertheless high. I interpret this as a proof that *есть* is stable in these person-and-number forms.

Tab. 8. The distribution of *есть* and *суть*. The figures show the number of contact sequences with subject pronouns in the main corpus of RNC from 1800 A.D. The percentage shows the ratio of content verb vs copular uses.

	SG		PL	
	Content verb	Copula	Content verb	Copula
1p.	440 (56.85%)	334 (43.15%)	191 (84.1%)	67 (25.9%)
w. <i>суть</i>	440 (56.13%)	337 (43.77%)	191 (70.75%)	79 (29.25%)
2p.	345 (48%)	375 (52%)	147 (61%)	94 (39%)
w. <i>суть</i>	346 (47.76%)	377 (52.14%)	147 (60%)	98 (40%)
3p.	All gender forms: 3,023 (72.68%) 3Sg.M 1194 (79.51%) 3Sg.F 1111 (75.48%) 3Sg.N 718 (61.27%)	All gender forms: 1,144 (27.32%) 3Sg.M 327 (21.49%) 3Sg.F 363 (24.62%) 3Sg.N 454 (38.73%)	1,243 (95.76%)	55 (4.24%)
w. <i>суть</i>	3,023 (72.48%)	1,148 (27.52%)	1,269 (83.99%)	242 (16.01%)
Total: 7,458	3,808 (67.27%)	1,853 (32.73%)	1,581 (77.98%)	216 (12.02%)
w. <i>суть</i> : 7,697	+ 1 3,809 (67.17%)	+ 9 1,862 (32.83%)	+ 26 1,607 (79.32%)	+ 203 419 (20.68%)

The approximation does not provide absolute figures for $\emptyset^{BE.PRES}$, but heuristic estimates can be given. One of them is based on the next-neighbor method, which requires a lookup of the left and right context for the pivotal subject element ...X... in order to check whether the right or left neighbor of X is its complement in the verb phrase $[\emptyset^{BE.PRES}-Y]_i$ linked with X_i . If the search is oriented to identifying the right neighbor as complement of the silent head $\emptyset^{BE.PRES}$ and X is the 1Sg subject pronoun я_{1SG} 'I', sentences like *Это $\emptyset^{BE.PRES}$ я $_{1SG}$, Иван $_{NOM}$* 'That is me, **John**' will return 'false', while sentences like *Все-таки я $_{1SG}$ — $\emptyset^{BE.PRES}$ дурак* 'Still, I am a fool' return 'true'. If one takes the context subject pronoun + noun in the nominative case in the window $\langle 1; 1 \rangle$, the expectancy that these elements are part of the structure $S_{pron}-\emptyset^{BE.PRES}-S_{NOM}$, where the pronoun is the subject and its right neighbour it is part of its nominal complement can be measured. I checked sequences of the type 1Sg subject personal pronoun я $_{1SG}$ + noun in the nominative case: the RNC search returns 78,676 raw hits. A test sample of 2,000 sentences dated with 1987–2015 was processed. The input had

wrong morphological tags fixed by the annotator manually. The lexical-grammatical search in RNC returns all elements which can be analyzed as nouns in the nominative case, cf. the adjective *рада* ‘glad’ (cf. the noun *рада* ‘Ukrainian parliament’), adverb *дома* ‘at home’, cf. the noun *дома* ‘houses’, preposition *перед* ‘in front of’, cf. the noun *перед* ‘the front end’ as well all syncretic forms that can either stand for nominative or some other case. The sample also included sentences where *я* and its right neighbor belong to different clauses and other structures that do not match the pattern $S_{\text{pron}} - \emptyset^{\text{BE.PRES}} - S_{\text{NOM}}$. Sentences where the entire structure $[\emptyset^{\text{BE.PRES}} - S_{\text{NOM}}]$ was located to the left from *я* were filtered out, since the right neighbor of *я* is not a complement of $\emptyset^{\text{BE.PRES}}$. At the same time, blind hits with expressions wrongly tagged as S_{NOM} cf. *Я рада* _{ADJ.SG.F} ‘I am glad’ or *Я дома* _{ADV} ‘I am at home’ were rendered positive, if they matched the pattern with the proviso that the predicate complement is not a noun but an non-verbal element of different morphology. The trimmed sample returns 49.25% positive examples (985 from 2,000). If this ratio holds for the whole RNC collection in the searched context, it should include 38,748 sentences with the subject in 1Sg, zero copula $\emptyset^{\text{BE.PRES}}$ and the word order $Я_{\text{1SG}} - \emptyset^{\text{BE.PRES}} \dots S_{\text{NOM}} / \text{PRED}$.

One more estimate is based on the proportional usage of past and present tense forms of *быть*. As stated in 1.2., overt past tense forms *был*, *была*, *было*, *были* partly correspond to overt present forms *есть*₂ and *суть*₂ (Type I contexts) partly—to $\emptyset^{\text{BE.PRES}}$ (Type II contexts). Let us assume that RNC has at least as much Type I sentences in the past tense as in the present tense. Let us also assume that all uses of *есть*₂ and *суть*₂ pattern with Type I structures and all uses of $\emptyset^{\text{BE.PRES}}$ pattern with Type II structures: this simplification maximizes the number of sentences with *есть*₂ and *суть*₂. If there are *m* sentences with *есть*₂ and *суть*₂ and *n* sentences with *был*, *была*, *было*, *были*, the number of sentences with $\emptyset^{\text{BE.PRES}}$ is $n - m = k$. The 4 forms *был*, *была*, *было*, *были* return 2,267,476 raw hits. These verb forms have two homonyms—the particle *было*₂ ‘marker for a canceled event’ and the nominal form *были* from the lemma *быль* ‘legend’. Both are low frequent words: let us assume that they take maximum 7,476 hits, which is actually above than their frequency. Then we get 2,600,000 uses of the past tense forms of *быть* after the disambiguation. The present form *есть*₂ (393,200 raw hits) has homonyms *есть*₁ ‘to eat’ and *есть*₃ ‘I obey’ [military command]: the exact figures are not available, since the search returns the homonyms, but one can assume that the frequency of *есть*₁ corresponds to the frequency of its perfective correlate *съест* ‘to eat up’ (2,268 hits) and *есть*₁ and *есть*₃ total maximum 3,200 hits. Then *есть*₂ gives 390,000 hits after the disambiguation. The present form *суть*₂ (16,088 raw hits) has a homonymic noun *суть*₁ ‘essence’, they have a comparable frequency. Let us assume that *суть*₁ takes maximum 8,088 hits. Then we are left with 8,000 hits of *суть*₂. With these stipulations, RNC should feature at least $k = 1,862,000$ sentences with $\emptyset^{\text{BE.PRES}}$, since $n = 2,260,000$ and $m = 398,000$. With the stipulations made, all these RNC sentences with $\emptyset^{\text{BE.PRES}}$ will be interpreted as copular, though in reality a minor part from 1,862,000 sentences are structures with a zero content verb in contexts like *У него* $\emptyset^{\text{BE.PRES}}$ *много книг* ‘He has many books’.

Finally, the estimates for sentences with $\emptyset^{\text{BE.PRES}}$ in large corpora can be derived on the basis of tree banks with syntactic annotation. Such estimates however reflect the architecture of the parser. Apresjan’s model of *быть* adopted in this paper

is implemented in the ETAP-3 parser [Apresjan et alii 2003]. The present forms \emptyset and *есть* are recognized here as separate elements, but both of them belong to the lemma *ЕСТЬ*, while all non-present forms of BE are linked to a different lemma—*БЫТЬ*⁹. In a parser based on Ščerba's model, $\emptyset^{\text{BE.PRES}}$ and *есть* will be linked to different lemmas. Since the notion of the zero element is non-neutral, any technical decision has impact on processing the coverage of $\emptyset^{\text{BE.PRES}}$.

3.2. Morphological paradigms and historical corpora

The present tense paradigm of *быть* 'be'—{ $\emptyset^{\text{BE.PRES}}$, *есть*, *суть*} is historically a transition from an agreement system characteristic of Old Russian to a system without overt present BE-forms. It is surprisingly stable: the overt forms *суть* and *есть* did not disappear during the last 300 years. The historical 1–2 p. forms of *быть* behave as borrowed elements already in the XVIII century. The loss of number agreement in the 3rd p. is not a new phenomenon either. The XVII century traveler Pjotr Tolstoj (b. 1645) in his diary included in the Historical corpus of RNC uses *суть*₂ 4 times with plural nouns and 6 times with singular nouns, cf. (15).

- (15) **Варшава**_{SG} **суть**^{BE.PRES} **место**_{SG} **великое**_{SG},
на левом берегу реки Вислы положенное.
[Путешествие стольника П. А. Толстого по Европе. 1697–1699 (1699)]
'**Warsaw** is [lit: are] a big city founded on the left bank of the Wisla-river.'

P. Tolstoj's treatment of *суть*₂ as an emphatic copula does not differ from the XIX–XX century examples (11) and (12). It is plausible that an idiom of Russian with such settings for *суть*₂ existed during a long time but was suppressed by Church Slavonic which only approved *суть*₂ in ЗПИ.

The history of Russian 'BE' can be modeled on the basis of its usage in the Modern Russian period, if one takes into account three blocks of input data for each present BE-form: 1) frequency and number of texts, where this BE-form is attested; 2) even vs uneven distribution of BE-forms for different persons and numbers; 3) even vs uneven distribution of the copular vs content verb uses for each BE-form. If one adopts the hypothesis that the Russian present tense BE-paradigm { $\emptyset^{\text{BE.PRES}}$, *есть*, *суть*} originates from a paradigm where all elements were genuine agreement markers, its restructuring follows three steps:

- I. The 1–2 p. forms disappear first;
- II. The uses of *суть* get restricted by the pair { $\emptyset^{\text{BE.PRES}}$, *есть*};
- III. The uses of *есть* get restricted by $\emptyset^{\text{BE.PRES}}$.

This model allows making two predictions concerning the past and the future of the BE-paradigm:

- (i) $\emptyset^{\text{BE.PRES}}$ is historically a 3rd person form and an inherited part of the BE-paradigm.
- (ii) The form *есть*₂ will disappear from the paradigm of BE in the future.

⁹ This decision is commented in [Apresjan 1996: 528].

The hypothesis (i) is in line with historical linguistics: the latter confirms that Old Russian, unlike Old Church Slavonic lacked overt copulas in the 3rd p. in Type III contexts (л-перфект) [Borkovskij, Kusnetsov 1963: 203] and partly also in the Type II contexts (copular structures with nominal predicates) [Zaliznjak 2008: 259–261]. The prediction (ii) is in line with Russian dictionaries, which claim that *быть* has no present forms except for the *есть*₂ which probably is a separate verb. This description does not hold for the present-day Russian BE-paradigm but anticipates its future.

4. Conclusions

The undertaken study has shown that in a language where zero syntactic forms gradually replace overt forms the status of endangered forms is revealed by two measures: 1) low frequency and uneven distribution in the texts; 2) uneven distribution in different persons and numbers. The history of the Russian BE-paradigm requires a third one and a more specific measure—3) uneven distribution of copular vs non-copular uses in each person and number form. The coverage of the zero copula $\emptyset^{\text{BE.PRES}}$ in Modern Russian can be processed in corpora without syntactic annotation on the basis of systemic proportions between different types of syntactic contexts.

References

1. *Apresjan Ju. D.* (1996), Lexicographical portraits. On the basis of the verb *быть* [Lexikografičeskie portrety (na materiale glagola *быть*)], Integral description of language and systemic lexicography [Integral'noe opisanie jazyka i sistemnaja leksikografija]. LRC publishing house, Moscow, 1996, pp. 518–537.
2. *Apresjan Ju. D., Boguslavsky I., Iomdin L., Lazurskij A. V., Sannikov V. Z., Sizov V. G., Tsinman L. L.* (2003), ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT, First International Conference on Meaning-Text Theory (MTT'2003). June 16–18, 2003. Paris: Ecole Normale Supérieure, 2003, pp. 279–288.
3. *Arutyunova N. D., Širyaev E. N.* (1983) Russian sentence. The existential type [Russkoe predloženie. Bytijnyj tip], Nauka, Moscow.
4. *Bánhidí Z., Jókay Z., Szabó D.* (1965). Learn Hungarian. 3rd ed. Budapest.
5. *Benvenist E.* (1960). 'Etre' et 'avoir' dans leur fonctions linguistiques, Bulletin de la Société de linguistique, 1960, tome LV.
6. *Borkovskij V. I., Kusnetsov P. S.* (1963), Historical Russian grammar [Istoričeskaja grammatika russkogo jazyka], Nauka, Moscow.
7. *Dal V. I.* (1880), Explanatory dictionary of the living Great Russian language [Tolkovyj slovar' živogo velikoruskogo jazyka]. Vols. I–IV. 2nd ed. Moscow: M. O. Wolf publishing house, Moscow.
8. *Dymarskij M. Ja.* (2018), A u menja v karmane gvozd'. Zero copula or predicate ellipsis? [A u menja v karmane gvozd'. Nulevaja svjazka ili ellipsis skazuemogo?], World of the Russian word [Mir russkogo slova], 2018, № 3, pp. 5–12.
9. *Jevgenjeva A. P.* (1999), A dictionary of Russian [Slovar' russkogo jazyka], in 4 vols. 4th ed. Russkij jazyk, Moscow.

10. *Letučij A. B.* (2018), Zero copula [Nulevaja svjazka], Russian corpus grammar [Russkaja korpusnaja grammatika] <http://rusgram.ru/>.
11. *Lyščevskaja O. N., Šarov S. A.* (2009). Frequency dictionary of Modern Russian (on the basis of Russian national corpus) [Častotnyj slovar' sovremennogo russkogo jazyka (na material National'nogo korpusa russkogo jazyka)], Azbukovnik press, Moscow.
12. *Ožegov S. I., Švedova N. Ju.* (1992). Explanatory dictionary of Russian [Tolkovyj slovar' russkogo jazyka], Moscow.
13. *Peškovskij A. M.* (1928), Russian syntax in a scientific perspective [Russkij sintaksis v naučnom osveščanii]. 3rd ed., Moscow-Leningrad.
14. *Pustet R.* (2003), Copulas. Universals in the Categorization of the Lexicon. OUP, Oxford.
15. *Ščerba L. V.* (1928), On parts of speech in Russian [O častjah reči v russkom jazyke], Russian speech, [Russkaja reč], New series, II. Academia publishing house, Leningrad, pp. 5–27.
16. *Stassen L.* (1994), Typology versus mythology: the case of the zero copula, *Nordic Journal of Linguistics*, 17, pp. 105–126.
17. *Testelets Ja. G.* (2008), Syntactic structures with a covert copula in Russian [Struktura predloženíj s nevyraženoj svjazkoj v russkom jazyke], *Dynamic models: Sentence and text. In honour of E. V. Padučeva [Dinamičeskie modeli: Predloženie. Tekst. Sbornik statej v čest' E. V. Padučevoj]*, LRC publishing house, Moscow, pp. 773–789.
18. *Ušakov D. N.* (1935–1940), Explanatory dictionary of Russian [Tolkovyj slovar' russkogo jazyka], Vols. I–IV. Moscow.
19. *Yanko T. J.* (2000), Existence and possession: constructions with the verb *byt'* [Bytovanie i obladanie: konstrukcii s glagolom *byt'*], N. D. Arutynova, I. B. Levontina (eds.). *Logical analysis of language. Languages of space. [Logičeskij analiz jazyka. Jazyki prostranstv]*, Indrik publishing house, Moscow, pp. 198–211.
20. *Zaliznjak A. A.* (1977), Grammatical dictionary of Russian [Grammatičeskij slovarj russkogo jazyka]. Nauka publishing house, Moscow.
21. *Zaliznjak A. A.* (2004), Old Novgorod dialect [Drevnenovgorodskij dialect], 2nd ed. MLRC publishing house, Moscow.
22. *Zaliznjak A. A.* (2008), Old Russian enclitics [Drevnerusskie enklitiki], LRC publishing house, Moscow.
23. *BNC*—British National Corpus: <https://www.english-corpora.org/bnc/>. Accessed at 31.03.2020.
24. *COCA*—Corpus of Contemporary American English: <https://www.english-corpora.org/coca/>.
25. *Gigafida*—<http://www.gigafida.net/>. Accessed at 31.03.2020.
26. *ETAP-3*—Linguistic processor ETAP-3: <http://proling.iitp.ru/ru/etap3>.
27. *RNC*—Russian National Corpus: <http://www.ruscorpora.ru/>. Accessed at 31.03.2020.

THE PERCEPTION OF ROBOT'S EMOTIONAL GESTURES AND SPEECH BY CHILDREN SOLVING A SPATIAL PUZZLE¹

Zinina A. A. (zinina_aa@nrcki.ru),

Zaidelman L. Y. (zaydelman_ly@nrcki.ru),

Kotov A. A. (kotov_aa@nrcki.ru),

Arinkin N. A. (arinkin_na@nrcki.ru)

Kurhcato Institute; Russian State University for the Humanities, Moscow, Russia

The emotional behavior of a companion robot is important for human-robot interaction in the situation of training tasks. We examined the influence of emotional gestures and emotional speech of the robot on its perception by primary school students (N=52, male, female, mean age 9.8) in the situation of joint solution of the spatial Tangram puzzle. It was shown that emotional gestures make a significant contribution to the attractiveness of the robot for the child. It was also found that test subjects prefer the robot with emotional gestures and speech over the robot with neutral gesture and speech behavior. The study also analyzed the communicative behavior of children, identified typical communicative signs that are typical for interaction start with the robot, for monitoring the game and for difficult situations. We described typical mistakes that children make when assembling a puzzle together with the robot.

Keywords: multimodal communication, robot-to-human interaction, affective robot tutors

DOI: 10.28995/2075-7182-2020-19-811-826

ВОСПРИЯТИЕ ЭМОЦИОНАЛЬНЫХ ЖЕСТОВ И РЕЧИ РОБОТА ДЕТЬМИ, РЕШАЮЩИМИ ПРОСТРАНСТВЕННУЮ ГОЛОВОЛОМКУ

Зинина А. А. (zinina_aa@nrcki.ru),

Зайдельман Л. Ю. (zaydelman_ly@nrcki.ru),

Котов А. А. (kotov_aa@nrcki.ru),

Аринкин Н. А. (arinkin_na@nrcki.ru)

Курчатовский институт; Российский государственный гуманитарный университет, Москва, Россия

¹ The research is supported by the Russian Science Foundation (project No 19-78-00113).

Эмоциональное поведение робота-компаньона важно при взаимодействии человека и робота в ситуации решения обучающих задач. В рамках эксперимента (N=52, муж., жен., средний возраст 9,8) исследовалось влияние эмоциональных жестов и эмоциональной речи на восприятие робота младшими школьниками. Было показано, что именно эмоциональные жесты вносят существенный вклад в привлекательность робота для ребенка. Также обнаружено, что испытуемые скорее предпочитают робота с эмоциональными жестами и речью по сравнению с роботом с нейтральным жестовым и речевым поведением. В рамках исследования также было проанализировано коммуникативное поведение детей при решении пространственной головоломки Танграм, выделены типичные коммуникативные знаки у детей, характерные для начала взаимодействия с роботом, контроля хода игры, затруднительных ситуаций. Описаны типичные ошибки при совместной сборке головоломки с роботом.

Ключевые слова: мультимодальная коммуникация, взаимодействие между роботом и человеком, роботы-наставники

1. Introduction

Robots with social skills are increasingly used in various fields of education. Learning with robots is called “r-Learning” [Han et al., 2008]. In a situation of learning, robot can act as a mentor, a tool, or an equal partner [Mubin et al., 2013]; [Kanda et al., 2004]. Robots are used for teaching natural sciences [Shiomi et al., 2015], mathematics [Brown, Howard, 2014] and music [Han et al., 2009]. It is noted that robots are effective for language learning [van den Berghe et al., 2019]; [Belpaeme et al., 2018b]. In [Park et al., 2017] it is shown that interaction with a social robot acting as an equal partner (peer) can contribute to the development of cognitive skills among children. In many interaction situations, robots are more interactive and user-friendly than computers [Kanda et al., 2004]; [2014]; [Alemi et al., 2014]; [Belpaeme et al., 2018a]. As compared to software applications, social robots greater involve children in learning, increase their motivation and curiosity, as well as the number of emotional responses [Kanda et al., 2004]; [Gordon et al., 2015]; [Wainer et al., 2006]. Robots can demonstrate various means of communication: using gestures, movements, and facial expressions, which also helps to increase the interest and motivation of users to any training. In [Han et al., 2008], using the example of English lessons, it is shown that home robots are more effective than other types of learning tools (such as books or audio recordings): children show interest in learning, better concentration, and increased academic performance. In addition, children feel that a home robot is more friendly than other educational material. In [Han, Kim, 2009], it is shown that the robot’s cheering is especially important for children. In [Brown, Howard, 2014], the robot uses different verbal encouragement strategies when teaching maths to maintain the level of student engagement. The authors assume that educational robots should monitor student involvement and apply behavioral strategies (verbal or nonverbal cues) when involvement decreases. In our study, the robot acts as a child’s assistant when solving a puzzle. Thus, the robot controls the solution of the problem: it introduces puzzle to the child, gives instructions, and

monitors the progress. Here we follow [Leyzberg et al., 2012], where a robot also acted as an instructor in solving the puzzle. The authors showed that a physical robot is quite effective as a mentor. In addition, the advantages of a physical robot are highlighted in [Wainer et al., 2006] when performing in Hanoi puzzle.

Several researchers emphasize the importance of nonverbal behavior for educational robots. In [Kennedy et al., 2015b], the robot helped children determine whether numbers are prime. The authors showed that children who interact with a robot that exhibits more complex nonverbal behavior show great learning success. On the other hand, the same authors [Kennedy et al., 2015a] note that while the robot's learning strategy is effective, adding social and adaptive behaviors does not lead to a significant increase in efficiency. Therefore, although the presence of a physical robot leads to the improved learning progress, the performance of social behavior by a robot in the context of learning requires caution.

Even though affective robot mentors have great potential in the field of education, many researchers pay special attention to the development of an emotional robot model [Cuadrado et al., 2019]; [Schodde et al., 2017]. For example, in [Leite, Pereira, 2007] iCat robot played chess with a child. The robot's emotional state and expression is affected by every move of the child. The authors showed that in this situation "children may then interpret the characters affective behavior and by that acquire additional information to understand the game better and thus to improve their chess skills". In another paper [Pereira et al., 2010], the robot comments on the moves of two chess players, showing sympathy for one of them and a neutral attitude to the other. The results of an experimental study show that users who were empathized with the robot were more likely to perceive the robot as a friend. Thus, modeling emotional and empathic behavior is one of the key characteristics of an educational robot. Moreover, emotional behavior can be modeled not only by emotional statements, but also by emotional mimic and gesture reactions. In our work, we decided to test the effect of robot's emotional gestures and speech on participants in an experimental study. We assumed that participants would prefer the robot that helps them and uses emotional communicative actions. The study was conducted using F-2 robot, an experimental platform for studying the interaction between humans and robots [Zinina et al., 2019].

2. Research procedure

We conducted an experiment in which we evaluated the impact of emotional gestures and emotional speech on the attractiveness of the robot to a user. A separate goal of the experiment was to investigate the complex communicative behavior of test subjects when solving a spatial puzzle. In this experiment two robots helped participants to complete a Tangram puzzle. This puzzle is a well-known experimental media for studies of natural human communication [Clark, Wilkes-Gibbs, 1986], development of linguistic resources [Shore et al, 2018] as well as for the design of robot communicative strategies [Kirschner et al., 2016]. This puzzle is also familiar to primary school students, which helps to increase their interest and motivation to learn. The puzzle consists of 7 elements of different color, shape, and size (two big triangles: red and blue; two small triangles: yellow and dark blue; a middle green triangle, an orange parallelogram, and

a purple square). The task of a participant was to arrange the elements within a given contour on a white sheet. During the experiment, the participant was to complete: *Parallelogram*, *Fish*, *Triangle*, and *Ship* figures. The robot was located on the table in front of the participant and used speech instructions, gestures, and gaze, instructing the participant to put a certain element on a certain place. A human assistant was present in the experiment situation, but his communication with the subjects was minimal: the robot itself greeted a child, repeated the rules of the game, gave instructions, and asked to move to another robot after a figure is finished. At the beginning of the experiment, the test subject sat in front of one of the robots (Fig. 1), the robot was selected randomly. In the experiment the robot introduced the rules of Tangram puzzle, pointed to each element of the puzzle, and repeated its name, for example, “on the right there are two big triangles and a parallelogram”. Then the robot helped the child to solve two tasks. For example, the robot might say “take this big triangle and put it in the lower right corner”. Then the robot asked him to move to another robot. Another robot also introduced itself the child, helped him to solve the other two tasks and completed the experiment. The experiment was conducted in Russian. The behavioral scripts of the robot were organized as sequences of BML (Behavior Markup Language) packets: one sequence per task. The average time of the experiment was 15 minutes. After interacting with robots, the experimenter asked the children to choose the preferred condition (robot) and rate the robots in each of the two experimental conditions on five-point semantic differential scales, and then questioned about the difference in robot’s behavior in detail. In addition, the objective indicators were recorded— the speed of solving each task and the number of participants’ mistakes. The whole experiment was recorded from two viewpoints: front view of the experimental situation (Fig. 1a) and the side view of a player (Fig. 1b).

The experiment was organized in the paradigm of the Wizard of Oz [Kelley, 1984] in which the moves by the player were evaluated as successful or not by a remote human operator. After the operator had indicated a successful move, the robot chose a communicative pattern to praise the user and explained the next move. If a participant was mistaken, the operator indicated that to the robot, and the robot chose a communicative pattern to indicate an error and repeated the instruction. If the user solved the whole figure before the end of BML protocol, the operator gave a command to the robot, and the robot chose a communicative pattern to praise the participant for the successful task.



Fig. 1 (a). Experimental setup: front view



Fig. 1 (b). Experimental setup: side view

3. Experimental conditions

In the experiment two independent variables varied, each with two levels: robots demonstrated expressive (emotional) or neutral gestures, robots could also react with emotional or neutral statements (Fig. 2).

In one experimental series the child interacted with two robots whose behavior was different:

- (i) **emotionality of gestures**—we compared the conditions: (a) × (c) and (b) × (d). The robot with neutral gestures was opposed to the robot with emotional gestures, while the speech of two robots could be emotional or neutral, but the same for the two robots in each trial. Gestures varied only in the communicative reaction to the subject's actions, in the part where the robots gave instructions, the gestures did not differ.
- (ii) **emotionality of speech**—we compared the conditions: (a) × (b) and (c) × (d). The robot with neutral speech was opposed to the robot with emotional speech, while the gestures of two robots could be emotional or neutral, but the same for the two robots in each trial.
- (iii) **emotionality of gestures and speech**—we compared the conditions: (a) and (d). The robot that demonstrated emotions in gestures and speech was opposed to the robot with neutral gestures and speech.

Tasks with different experimental conditions were presented in random order.

Robot with **emotional speech**. This robot used emotional interjections, expressed emotional evaluation of user's actions and game situation (*Ah, what a beautiful figure we have got!*), or indirectly evaluated the interlocutor (*Wow, you are doing great!*). Negative politeness strategies were used in speech responses to erroneous user actions [Brown, Levinson, 1987]. For example, *We have the wrong element* ("rise common ground" strategy), *Oh, a bit wrong* ("reduce the imposition" strategy).

In its **neutral speech**, the robot only briefly evaluated the subject's actions using particles or predicates: for example, *Yes, Aha, Ok* and *Wrong, No, Not here*. The robot did not evaluate the child and did not describe his own emotions.

Emotional gestures for robots were developed based on the REC corpus. In this experimental condition, the robot accompanied speech with actions directed at the addressee and used gestures with such communicative functions as *joy, appeal* and *demonstration*. In a condition with neutral gestures, the robot accompanied speech with the imitated breathing and gestures typical to communicative function *compensation-closing*—e. g. joined hands in front of the body, imitated scratching of the body.

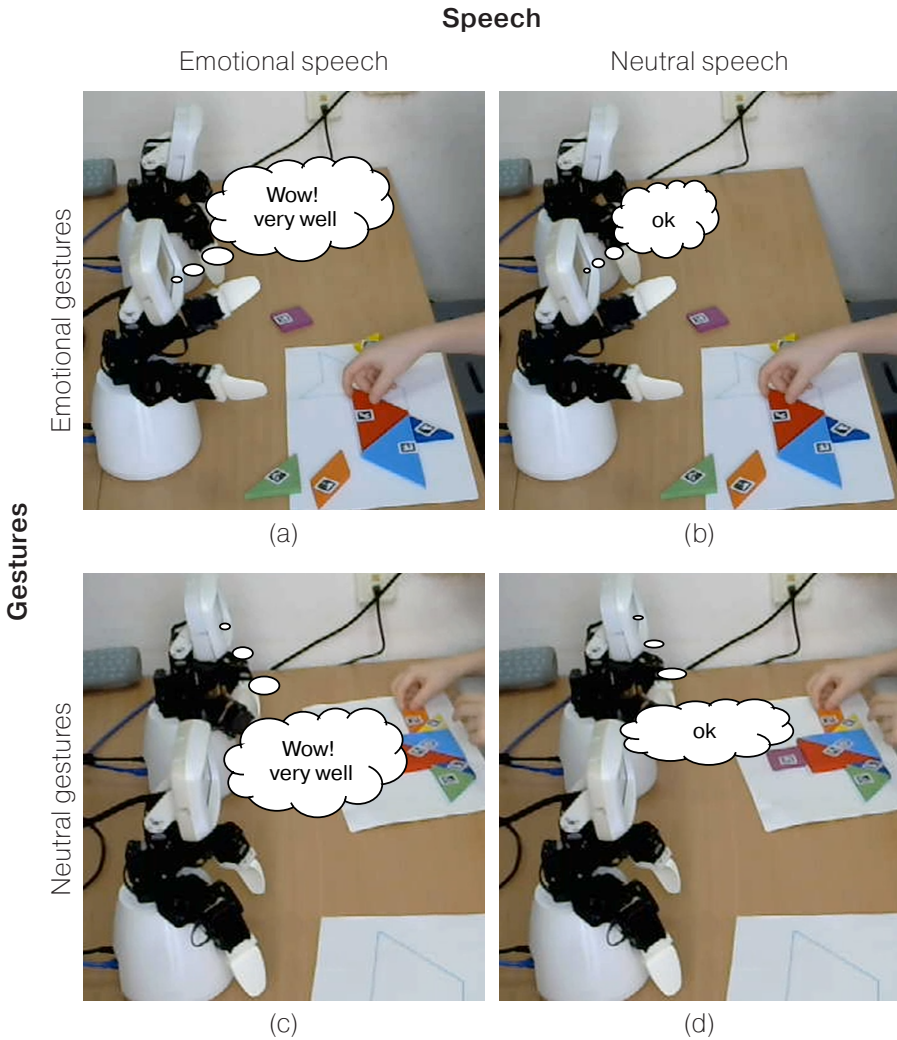


Fig. 2. Experimental conditions

4. Participants

Participants of the experiment were 4th degree pupils of an elementary school: 52 children, 18 girls and 34 boys, most of the participants were 10 years old, min. age 8, max. age 11. All subjects received informed consent from their parents. After the experiment, the children received a sweet prize. In the experimental study we focused on children of primary school age: this age is important because at this age educational activity replaces playing activity as the leading activity according to the modern periodization of mental development [Elkonin, 1951]. Also, 10-year-old schoolchildren

seem to be an interesting group for analysis since not all of them have fully formed the required shape rotation operations—therefore, the robot acts as a valuable assistant and mentor. This group, on the one hand, is focused on getting new knowledge and on the other hand, younger students are easily carried away by an interesting, playful, but rather difficult task.

5. Reaction to the emotional robot

According to the analysis of variance, **gestures** are the key factor that influences the attractiveness of the robot for the children: $F = 5.3$, $p = 0.02$ (Fig. 3). Among the participants who compared robots with different gesture strategies, this difference was correctly described in 31.6% of cases: where the subjects have noted that two robots differ in gestures (only 10.5% of the subjects reported that the robots did not differ). At the same time the participants have noted in the reports that the robot with emotional gestures seems to be more sociable, funny and helpful, explains more clearly and is generally kinder.

When varying the **emotionality of speech**, 63.2% of children correctly described the difference: they noted that robots spoke differently and used different words (only 15.8% of subjects reported that they did not notice a difference in the behavior of robots). The subjects also noted that robots have different characters: the robot with emotional speech is more kind and polite. The participants suggested that the robot resembles a child. Some children noted that the robot with neutral speech is more rigorous and should communicate in a “children’s language”. Other subjects described the robot with neutral speech as more modern, close to the speech of school children. Even though children more definitively noted the difference between robots with emotional and neutral speech, this did not affect their preferences. The differences between the groups of subjects who prefer a robot with emotional speech and a robot with neutral speech were not significant.

When combining the two experimental conditions (fully emotional and fully non-emotional robot), 64.3% of children correctly described the difference between the robots. None of the subjects said that the robots behaved the same way. The children noted that the robot with emotional speech and gestures is more kind, empathic, uses interesting words. Differences in evaluations of the robot with emotional gestures and speech as compared to the robot with neutral behavior are statistically significant ($t = 2.4$, $p = 0.02$) (Fig. 4).

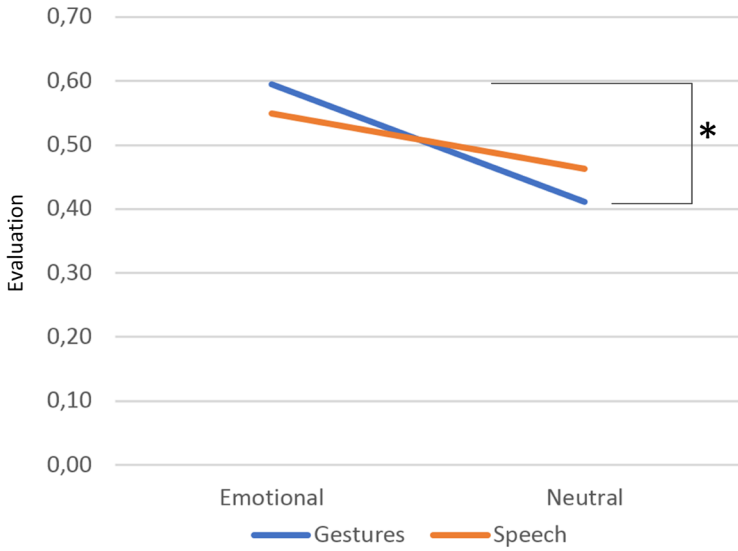


Fig. 3. Differences between robots with emotional performance in gestures or speech ($p < 0,05$)

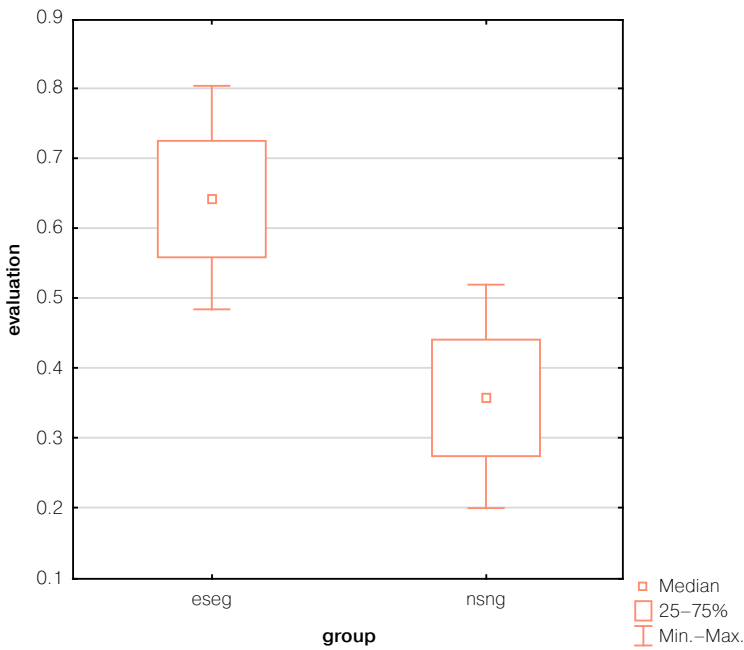


Fig. 4. Differences between the robot with emotional gestures and speech (*eseq*) and the robot with neutral gestures and speech (*nsng*) ($p < 0.05$)

Based on the obtained data, we can conclude that the difference in speech is more obvious to the participants. The difference in gestures is more difficult to recognize. At the same time, the difference in gestures has the major contribution to the attractiveness of a robot to a child: children more often prefer the robot that demonstrates emotional gestures, even if they cannot accurately describe the difference between the experimental conditions. The insensitivity of user preferences to varying emotionality in speech can be explained as follows. Regardless of the emotionality of speech, the robot performed its functions as a tutor—the robot informed the person of correct or erroneous actions and, as a result, led the person to solving the puzzle. The emotionality of speech was well differentiated by the users but was considered secondary for solving the problem. The use of gestures by the robot was recognized by the subjects less definitively, it was not essential for the process of solving the problem, but it influenced the assessment of the robot by a person. The emotionality of gestures did not affect the process of solving the puzzle, but it affected human evaluation of the robot. This may indicate a more subtle, implicit influence of gestures on the preferences of subjects.

The proportion between the number of subjects who noted the difference between the robots and the number of subjects who were able to describe this difference is quite interesting (100% of children said that there is a difference, 64.3% correctly described this difference). It can be assumed that the execution of the experiment at school affects children's answers: to the question *Did the robots behave differently?* the pupils usually assumed that the difference did exist. Perhaps it is advisable to evaluate the distribution of estimates on robots that are completely identical.

We assume that weak influence of the robot's emotional speech on the subjects' preferences can be explained by the theory of politeness [Brown, Levinson, 1987]. According to the theory, in a critical situation one uses fewer politeness strategies and more categorical instructions. Therefore, the restrained (non-emotional) speech behavior of the robot is acceptable and correct in the urgent situation of problem solving—and is not considered as rude. In addition, some subjects noted that the robot with non-emotional speech looks more “modern”.

6. Children's communicative behavior during interaction with the robot

To assess the affective reaction of children to the robots, we calculated the number of smiles in each of the experimental conditions. It turned out that the condition with emotional gestures and neutral speech has the greater number of smiles among other conditions. The robot using emotional speech also causes on average the greater number of smiles than the robot using neutral speech in combination with neutral gestures (Fig. 5). It is also important to note that the first robot in various experimental conditions caused more smiles.

While assisting a user in solving a puzzle, the robot tells which element to take and where to place it. Therefore, subjects can request feedback when they select a game element and when they put it on the game field. To control the correct choice of the game element, the children can raise and show the element to the robot. At the same time, respondents could perform actions specific to the feedback request, for

example, tilt their head, raise their eyebrows, tilt the whole body towards the robot. In addition, children could raise the element and shake it several times in the air, trying to attract the attention of the robot, or directly ask the robot *This?* (Fig. 6).

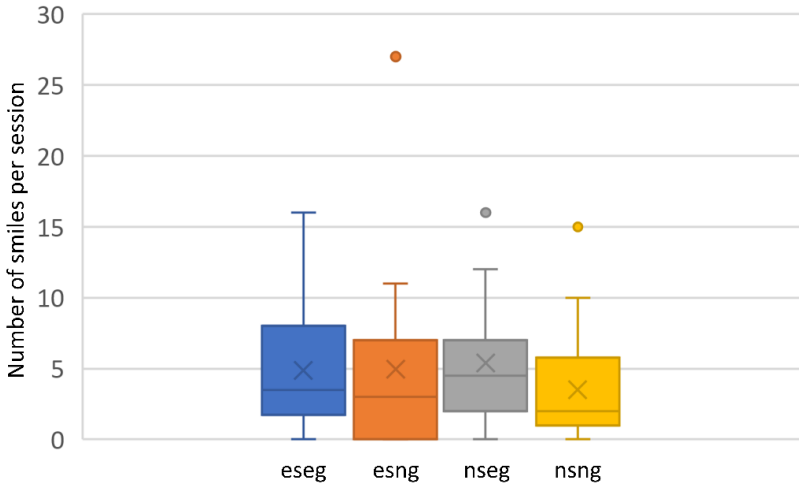


Fig. 5. Distribution of smiles in different experimental conditions: *eseg*—emotional speech and gestures; *esng*—emotional speech and neutral gestures; *nseg*—neutral speech and emotional gestures; *nsng*—neutral speech and gestures



Fig. 6. Children try to attract the robot's attention

Participants also mark up the placement of the game element on the playing field with special gestures to inform the robot that the action has ended.

- (i) Children retract hands from the playing field or clasp hands at the chest level (Fig. 7).
- (ii) After placing the game element, the participants additionally touched it with their fingers (with the same or different hand). The element did not move, so these gestures are communicative (Fig. 8).



Fig. 7. Children try to inform the robot that the action is complete



Fig. 8. Children try to inform the robot that the action is complete by additionally touching the element

7. Typical mistakes

Fish figure took the longest time to complete (on average 2 minute 16 seconds). The simplest figure for the participants was the *Triangle* (1 minute 32 seconds) (Fig. 9).

Fish was also the most complex figure—the participants made most of the mistakes while completing this figure—59 errors. In the parallelogram, 26 errors were made, in the ship—24, in the triangle 20 (Fig. 10).

The parallelogram is the only non-symmetrical element in the Tangram puzzle: after flipping it has a different shape, which is used in most of game figures. We found that the flipping of parallelogram causes the difficulties to many children. This may indicate the lack of skill to flip the figure in order to get a different shape. Sometimes it is more acceptable for subjects to change the arrangement of the game elements on the field than to flip the parallelogram to the other side (Fig. 11). The detection of such missing skill and their development may serve as one the basic strategies of r-Learning.

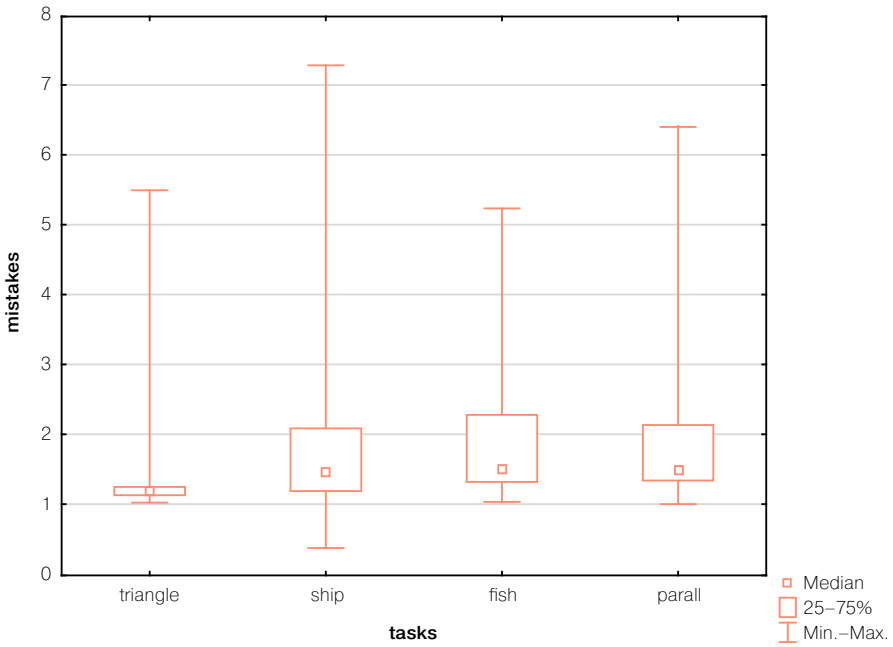


Fig. 9. Tasks solving time (min)

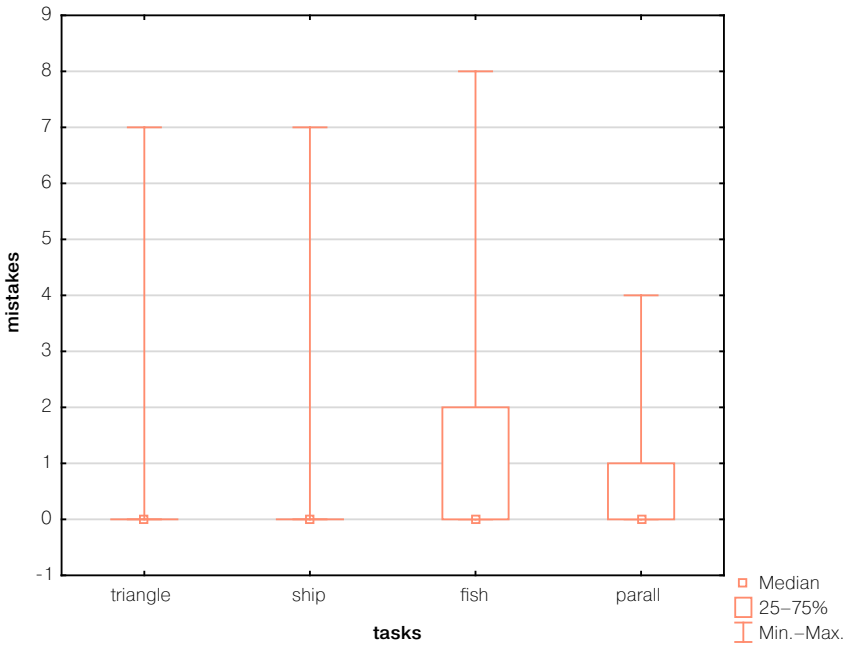


Fig. 10. The number of mistakes



Fig. 11. Mistakes with the parallelogram

The child cannot place the parallelogram on the left side of the figure without flipping it over to the other side. He prefers to move already assembled elements from the right side to the left side in order to place the parallelogram on the right in its current orientation

The second type of mistakes comes from the difficulties in formation of chunks of several game elements. For example, when assembling the *Ship*, children need to compose a big square of two big triangles and place it in the center of the contour. The subjects either cannot cope with the first instruction, or they cannot put the resulting combination in the right place of the contour.

The third type of mistakes is related to the ambiguity of the instructions. For example, when the robot asked a subject to put the parallelogram to the *middle* triangle (the middle-sized triangle in the game set), children often leaned it against the triangle that was *in the middle* of the present assembly (**Fig. 12**).

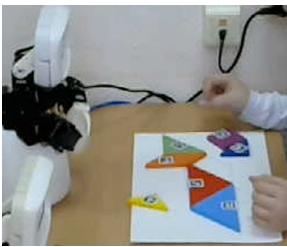


Fig. 12 (a). Correct position of the parallelogram



Fig. 12 (b). Incorrect positions of the parallelogram

Regular errors associated with misunderstanding or lack of skills can serve as the basis for the development of training programs with the use of robots. In this case, the

robots could detect the lack of formation or certain skills and offer simpler tasks for this particular skill until it is developed.

8. Conclusion

We conducted a study that provides further insights into the question of robot-to-human interaction. Emotional gestures were particularly important to form positive impression of the interaction with the robot. However, emotional speech causes positive facial reactions such as smiles. During the interaction, the subjects demonstrate stable behavioral patterns: mark the choice of the game element and the completion of manipulations with a game element. Difficulties that regularly arise during the solution can be used in developing strategies for teaching a person by robot. For example, the robot should not only prompt the figure and its location in the contour, he also should advise turning the figure over (in particular, a parallelogram), skipping instructions if the child wants to assemble a puzzle by himself. The robot also has to detect certain behavioral cues: track the user's mimic and gesture reactions, marking of the choice of the game element, completion of actions or expression of the overall emotional reaction of the child.

Summarizing, we can say that the interaction of the robot with children of primary school age contributes to their development and learning. When the robot acts as a source of knowledge, younger schoolchildren show increased interest. The use of robots as learning assistants is an interesting and promising direction in the development of robotics.

References

1. *Alemi, M., Meghdari, A., Ghazisaedy, M.* (2014). Employing humanoid robots for teaching English language in Iranian junior high-schools. In *International Journal of Humanoid Robotics*, 11(03), 1450022.
2. *Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., Tanaka, F.* (2018 a). Social robots for education: A review. In *Science robotics*. V. 3(21), eaat5954.
3. *Belpaeme, T., Vogt, P., Van den Berghe, R., Bergmann, K., Göksun, T., De Haas, M., Papadopoulos, F.* (2018b). In *Guidelines for designing social robots as second language tutors*. In *International Journal of Social Robotics*. V. 10(3), pp. 325–341.
4. *Brown, L. N., Howard, A. M.* (2014). The positive effects of verbal encouragement in mathematics education using a social robot. In *IEEE integrated STEM education conference*, pp. 1–5.
5. *Brown, P., Levinson, S. C.* (1987) *Politeness: Some universals in language usage*. Cambridge University Press.
6. *Clark H. H., Wilkes-Gibbs D.* (1986) Referring as a collaborative process, *Cognition*, pp. 1–39.
7. *Cuadrado, L. E. I., Riesco, Á. M., de la Paz López, F.* (2019). FER in Primary School Children for Affective Robot Tutors. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pp. 461–471. Springer, Cham.

8. *Elkonin D. B.* (1951) Thinking of primary school children [Myshlenie mladshogo shkol'nika]. Essays on child psychology [Ocherki psihologii detej]. Moscow
9. *Gordon, G., Breazeal, C., Engel, S.* (2015). Can children catch curiosity from a social robot?. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 91–98.
10. *Han, J. H., Jo, M. H., Jones, V., Jo, J. H.* (2008). Comparative study on the educational use of home robots for children. In Journal of Information Processing Systems, 4(4), 159–168.
11. *Han, J. H., Kim, D. H., Kim, J. W.* (2009). Physical learning activities with a teaching assistant robot in elementary school music class. In 2009 Fifth International Joint Conference on INC, IMS and IDC, pp. 1406–1410.
12. *Han, J., Kim, D.* (2009). r-Learning services for elementary school students with a teaching assistant robot. In 2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 255–256.
13. *Kanda, T., Hirano, T., Eaton, D., Ishiguro, H.* (2004). Interactive robots as social partners and peer tutors for children: A field trial. In Human-Computer Interaction. V. 19(1–2), pp. 61–84.
14. *Kelley J. F.* (1984) An iterative design methodology for user-friendly natural language office information applications, ACM Transactions on Office Information Systems, March 1984, V. 2(1), pp. 26–41.
15. *Kennedy, J., Baxter, P., Belpaeme, T.* (2015 a). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 67–74.
16. *Kennedy, J., Baxter, P., Senft, E., Belpaeme, T.* (2015b). Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions. In International conference on social robotics, pp. 327–336. Springer, Cham.
17. *Kirschner, D., Velik, R., Yahyanejad, S., Brandstötter, M., Hofbauer, M.* (2016). YuMi, come and play with Me! A collaborative robot for piecing together a tangram puzzle. In International Conference on Interactive Collaborative Robotics, pp. 243–251.
18. *Leite, I., Pereira, A.* (2007). Icat, the affective chess player. In Proceedings of the systems demonstrations. Second international conference on Affective Computing and Intelligent Interaction (ACII 2007), pp. 29–33).
19. *Leyzberg, D., Spaulding, S., Toneva, M., Scassellati, B.* (2012). The physical presence of a robot tutor increases cognitive learning gains. In Proceedings of the annual meeting of the cognitive science society (Vol. 34, No. 34).
20. *Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., Dong, J.-J.* (2013). A review of the applicability of robots in education. In British Journal of Educational Technology. V. 1, pp.13.
21. *Park, H. W., Rosenberg-Kima, R., Rosenberg, M., Gordon, G., Breazeal, C.* (2017). Growing growth mindset with a social robot peer. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 137–145.
22. *Pereira, A., Leite, I., Mascarenhas, S., Martinho, C., Paiva, A.* (2010). Using empathy to improve human-robot relationships. In International Conference on Human-Robot Personal Relationship, pp. 130–138. Springer, Berlin, Heidelberg.

23. *Schodde, T., Hoffmann, L., Kopp, S.* (2017). How to manage affective state in child-robot tutoring interactions? In 2017 International Conference on Companion Technology, pp. 1–6.
24. *Shiomi, M., Kanda, T., Howley, I., Hayashi, K., Hagita, N.* (2015). Can a social robot stimulate science curiosity in classrooms?. In International Journal of Social Robotics. V. 7(5), pp. 641–652.
25. *Shore, T., Androulakaki, T., Skantze, G.* (2018). KTH Tangrams: A Dataset for Research on Alignment and Conceptual Pacts in Task-Oriented Dialogue. In Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 768–775.
26. *van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., Leseman, P.* (2019). Social robots for language learning: A review. In Review of Educational Research. V. 89(2), pp. 259–295.
27. *Wainer, J., Feil-Seifer, D. J., Shell, D. A., Mataric, M. J.* (2006). The role of physical embodiment in human-robot interaction. In ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication, pp. 117–122.
28. *Zinina A., Zaydelman L., Arinkin N., Kotov A.* (2019). Non-verbal behavior of the robot companion: a contribution to the likeability. In Procedia Computer Science. Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA. Elsevier.

Abstracts

EXPLORING PRETRAINED MODELS FOR JOINT MORPHO-SYNTACTIC PARSING OF RUSSIAN

Anastasyev D. G. (dan-anastasev@yandex-team.ru), Yandex, Moscow, Russia

In this paper, we build a joint morpho-syntactic parser for Russian. We describe a method to train a joint model which is significantly faster and as accurate as a traditional pipeline of models. We explore various ways to encode the word-level information and how they can affect the parser's performance. To this end, we utilize learned from scratch character-level word embeddings and grammeme embeddings that have shown state-of-the-art results for similar tasks for Russian in the past. We compare them with the pretrained contextualized word embeddings, such as ELMo and BERT, known to lead to the breakthrough in miscellaneous tasks in English. As a result, we prove that their usage can significantly improve parsing quality.

WORD2VEC NOT DEAD: PREDICTING HYPERNYMS OF CO-HYPONYMS IS BETTER THAN READING DEFINITIONS

Arefyev N. V. (nick.arefyev@gmail.com)^{1,2,3}, **Fedoseev M. V.**

(maxim.fedoseev13@gmail.com)¹, **Kabanov A. V.** (arshehremen@gmail.com)¹, **Zizov V. S.**

(vzs815@gmail.com)¹; ¹Lomonosov Moscow State University, ²Samsung R&D Institute Russia,

³National Research University Higher School of Economics, Moscow, Russian Federation

Expert-built lexical resources are known to provide information of good quality for the cost of low coverage. This property limits their applicability in modern NLP applications. Building descriptions of lexical-semantic relations manually in sufficient volume requires a huge amount of qualified human labour. However, given some initial version of a taxonomy is already built, automatic or semi-automatic taxonomy enrichment systems can greatly reduce the required efforts. We propose and experiment with two approaches to taxonomy enrichment, one utilizing information from word definitions and another from word usages, and also a combination of them. The first method retrieves co-hyponyms for the target word from distributional semantic models (word2vec) or language models (XLM-R), then looks for hypernyms of co-hyponyms in the taxonomy. The second method tries to extract hypernyms directly from Wiktionary definitions.

The proposed methods were evaluated on the Dialogue-2020 shared task on taxonomy enrichment. We found that predicting hypernyms of co-hyponyms achieves better results in this task. The combination of both methods improves results further and is among 3 best-performing systems for verbs. An important part of the work is detailed qualitative and error analysis of the proposed methods, which provide interesting observations of their behaviour and ideas for the future work.

EXPLORING SEMANTIC CONCRETENESS AND ABSTRACTNESS FOR METAPHOR IDENTIFICATION AND BEYOND

Badryzlova Yu. G. (yuliya.badryzlova@gmail.com), National Research University Higher School of Economics, Moscow, Russia

The paper presents a method for computing indexes of semantic concreteness and abstractness in two languages (Russian and English). These indexes are used in metaphor identification experiments in both languages; the results are either comparable to or surpass previous work and the baselines. We analyze the obtained indexes of concreteness and abstractness to see how they align with the linguistic intuitions about the corresponding semantic categories. The results of the analysis may have broader implications for computational studies of the semantics of concreteness and abstractness.

STYLE DYNAMICS OF THE RUSSIAN WRITTEN SPEECH OF THE 19TH CENTURY: A CORPUS STUDY

Baranov A. N. (baranov_anatoly@hotmail.com), **Dobrovol'skij D. O.** (dobrovol'skij@gmail.com), Russian Language Institute of the RAS, Institute of Linguistics of the RAS, Moscow, Russia

The starting point of the present paper is the hypothesis that the distribution of discursive words characterizes the trends in the development of the writing style of the 19th century. The paper presents and discusses the results of an experiment based on the data of the Russian National Corpus on the frequency of using discursive words with the semantics of epistemic modality, such as *konechno*, *razumeetsya* (both roughly meaning 'of course'), *po-vidimomu* 'apparently', *kak kazhetsya*, *kazalos' by* (both ≈ 'it would seem'), *naverno* ≈ 'as it were', *veroyatno* 'probably', *pozhaluy* ≈ 'maybe', *deystvitel'no* 'really', etc. We show that the frequency of this group of expressions increases in the second half of the 19th century. A similar trend is also observed for some syntactic constructions with the same semantics: (*ya dumayu*, *chto...* '(I) think that...'; (*ya schitayu*, *chto...* '(I) believe that...'; (*mne kazhetsya*, *chto...* 'it seems to me that'. The revealed regularity is considered as a discursive practice in changing the style of fiction, which consisted in expanding the modus part of the utterance as compared to the earlier period.

The discursive practice of expanding the modus was inherent only to a group of innovative writers (first of all, F. M. Dostoevsky, M. E. Saltykov-Shchedrin, L. N. Tolstoy, I. A. Goncharov, A. F. Pisemsky, P. I. Melnikov-Pechersky, N. S. Leskov, and I. S. Turgenev), who, however, due to their talent, social significance, and the number of published texts, had a significant impact on the language of fiction.

The task of studying the dynamics of artistic style is to identify and describe a set of discursive practices that establish written discourse as such.

WEB-CORPUS AS A TOOL FOR LINGUISTIC RESEARCH: DIFFERENTIATION, AUTHORIZATION, THEMATIC BIASES (OR CORPORA WE WANT SO MUCH TO BELIEVE)

Belikov V., ABBYY Lab, MIPT, **Selegey V.** (vladimir.selegey@abbyy.com), ABBYY, **Selegey D.** (daniil_s@abbyy.com), ABBYY Lab, MIPT

The paper presents the General Internet Corpus of the Russian Language (GICR) as a tool for linguistic research. Problems are identified that are common to any WEB-corpus that affect the reliability of such research. Among the problems considered: the importance of taking into account sociolinguistic variability, the influence of falsely attributed texts, thematic biases, the prospects and disadvantages of new methods for corpora output aggregation. A distinctive feature of our approach is the emphasis on linguistic significance, reliability, and interpretability of the results obtained.

MODELING LEMMA FREQUENCY BANDS FOR LEXICAL COMPLEXITY ASSESSMENT OF RUSSIAN TEXTS

Blinova O. V. (o.blinova@spbu.ru; ovblinova@hse.ru), Saint Petersburg State University; HSE University, Saint Petersburg, Russia, **Tarasov N. A.** (tarasovn2468@yandex.ru), **Modina V. V.** (st072157@student.spbu.ru), **Blekanov I. S.** (i.blekanov@spbu.ru), Saint Petersburg State University, Saint Petersburg, Russia

The paper is devoted to the problem of modeling general-language frequency using data of large Russian corpora. Our goal is to develop a methodology for forming a consolidated frequency list which in the future can be used for assessing lexical complexity of Russian texts.

We compared 4 frequency lists developed from 4 corpora (Russian National Corpus, ruTenTen11, Araneum Russicum III Maximum, Taiga). Firstly, we applied rank correlation analysis. Secondly, we used the measures "coverage" and "enrichment". Thirdly, we applied the measure "sum of minimal frequencies". We found that there are significant differences between the compared frequency lists both in ranking and in relative frequencies. The application of the "coverage" measure showed that frequency lists are by no means substitutable. Therefore, none of the corpora in question can be excluded when compiling a consolidated frequency list.

For a more detailed comparison of frequency lists for different frequency bands, the ranked frequency list, based on RNC data, was divided into 4 equal parts. Then 4 random samples (containing 20 lemmas from each quartile) were formed.

Due to the wide range of values, accepted by ipm measure, relative frequency values are difficult to interpret. In addition, there are no reliable thresholds separating high-frequency, mid-frequency, and low-frequency lemmas. Meanwhile, to assess the lexical complexity of texts, it is useful to have a convenient way of distributing lemmas with certain frequencies over the bands of the frequency list. Therefore, we decided to assign lemmas “Zipf-values”, which made the frequency data interpretable because the range of measure values is small.

The result of our work will be a publicly accessible reference resource called “Frequentator”, which will allow to obtain interpretable information about the frequency of Russian words.

THE RUSSIAN LANGUAGE PIPELINE IN THE LIMA MULTILINGUAL ANALYZER

Bocharov V. V. (victor.bocharov@cea.fr), **de Chalendar G.** (gael.de-chalendar@cea.fr),
CEA LIST, LASTI, Gif-sur-Yvette, France

In this paper we describe the implementation of Russian language pipeline in LIMA multilingual analyzer and the results obtained in GramEval-2020 shared task. LIMA is a modular pipeline that implements rule-based and machine learning analysis components. Russian language pipeline includes deep neural networks based modules for tokenization, sentence segmentation, part of speech tagging, lemmatization and dependency parsing. Part of speech tags, feature tags and dependency trees conform to Universal Dependencies rules.

FULL-FLEDGED SEMANTIC ANALYSIS AS A TOOL FOR RESOLVING TRIANGLE-COPA SOCIAL SCENARIOS

Boguslavsky I. M. (bogus@iitp.ru), A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; Universidad Politécnica de Madrid, Madrid, Spain, **Dikonov V. G.** (sdcikonov@mail.ru), **Frolova T. I.** (tfrolova@gmail.com), **Iomdin L. L.** (iomdin@gmail.com), **Lazursky A. V.** (lazursky@mail.ru), **Rygaev I. P.** (irygaev@gmail.com), **Timoshenko S. P.** (nyrestein@gmail.com), A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

Text interpretation often requires common sense knowledge and reasoning. A convenient tool for developing methods of common sense reasoning are special sets of challenge problems whose interpretation requires sophisticated reasoning. An interesting example is a recently published data set called Triangle Choice of Plausible Alternatives (Triangle-COPA), which contains 100 multiple-choice problems that test the interpretation of social scenarios. Each problem includes a statement and two alternatives. The task is to identify the more plausible alternative. For processing Triangle-COPA data we use SemETAP, a general purpose semantic analyzer. We implement the full scenario of NL understanding starting from NL texts and not from manually composed simplified logical formulas, which is a common practice in logic-based approaches to common sense reasoning. We produce Enhanced Semantic Structures of the statement and both alternatives and check which alternative manifests more semantic agreement with the statement in terms of inferences.

GENERATING TRAINING DATA FOR WORD SENSE DISAMBIGUATION IN RUSSIAN

Bolshina A. S. (angelina_ku@mail.ru), Lomonosov Moscow State University, Moscow, Russia, **Loukachevitch N. V.** (louk_nat@mail.ru), Kazan Federal University, Kazan, Russia

The best approaches in Word Sense Disambiguation (WSD) are supervised and rely on large amounts of hand-labelled data, which is not always available and costly to create. For the Russian language there is no sense-tagged resource of the size sufficient to train supervised word sense disambiguation algorithms. In our work we describe an approach that is used to create an automatically labelled collection based on the monosemous relatives (related unambiguous entries). The main contribution of our work is that we extracted monosemous relatives that can be located at relatively long distances from a target ambiguous word and ranked them according to the similarity measure to the target sense. The selected candidates are then used to extract training samples from the news corpus. We evaluated word sense disambiguation models based on a nearest neighbor classification on BERT and ELMo embeddings. Our work relies on the Russian wordnet RuWordNet.

REFERENTIAL PHENOMENA IN SPEAKERS KINETIC CHANNELS

Budennaya E. V. (jane.sdrv@gmail.com), Higher School of Economics / Institute of Linguistics, RAS, Moscow, Russia, **Evdokimova A. A.** (arochka@gmail.com), Institute of Linguistics, RAS, Moscow, Russia, **Nikolaeva Ju. V.** (julianikk@gmail.com), Lomonosov Moscow State University, Moscow, Russia, **Sukhova N. V.** (sukhova.natalya@gmail.com), National University of Science and Technology MISiS, Moscow, Russia

The article addresses the relation of referential expressions and co-occurring kinetic phenomena (hand and head gestures) on the material of the RUPEX multimodal corpus. The results reflect significant differences in how individual movements and gestures are aligned with two major types of reference (full NPs vs. reduced expressions). It was initially assumed that full NPs are more often accompanied by a gesture. Our data support this hypothesis not only through the material of hand gestures, but also through head movements. Moreover, full NPs are more likely to be accompanied by downward movements in both manual and cephalic channels, as well as by metadiscourse gestures, in comparison to reduced referential units (personal and demonstrative pronouns). In addition, pronouns are more likely to be aligned with pointing hand gestures and zero reference is often accompanied by descriptive hand gestures. However, the kinetic behavior of the interlocutors is determined by a variety of factors, including the topic of the conversation, which predisposes to certain types of gestures and the relative position of the interlocutors.

WHAT DO WE LEARN FROM MISTAKES: PROCESSING DIFFICULTIES WITH FREQUENTLY MISSPELLED WORDS

Chernova D. A. (chernovadasha@yandex.ru), St. Petersburg State University, St. Petersburg, Russia, **Alexeeva S. V.** (mail@s-alexeeva.ru), St. Petersburg State University, St. Petersburg, Russia, **Slioussar N. A.** (slioussar@gmail.com), HSE, Moscow, and St. Petersburg State University, St. Petersburg, Russia

Even if we know how to spell, we often see words misspelled by other people — especially nowadays when we constantly read unedited texts on social media and in personal messages. In this paper, we present two experiments showing that the incidence of orthographic errors reduces the quality of lexical representations in the mental lexicon—even if one knows how to spell a word, repeated exposure to incorrect spellings blurs its orthographical representation and weakens the connection between form and meaning. As a result, it is more difficult to judge whether the word is spelled correctly, and—more surprisingly—it takes more time to read the word even when there are no errors. We show that when all other factors are balanced the effect of misspellings is more pronounced for the words with lower frequency.

We compare our results with the only previous study addressing the problem of misspellings' influence on the processing of correctly spelled words—it was conducted on the English data. It may be interesting to explore this issue in a cross-linguistic perspective. In this study, we turn to Russian, which differs from English by a more transparent orthography. Much larger corpora of unedited texts are available for English than for Russian, but, using a different way to estimate the incidence of misspellings, we obtained similar results and could also make some novel generalizations.

In Experiment 1 we selected 44 words that are frequently misspelled and presented in two conditions (with or without spelling errors) and were distributed across two experimental lists. For every word, participants were asked to determine whether it is spelled correctly or not. The frequency of the word and the relative frequency of its misspelled occurrences significantly influenced the number of incorrect responses: not only it takes longer to read frequently misspelled words, it is also more difficult to decide whether they are spelled correctly.

In Experiment 2 we selected 30 words from the materials of Experiment 1 and for every selected word, we found a pair that is matched for length and frequency, but is rarely misspelled due to its orthographic transparency. We used a lexical decision task, presenting these 60 words in the correct spelling, as well as 60 nonwords. We used LMMs for statistics. Firstly, the word type factor was significant: it takes more time to recognize a frequently misspelled word, which replicates the results obtained for English. Secondly, the interaction between the word type factor and the frequency factor was significant: the effect of misspellings was more pronounced for the words of lower frequency. We can conclude that high frequency words have more robust representations that resist blurring more efficiently than low frequency ones.

Finally, we conducted a separate analysis showing that the number of incorrect responses in Experiment 1 correlates with RTs in Experiment 2. Thus, whether we consciously try to find an error or simply read words orthographic representations blurred due to exposure to frequent misspellings make the task more difficult.

ON THE SECONDARY IMPERFECTIVATION OF PO-PERFECTIVES IN RUSSIAN

Chuikova O. Iu. (ochuikova@herzen.spb.ru, oxana.chuykova@gmail.com), Herzen State Pedagogical University of Russia, St. Petersburg, Russia

The paper deals with a number of characteristics of the secondary imperfectivation of *po*-perfectives in Russian. The study is based on the analysis of the level of imperfectivability of Russian perfective verbs with the prefix *po*- compared to a number of other prefixed perfective verb groups (e. g. the verbs with such perfectivizing prefixes as *na*-, *za*-, etc.) according to the Dictionary of Russian Language, the Russian National Corpus and the Russian-language Internet (Runet). It is shown that the discussed perfective verb group is specific as a whole as well as with respect to its subgroups, i.e., deperfective perfective verbs and morphologically marked Aktionsarten. *Po*-perfectives demonstrate a low average imperfectivability in comparison to corresponding figures for other prefixed verb groups. For the subgroup of deperfective (formed from perfective stems) verbs the level of imperfectivability is also unusually low. The delimitative Aktionsart shows a higher imperfectivability than other morphologically marked Aktionsarten do. Possible explanations for the peculiarities of imperfectivability of *po*-perfectives rather confirm than contradict the hypothesis about the regularity of the secondary imperfectivation in Russian.

A SIMPLE SOLUTION FOR THE TAXONOMY ENRICHMENT TASK: DISCOVERING HYPERNYMS USING NEAREST NEIGHBOR SEARCH

Dale D. S. (dale.david@yandex.ru), Yandex, Moscow, Russia

In this paper, we present the system we used in the Taxonomy Enrichment for the Russian Language evaluation campaign. The goal of this challenge is to predict hypernyms for the words not included in the taxonomy. Our approach was to generate and score candidate hypernyms by word embedding similarity of the input words and concepts already in the taxonomy. Despite being very simple, our system was ranked first on the verbs track.

RENERSANS: RELATION EXTRACTION AND NAMED ENTITY RECOGNITION AS SEQUENCE ANNOTATION

Davletov A. A. (davletov-aa@ranepa.ru)^{1,3}, **Gordeev D. I.** (gordeev-di@ranepa.ru)¹,

Rey A. I. (rey-ai@ranepa.ru)¹, **Arefyev N. V.** (nick.arefyev@gmail.com)^{2,3,4}; ¹RANEPA,

Moscow, Russia, ²Samsung R&D Institute Russia, Moscow, Russia, ³Lomonosov Moscow

State University, Moscow, Russia, ⁴National Research University Higher School of Economics

Moscow, Russian Federation

In this work we present our system for RuREBus shared task held together with Dialog 2020 conference. The task consisted of 3 subtasks: named entity recognition, relation extraction with provided named entity tags and end-to-end relation extraction. Our system took the first and the second place in the first and the second subtasks respectively. For the third subtask we submitted our solution only in the post-evaluation phase, however, it was among the top 2 best performing systems. The systems for all tasks are based on Transformer models. Relation extraction was solved as a sequence labelling problem. We also used joint task named entity and relation extraction learning.

STABILITY OF TOPIC MODELING VIA MODALITY REGULARIZATION

Derbanosov R. (derbanosov@gmail.com), National Research University Higher School of Economics, Moscow, Russia, **Bakhanova M.** (marybakhanova@gmail.com), Skolkovo Institute of Science and Technology, National Research University Higher School of Economics, Moscow, Russia

Probabilistic topic modeling is a tool for statistical text analysis that can give us information about the inner structure of a large corpus of documents. The most popular models—Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation—produce topics in a form of discrete distributions over the set of all words of the corpus. They build topics using an iterative algorithm that starts from some random initialization and optimizes a loss function. One of the main problems of topic modeling is sensitivity to random initialization that means producing significantly different solutions from different initial points.

Several studies showed that side information about documents may improve the overall quality of a topic model. In this paper, we consider the use of additional information in the context of the stability problem. We represent auxiliary information as an additional modality and use BigARTM library in order to perform experiments on several text collections. We show that using side information as an additional modality improves topics stability without significant quality loss of the model.

DIFFERENTIAL SEMANTIC SKETCHES FOR RUSSIAN INTERNET-CORPORA

Detkova J. (julia.detkova@abbyy.com), ABBYY Lab MIPT, **Novitskiy V.** (valeriy.novitskiy@abbyy.com), **Petrova M.** (m.petrova@abbyy.com), **Selegey V.** (vladimir.selegey@abbyy.com), ABBYY

The current paper suggests a new representation type of word collocations—the semantic sketches. It was first tested on one of the subcorpora of the General Internet-Corpus of Russian. The semantic sketches continue the idea of word sketches based on grammatical relations between words and expand it by adding the semantic information—word meanings and semantic relations between words. Moreover, the sketches can be additionally provided with metatextual characteristics.

Certainly, building such sketches demands the semantic markup of the corpora. Therefore, we have used partial semantic analysis of the Comreno parser for our purposes. The paper demonstrates the examples of the sketches, provides the quality evaluation of the markup they are based on, and shows the advantages and disadvantages of the given approach.

ANNOTATING AND EXPLORING CODE-SWITCHING IN FOUR CORPORA OF MINORITY LANGUAGES OF RUSSIA

Dyachkov V. V. (hyppocentaurus@mail.ru), Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia, **Khomchenkova I. A.** (irina.khomchenkova@yandex.ru), Russian Language Institute, Russian Academy of Sciences & Lomonosov Moscow State University, Moscow, Russia, **Pleshak P. S.** (polinapleshak@yandex.ru), University of Maryland, College Park, USA, **Stoyanova N. M.** (stoyanova@yandex.ru), Russian Language Institute, Russian Academy of Sciences & NRU Higher School of Economics, Moscow, Russia

This paper describes code-switching with Russian in four spoken corpora of minority languages of Russia: two Uralic ones (Hill Mari and Moksha) and two Tungusic ones (Nanai and Ulch). All narrators are bilinguals, fluent both in the indigenous language (IL) and in Russian; all the corpora are comparable in size and genres (small field collections of spontaneous oral texts, produced under the instruction to speak IL); the languages are comparable in structural (dis) similarity with Russian. The only difference concerns language dominance and the degree of language shift across the communities. The aim of the paper is to capture how the degree of language shift influences the strategy of code-switching attested in each of the corpora using a minimal additional annotation of code-switching. We added to each corpus a uniform annotation of code-switching of two types: first, a simple semi-automatic word-by-word language annotation (IL vs. Russian), second, a manual annotation of structural code-switching types (for smaller sub-corpora). We compared several macro-parameters of code-switching by applying some existing simple measures of code-switching to the data of annotation 1. Then we compared the rates of different structural types of code-switching, basing on annotation 2. The

results of the study, on the one hand, verify and enhance the existing generalizations on how language shift influences code-switching strategies, on the other hand, they show that even a very simple annotation of code-switching integrated to an existing field records collection appears to be very informative in code-switching studies.

THE RUSSIAN CONSTRUCTICON: A NEW LINGUISTIC RESOURCE, ITS DESIGN AND KEY CHARACTERISTICS

Endresen A. A. (anna.endresen@uit.no) , UiT The Arctic University of Norway, **Zhukova V. A.** (valentina.zh96@gmail.com), National Research University Higher School of Economics, **Mordashova D. D.** (mordashova.d@yandex.ru), Institute of Linguistics (Russian Academy of Sciences), Lomonosov Moscow State University, **Rakhilina E. V.** (rakhilina@gmail.com), **Lyashevskaya O. N.** (olesar@yandex.ru), National Research University Higher School of Economics; Vinogradov Institute for Russian language (Russian Academy of Sciences)

We present a new open-access electronic resource named the Russian Constructicon that offers a searchable database of Russian constructions accompanied by descriptions of their properties and illustrated with corpus examples. The project was carried out over the period 2016–2020 and at present contains an inventory of over 2200 multi-word constructions of Contemporary Standard Russian. We prioritize “partially schematic” constructions that lie between the two extremes of fully compositional syntactic sequences on the one hand and fully idiomatic (phrasological) expressions on the other hand. Constructions of this type are difficult to account for in terms of either lexicon or grammar alone, and are often underrepresented in reference works of Russian. A typical construction in our database contains a fixed part (anchor words) and an open slot that can be filled with a restricted set of lexemes. In this paper we first focus on key characteristics of this resource that make it different from existing constructicons of other languages. Second, we describe how the new interface will be designed and how it will serve the needs of both linguists and L2 learners of Russian. In particular, we discuss various search possibilities relevant for different users and those parameters that are available for specifying the retrieval output. An example of an entry is given to show how the information about each construction is structured and presented. Third, we provide an overview of our multi-level semantic classification of constructions. We argue that our system of semantic and syntactic tags subdivides our items into meaningful classes and smaller groups and eventually facilitates the identification of constructional families and clusters. This methodology works well in turning the initial list of constructions as unrelated units into a structured network and makes it possible to refine and expand the collected inventory of constructions in a systematic way.

QUANTILE-BASED APPROACH TO ESTIMATING COGNITIVE TEXT COMPLEXITY

Eremeev M. A. (maks5507@yandex.ru), Lomonosov Moscow State University (MSU), Moscow, Russia, **Vorontsov K. V.** (vokov@forecsys.ru), Moscow Institute of Physics and Technology (MIPT), Moscow, Russia

This paper introduces an approach to measuring the cognitive complexity of texts on various language levels. While standard readability indices are based on the linear combination of primary statistics, our general approach allows us to estimate complexity on morphological, lexical, syntactic, and discursive levels. Each model is defined by the tokens for the specific language level and the complexity function of a single token. We then use the reference collection of moderately complex texts and the quantile-based approach to spot the abnormally rare tokens. The proposed supervised ensemble, based on the ElasticNet model, incorporates models from all language levels. Having collected a labeled dataset through crowdsourcing, consisting of pairs of articles from the Russian Wikipedia, we consider several models and ensembles and compare them to common baselines. Suggested models are flexible due to the freedom in choosing the reference collection. The described experiments confirm the competitiveness of the proposed approach, as the ensembles demonstrate the best target metric value.

SPARQL QUERY GENERATION FOR COMPLEX QUESTION ANSWERING WITH BERT AND BILSTM-BASED MODEL

Evseev D. A. (dmitrij.euseew@yandex.ru), **Arkhipov M. Yu.** (arkhipov@yahoo.com), Neural Networks and Deep Learning Lab, Moscow Institute of Physics and Technology, Moscow, Russia

In this paper we describe question answering system for answering of complex questions over Wikidata knowledge base. Unlike simple questions, which require extraction of single fact from the knowledge base, complex questions are based on more than one triplet and need logical or comparative reasoning. The proposed question answering system translates a natural language question into a query in SPARQL language, execution of which gives an answer. The system includes the models which define the SPARQL query template corresponding to the question and then fill the slots in the template with entities, relations and numerical values. For entity detection we use BERT-based sequence labelling model. Ranking of candidate relations is performed in two steps with BiLSTM and BERT-based models. The proposed models are the first solution for LC-QUAD2.0 dataset. The system is capable of answering complex questions which involve comparative or boolean reasoning.

COMBINING FACTS, SEMANTIC ROLES AND SENTIMENT LEXICON IN A GENERATIVE MODEL FOR OPINION MINING

Feldman D. G. (daniil.feldman@phystech.edu), Moscow Institute of Physics and Technology, Moscow, Russia, **Sadekova T. R.** (str12.01.94@gmail.com), Lomonosov Moscow State University, Moscow, Russia, **Vorontsov K. V.** (vokov@forecsys.ru), Moscow Institute of Physics and Technology, Moscow, Russia

Opinion mining is a popular task, that is applied, for example, to determine news polarisation and identify product review classes. Our task is unsupervised clusterization of opinionated texts, in particular news on political events. Many papers that tackle this issue use generative models based on lexical features. Our goal is to determine the entities defying an opinion amongst lexical, syntactic and semantic features as well as their compositions. More specifically, we test the hypothesis that an opinion is determined by the composition of the mentioned facts (SPO triples), the semantic roles of the words and the sentiment lexicon used in it. In this paper we formalise this task and prove that using a composition of the above features provides the best quality when clusterising opinionated texts. To test this hypothesis we have gathered and labelled two corpuses of news on political events and proposed a set of unsupervised algorithms for extracting the features.

EVENT2MIND FOR RUSSIAN: UNDERSTANDING EMOTIONS AND INTENTS IN TEXTS. CORPUS AND MODEL FOR EVALUATION

Fenogenova A. S. (alenush@gmail.com), Sberbank, Moscow, Russia, **Tikhonova M. I.** (m_tikhonova94@mail.ru), Sberbank; National Research University Higher School of Economics, Moscow, Russia, **Filipetskaya D. V.** (dafi913@yandex.ru), Moscow Institute of Physics and Technology, Moscow, Russia, **Mironenko F. D.** (fomius2000@yandex.ru), Saint Petersburg State University, Saint-Petersburg, Russia, **Tabisheva A. O.** (anastasiatabisheva@yandex.ru), National Research University Higher School of Economics, Moscow, Russia

The paper provides a comprehensive overview of the corpus for the Russian language for the commonsense inference task. Namely, we construct event phrases, which cover a wide range of everyday situations with labelled intents and reactions of the event main participant and emotions of other people involved. The dataset consists of two parts: a crowdsourced corpus of 6,756 examples from Russian sources and a translated into Russian part of the original corpus of 23,409 examples. Apart from this, we use the collected data in order to train the event2mind model for the Russian language. The paper presents careful description of the best Russian model and the results of the conducted experiments.

IMPLICIT LOGICAL-SEMANTIC RELATIONS AND A METHOD OF THEIR IDENTIFICATION IN PARALLEL TEXTS

Goncharov A. A. (a.gonch48@gmail.com), Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia, **Inkova O. Yu.** (Olga.Inkova@unige.ch), Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland

One of the main characteristics of logical-semantic relations (LSRs) between two fragments of a text is that these relations can be either explicit (expressed by some marker, e.g. a connective) or implicit (derived from the interrelation of these fragments' semantics). Since implicit LSRs do not have any marker, it is difficult to find them in a text (whether automatically or not). In this paper, approaches to analysing implicit LSRs are compared, an original definition for them is offered and differences between implicit LSRs and LSRs expressed by non-prototypical means are described. A method is proposed to identify implicit LSRs using a parallel corpus and a supracorpora database of connectives. Based on the well-known statement that LSRs can be explicitated by adding connectives in the translation, it is argued here that through selecting pairs in which fragments where a connective is used to express an LSR in the translation correspond to those containing any of the translation stimuli standard for this connective in the source language, it is possible to get an array of contexts in which this LSR is implicit in the source text (or expressed by means other than connectives). This method is then applied to study the French causal connectives *car*, *parce que* and *puisque* using a Russian-French parallel corpus. The corpus data are analysed to obtain information about LSRs particularly about cases where the causal LSR in Russian is implicit, as well as about the use of causal connectives in French. These results are used to show that the method proposed allows to quickly create a representative array of contexts with implicit LSRs, which can be useful in both text analysis and in machine learning.

ASPECTUAL TRIPLETS OF THE RUSSIAN VERB IN DIACHRONY: EVIDENCE FROM THE RUSSIAN NATIONAL CORPUS

Gorbova E. V. (elena-gorbova@yandex.ru), Herzen State Pedagogical University of Russia, St. Petersburg, Russia

The paper deals with the so-called aspectual triplets of the Russian verb. Based on the data from the Russian National Corpus, it proposes a diachronic method to study triplets as well as a two-component model of the Russian aspect as an alternative to the traditional word-based classification model. The first component of the model is a morphological mechanism of the imperfectizing suffixation of prefixed verbs that is inflectional (*ras-kry-t*^{PFV} — *ras-kry-va-t*^{IPFV2} 'disclose, reveal'), but has a limited scope of action (prefixed verbs only). The second component of the model is the actionality (lexical aspect) with a maximal scope. Related to the verb class as a whole, it is especially crucial for non-prefixed simplexes. Actionality enables the functioning and perfective / imperfective characterization of simplexes which do not fall under the inflectional grammatical aspect. The analysis of ten biimperfective triplets resulted in several observations and conclusions. One of them concerns the role of a 'joker', which all imperfective simplexes (IPFV₁) have in the aspectual triplets as (quasi)synonyms for corresponding secondary imperfectives (IPFV₂). A working hypothesis on the predominance of IPFV₁ over PFV in every triplet, based on the broader polysemy of the former, has not been confirmed. However, the two-component model has explanatory power for the cases of reverse frequency (PFV over IPFV₁) through its lexical aspect component. Another working hypothesis on a possible increase or a decrease in the number of secondary imperfectives in diachrony was partially confirmed — an increase was noted for the 20/21st century.

RELATION EXTRACTION DATASET FOR THE RUSSIAN

Gordeev D. I. (gordeev-di@ranepa.ru), **Davletov A. A.** (davletov-aa@ranepa.ru), **Rey A. I.** (rey-ai@ranepa.ru), **Akzhigitova G. R.** (akzhigitova-gr@ranepa.ru), **Geymbukh G. A.** (geymbukh-ga@ranepa.ru), RANEPa, Moscow, Russia

There are few existing relation extraction datasets for the Russian language and they contain a rather small number of examples. Thus, we decided to create a new Ontonotes-based named entities and relation extraction sentence-level dataset called RURED. The dataset contains more than 500 annotated texts and more than 5,000 labelled relations. We also publish baseline models for relation extraction and named entity recognition trained on the dataset. Our models achieve 0.85 for named entity recognition and 0.78 for relation extraction in F1-score.

DIALOGUE MANAGEMENT USING EXTENDED DISCOURSE TREES

Ilvovsky D. A. (dilvovsky@hse.ru), Kazan Federal University, Kazan, Russia;
National Research University Higher School of Economics Moscow, Russia,
Galitsky B. A. (bgalitsky@hotmail.com), Oracle Inc, USA

In this paper we learn how to manage a dialogue relying on discourse of its utterances. We consider two complementary approaches of dialogue management based on the discourse text analysis to extend the abilities of the interactive information retrieval-based chat bot.

A QUANTITATIVE METHOD FOR ANALYSIS OF CONNECTIVES: THE “PORTRAIT” OF THE RUSSIAN CONJUNCTION *ILI* (OR) IN THE SUPRACORPORA DATABASE OF CONNECTIVES

Inkova O. Yu. (Olga.Inkova@unige.ch), Institute of Informatics Problems, FRC CSC RAS,
Moscow, Russia; University of Geneva, Geneva, Switzerland

The functional properties of the conjunction *ili* ‘or’ are quite well studied and discussed in grammars and number of specific studies. However, they were not subjected to multivariate quantitative analysis. The paper proposes this kind of analysis, carried out according to six parameters: i) logical-semantic relation expressed by the conjunction, ii) syntactic structure of the text fragment introduced by it, iii) position of the conjunction in this text fragment, iv) order of the text fragments connected by it, v) status of *ili* in the context (cf. its use as a particle mentioned in MAS), vi) disposition of the elements that make up the multiword connectives (cf. *ili ... ili prosto* ‘or ... or just’).

The analysis of the formal variants of *ili*, carried out using the Supracorpora database of connectives, made it possible to formulate the conclusions that follow and to confirm them with quantitative data. i) Always occupying the initial position in the text fragment introduced by him, *ili* is used mainly as a connective. ii) The most typical order of text fragments for formal variants with *ili* is the p CNT q scheme. iii) By its syntactic characteristics *ili* is directly opposite to *i* ‘and’: *ili* is used in written texts in most cases for a non-predicative coordination. iv) Formal variants with *ili* express the relation of alternative at all three semantic levels (propositional, illocutive, metalinguistic), as well as the relations of substitution, correction, and negative alternative, but with a clear predominance of propositional alternative. v) *Ili* forms multiword and two- and multicomponent connectives; their composition varies depending on the relation expressed by them.

AN EXPERIMENTAL RULE-BASED PARSER FOR RUSSIAN EMPLOYING THE NLP RESOURCES OF THE ETAP SYSTEM

Inshakova E. S. (e.s.inshakova@gmail.com), Laboratory of Computational Linguistics,
A. A. Kharkevich Institute for Information Transmission Problems; Institute of Linguistics
RAS, Moscow, Russia, **Sizov V. G.** (victor.sizov@gmail.com), Laboratory of Computational
Linguistics, A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia

This paper presents a rule-based dependency parser for Russian based on bottom-up approach. Its rules are partially rewritten ETAP syntagms, organized into groups that constitute a single pipeline. We demonstrate that such an organization enhances the performance of our parser relative to the ETAP system’s and enables it to successfully process long phrases (more specifically, heavy nominal and prepositional phrases at the current experimental stage of our work).

VALENCY STRUCTURE OF CERTAIN PREDICATES OF SPEECH: NEW FINDINGS

Iomdin B. L. (iomdin@ruslang.ru), **Iomdin L. L.** (iomdin@iitp.ru), V. V. Vinogradov Institute of
Russian Language, RAS; A. A. Kharkevich Institute for Information Transmission Problems, RAS

The paper discusses valency frames of a number of Russian verbal predicates whose semantics includes speech acts and, at a certain step of semantic decomposition, the negation, like *vozražat* ‘object, retort’, *vozmuščat’sja* ‘resent, be indignant’ or *izvinjat’sja* ‘apologize’. It is hypothesized that the frames of such predicates include a pair of propositional valencies distinctly

opposed to each other: (1) the valency of stimulus that expresses the state of events and (2) the valency of response that introduces a speech act performed by the subject as a reaction to this state of event and offering an explanation. For example, in the sentence *Ivan izvinilsja, čto ne prišel na moj den' roždenija* 'Ivan apologized that he did not come to my birthday party' the clause starting with *čto* 'that' represents the state of events, whilst in the sentence *Ivan izvinilsja, čto ploxo sebja čuvstvoval* 'Ivan apologized that he was not feeling well' the *čto*-clause introduces Ivan's response to the stimulus (e. g. of not coming to the birthday party). It is shown that these valencies cannot be adequately described with a single semantic role of content. The authors also give a generalization of this phenomenon, comparing it to other instances of valency pairs, and suggest the existence of predicates having two valency centers.

RUREBUS-2020 SHARED TASK: RUSSIAN RELATION EXTRACTION FOR BUSINESS

Ivanin V. A. (vitalii.ivanin@abbyy.com)^{1,2}, **Artemova E. L.** (Echernyak@hse.ru)³, **Batura T. V.** (tatiana.v.batura@gmail.com)^{4,7}, **Ivanov V. V.** (nomemm@gmail.com)^{5,7}, **Sarkisyan V. V.** (vsarkisyan@hse.ru)³, **Tutubalina E. V.** (elvtutubalina@kpfu.ru)^{6,7}, **Smurov I. M.** (ivan.smurov@abbyy.com)^{1,2}; ¹ABBY, ²Moscow Institute of Physics and Technology, ³National Research University Higher School of Economics, ⁴Novosibirsk State University, ⁵Innopolis University, ⁶Kazan Federal University, ⁷Lomonosov Moscow State University

In this paper, we present a shared task on core information extraction problems, named entity recognition and relation extraction. In contrast to popular shared tasks on related problems, we try to move away from strictly academic rigor and rather model a business case. As a source for textual data we choose the corpus of Russian strategic documents, which we annotated according to our own annotation scheme. To speed up the annotation process, we exploit various active learning techniques. In total we ended up with more than two hundred annotated documents. Thus we managed to create a high-quality data set in short time. The shared task consisted of three tracks, devoted to 1) named entity recognition, 2) relation extraction and 3) joint named entity recognition and relation extraction. We provided with the annotated texts as well as a set of unannotated texts, which could of been used in any way to improve solutions. In the paper we overview and compare solutions, submitted by the shared task participants. We release both raw and annotated corpora along with annotation guidelines, evaluation scripts and results at <https://github.com/dialogue-evaluation/RuREBus>.

COMPARATIVE ANALYSIS OF RHETORICAL AND ARGUMENTATIVE STRUCTURES IN THE STUDY OF POPULAR SCIENCE DISCOURSE

Kononenko I. S. (irina_k@cn.ru), **Sidorova E. A.** (Isidorova@iis.nsk.su), **Akhmadeeva I. R.** (ah.irishka@gmail.com), A. P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia

The proposed work is performed as a part of an on-going research project aimed at creation of discourse annotated corpus of popular science texts written in Russian. Annotation is carried out within the framework of a multi-level model of discourse, which considers the text from the perspective of genre, rhetorical and argumentative organization. We conduct a comparative study of the rhetorical and argument annotations, discuss their similarities and differences on the segment and structural levels and show them on the examples of standard schemes of reasoning described in D. Walton's theory of structured argumentation: "Argument from Expert Opinion", "Argument from Example", and "Argument from Cause to Effect". Special attention is paid to discourse markers registered during annotation as key indicators of discourse structure. We report the results of the experiment with argument indicator patterns, based on the list of rhetorical markers, and aimed at the extraction of "from Expert Opinion" arguments.

EXPLORING THE BERT CROSS-LINGUAL TRANSFER FOR READING COMPREHENSION

Konovalov V. P. (vaskoncv@phystech.edu)[†], **Gulyaev P. A.** (guliaev.pa@phystech.edu)[†], **Sorokin A. A.** (alexey.sorokin@list.ru)^{†*}, **Kuratov Y. M.** (yurii.kuratov@phystech.edu)[†], **Burtsev M. S.** (burtcev.ms@mipt.ru)[†]; [†]Moscow Institute of Physics and Technology, Dolgoprudny, Russia, ^{*}Moscow State University, Moscow, Russia

Multilingual BERT has been shown to generalize well in a zero-shot cross-lingual setting. This generalization was measured on POS and NER tasks. We explore the multilingual BERT cross-language transferability on the reading comprehension task. We compare different modes of training of question-answering model for a non-English language using both English and language-specific data. We demonstrate that the model based on multilingual BERT is slightly behind the monolingual BERT-based on Russian data, however, it achieves comparable results with the language-specific variant on Chinese. We also show that training jointly on English data and additional 10,000 monolingual samples allows it to reach the performance comparable to the one trained on monolingual data only.

DISFLUENCIES IN RUSSIAN SPOKEN MONOLOGUES: A DISTRIBUTIONAL ANALYSIS

Korotaev N. A. (n_korotaev@hotmail.com), **Podlesskaya V. I.** (vi_podlesskaya@il-rngu.ru), Russian State University for the Humanities, Moscow, Russia, **Smirnova K. V.** (kategold1@gmail.com), **Fedorova O. V.** (olga.fedorova@msu.ru), Lomonosov Moscow State University, Moscow, Russia

The paper addresses the overall distribution of speech disfluencies in Russian spoken monologic discourse: basing on corpus data, we investigate qualitatively and quantitatively how disfluencies of different types group (or do not group) with each other and how isolated disfluencies and their sequences are sandwiched with periods of fluent speech in the course of speech production. Self-repairs, filled and silent pauses, and instances of hesitation lengthening were annotated in a subcorpus of the “Russian Pears Chats and Stories” (RUPEX). A distribution-oriented typology of disfluencies was proposed that distinguishes between isolated disfluencies, disfluency clusters, and quasi-clusters. We claim that disfluency tokens tend to cluster, as isolated occurrences are significantly less frequent in our data than it could have been expected basing on the relative frequency of tokens. This finding contradicts previous studies that treated disfluency clusters as a more marginal phenomenon, and emphasizes the importance of a distributional, rather than merely structural, approach to annotating disfluencies. Furthermore, individual types of disfluency tokens demonstrate significantly different distributional patterns. Compared to other types, self-repairs occur more often in isolation, while words with hesitation lengthening appear predominantly in clusters, and filled pauses most often group with silent pauses to form quasi-clusters.

R-BERT FOR RELATIONSHIP EXTRACTION ON RUSSIAN BUSINESS DOCUMENTS

Korzun V. A. (sdernal2@gmail.com), MIPT ABBYY Lab, Moscow, Russia

This paper provides results of participation in the Russian Relation Extraction for Business shared task (RuREBus) within DialogueEvaluation 2020. Our team took the first place among 5 other teams in Relation Extraction with Named Entities task. The experiments showed that the best model is based on R-BERT model. R-BERT achieved significant result in comparison with models based on Convolutional or Recurrent Neural Networks on the SemEval-2010 task 8 relational dataset. In order to adapt this model to RuREBus task we also added some modifications like negative sampling. In addition, we have tested other models for Relation Extraction and Named Entity Recognition tasks.

TAXONOMY ENRICHMENT FOR RUSSIAN: SYNSET CLASSIFICATION OUTPERFORMS LINEAR HYPONYM-HYPERNYM PROJECTIONS

Kunilovskaya M. (maria.kunilovskaya@wlv.ac.uk), **Kutuzov A.** (andreku@ifi.uio.no), **Plum A.** (a.j.plum@wlv.ac.uk), University of Wolverhampton, UK, University of Oslo, Norway, University of Wolverhampton, UK

We present the description of our system that was ranked third in the noun sub-track of the Taxonomy Enrichment for the Russian Language shared task offered by Dialogue Evaluation 2020. Our best-performing system appears against the backdrop of other methods and their combinations attempted, and its results argue in favour of Occam's razor for this task. A simple supervised classifier was trained on static distributional embeddings of hyponym words as features and their numeric hypernym synset identifiers from the taxonomy as class labels. It outperformed more complicated approaches based on learning linear projections from hyponym embeddings to hypernym embeddings and returning synset identifiers for the nearest neighbours of the predicted vectors. Training specially tailored word embeddings for ruWordNet multi-word expressions proved to be one of the key factors for both approaches.

SEMANTIC EFFECTS OF VERB TENSE IN PARENTHETICAL CONSTRUCTIONS WITH MENTAL VERBS

Kustova G. I. (galinak03@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

Parenthetical constructions with verbs of opinion (*as I think*) are considered as the result of the reduction of the main clause: *Ja dumaju, chto priglashenie prislal professor Wiler Priglashenie, kak ja dumaju, prislal professor Wiler*. The meaning of the mental verb tense affects the interpretation of the sentence. In the present tense, construction *as I think* introduces an assumption with a neutral status: *Eto proizojdet, kak ja dumaju, v samom blizhajshem budushchem* [Ju. Semenov]—'no one knows, P or non-P'. In the past tense, construction *as I thought* introduces a wrong assumption: *Dj-adja, kotoryj, kak ja dumal, davno zabyl o podarenykh chasakh, vosprinjal etu novost' boleznenno*.

SHIFTRY: WEB SERVICE FOR DIACHRONIC ANALYSIS OF RUSSIAN NEWS

Kutuzov A. (andreku@ifi.uio.no), University of Oslo, Oslo, Norway, **Fomin V.** (wadimiusz@gmail.com), National Research University Higher School of Economics, Moscow, Russia, **Mikhailov V.** (Mikhailov.V.Nikola@sberbank.ru), National Research University Higher School of Economics; Sberbank, Moscow, Russia, **Rodina J.** (julia.rodina97@gmail.com), National Research University Higher School of Economics, Moscow, Russia

We present the ShiftRy web service. It helps to analyze temporal changes in the usage of words in news texts from Russian mass media. For that, we employ diachronic word embedding models trained on large Russian news corpora from 2010 up to 2019. The users can explore the usage history of any given query word, or browse the lists of words ranked by the degree of their semantic drift in any couple of years. Visualizations of the words' trajectories through time are provided. Importantly, users can obtain corpus examples with the query word before and after the semantic shift (if any). The aim of ShiftRy is to ease the task of studying word history on short-term time spans, and the influence of social and political events on word usage. The service will be updated with new data yearly.

SENTENCE COMPRESSION FOR RUSSIAN: DATASET AND BASELINES

Kuvshinova T. (tatiana.kuvsh@yandex.ru), Higher School of Economics, National Research University, Moscow, Russia

Sentence compression is the task of removing redundant information from a sentence while preserving its original meaning. In this paper, we approach deletion-based sentence compression for the Russian language. We use the data from the plagiarism detection corpus (ParaPlag) to create a corpus for sentence compression in Russian of almost 3,000 pairs of sentences. We align source sentences and their compressions using the Needleman-Wunsch algorithm and perform human-evaluation of the corpus by readability and informativeness.

Then we use bidirectional LSTM to solve sentence-compression task for Russian, which is a typical baseline for the problem. We also experiment with RuBert and Bert-multilingual. For the latter, we use transfer-learning, firstly pretraining the model on English data, which improves performance. We conduct human evaluation by readability and informativeness and do error analysis for the models. We are able to achieve f-measure of 74.8%, readability of 3.88 and informativeness of 3.47 (out of 5) on test data. We also implement post-hoc syntax-based evaluator, which can detect some of the wrong compressions, increasing overall quality of the system.

We provide the data and baseline results for future studies.

“UNDERSTATEMENT” AND SARCASM: LEXICALIZATION OF A RHETORICAL DEVICE

Levontina I. B. (irina.levontina@mail.ru), Russian Language (Vinogradov) Institute RAS

Understatement is a rhetorical device, based on making a statement weaker than it could be made in a given situation (i. e. underrating, less confident, presented as unimportant). In modern Russian, especially in colloquial speech, an extremely popular rhetorical figure is a combination of understatement and sarcasm; recently, several new ways of forming this figure have appeared: *na minutochku, esli chto, nichego chto..?* [Eto na minutochku moya professiya; Eto, esli chto, moya professiya; A nichego, chto eto moya professiya?] ([literally *This is my profession, for a minute; This is my profession, just in case; Doesn't it mean anything that this is my profession?*]). For some language units, the corresponding meaning is partially or completely lexicalized. So, *na minutochku* and *na sekundochku* do not initially possess a “degrading” sense (if it is not really about time, meaning that you need a tiny bit of time for something); they are always used sarcastically. That said, as opposed to *na minutochku* and *na sekundochku*, other word forms (*na minutu, na minutku, na sekundu, na mig, na mgnovenie*) are not used this way. Thus, here we have a completely lexicalized figure of speech. In general, sarcasm is extremely difficult to formalize. Therefore, detection of linguistic manifestations of sarcasm appears to be extremely valuable.

SENTIMENT FRAMES FOR ATTITUDE EXTRACTION IN RUSSIAN

Loukachevitch N. V. (louk_nat@mail.ru), Lomonosov Moscow State University, Moscow, Russia, **Rusnachenko N. L.** (kolyarus@yandex.ru), Bauman Moscow State Technical University, Moscow, Russia

Texts can convey several types of inter-related information concerning opinions and attitudes. Such information includes the author's attitude towards mentioned entities, attitudes of the entities towards each other, positive and negative effects on the entities in the described situations. In this paper, we described the lexicon RuSentiFrames for Russian, where predicate words and expressions are collected and linked to so-called sentiment frames conveying several types of presupposed information on attitudes and effects. We applied the created frames in the task of extracting attitudes from a large news collection.

GRAMEVAL 2020 SHARED TASK: RUSSIAN FULL MORPHOLOGY AND UNIVERSAL DEPENDENCIES PARSING

Lyashevskaya O. N. (olesar@yandex.ru), National Research University Higher School of Economics; V. V. Vinogradov Russian Language Institute of RAS, Moscow, Russia, **Shavrina T. O.** (rybolos@gmail.com), National Research University Higher School of Economics; Sberbank, Moscow, Russia, **Trofimov I. V.** (itrofimov@gmail.com), **Vlasova N. A.** (nathalie.vlassova@gmail.com), A. K. Ailamazyan Program Systems Institute of RAS, Pereslavl-Zalessky, Russia

The paper presents the results of GramEval 2020, a shared task on Russian morphological and syntactic processing. The objective is to process Russian texts starting from provided tokens to parts of speech (pos), grammatical features, lemmas, and labeled dependency trees. To encourage the multi-domain processing, five genres of Modern Russian are selected as test data: news, social media and electronic communication, wiki-texts, fiction, poetry; Middle Russian texts are used as the sixth test set. The data annotation follows the Universal Dependencies scheme. Unlike in

many similar tasks, the collection of existing resources, the annotation of which is not perfectly harmonized, is provided for training, so the variability in annotations is a further source of difficulties. The main metric is the average accuracy of pos, features, and lemma tagging, and LAS.

In this report, the organizers of GramEval 2020 overview the task, training and test data, evaluation methodology, submission routine, and participating systems. The approaches proposed by the participating systems and their results are reported and analyzed.

SUMMARY CONSTRUCTION STRATEGIES FOR HEADLINE GENERATION IN THE RUSSIAN

Malykh V. (valentin.malykh@phystech.edu), Kazan Federal University, Kazan, Russia,
Cherniavskii D., Valukov A., Moscow Institute of Physics and Technology, Moscow, Russia

RUSSE2020: FINDINGS OF THE FIRST TAXONOMY ENRICHMENT TASK FOR THE RUSSIAN LANGUAGE

Nikishina I. (Irina.Nikishina@skoltech.ru), **Logacheva V.** (V.Logacheva@skoltech.ru),
Panchenko A. (A.Panchenko@skoltech.ru), Skolkovo Institute of Science and Technology (Skoltech), Moscow, Russia, **Loukachevitch N.** (louk_nat@mail.ru), Lomonosov Moscow State University, Moscow, Russia

This paper describes the results of the first shared task on taxonomy enrichment for the Russian language. The participants were asked to extend an existing taxonomy with previously unseen words: for each new word their systems should provide a ranked list of possible (candidate) hypernyms. In comparison to the previous tasks for other languages, our competition has a more realistic task setting: new words were provided without definitions. Instead, we provided a textual corpus where these new terms occurred. For this evaluation campaign, we developed a new evaluation dataset based on unpublished RuWordNet data. The shared task features two tracks: “nouns” and “verbs”. 16 teams participated in the task demonstrating high results with more than a half of them outperforming the provided baseline.

CO-CONSTRUCTED SYNTACTIC UNITS IN DIALOGUES WITH INDIVIDUALS WHO STUTTER

Olenikova A. V. (tradaladei@gmail.com), **Fedorova O. V.** (olga.fedorova@msu.ru),
Lomonosov Moscow State University, Moscow, Russia

Dialogue implies a high degree of coordination between the interlocutors, which makes possible the existence of co-constructed turns used by speakers for various purposes. One of the reasons for them to appear is difficulties in articulation experienced by one of the participants and prompting the other participant interested in achieving the communicative goal to increase their own contribution to the dialogue. In conversations with people who stutter, co-constructions are more common than in conversations between people who have no diagnosed speech disorders; among them completions prevail, because one of the interlocutors more often spells out uncompleted constructions. The study of stuttering from a linguistic perspective is of considerable interest, since it provides an opportunity to study dialogue as a process including cooperation between participants. During this collaborative process one interlocutor’s contribution affects the contribution of another and can trigger non-standard turn-taking techniques.

DOC2VEC OR BETTER INTERPRETABILITY? A METHOD STUDY FOR AUTHORSHIP ATTRIBUTION

Pimonova E. (hpimonova@gmail.com), **Durandin O.** (oleg.durandin@gmail.com),
Malafeev A. (amalafeev@yandex.ru), National Research University Higher School of Economics, Nizhny Novgorod, Russia

In this work, we perform a method study for the problem of authorship attribution in Russian and English. The datasets used consist of 324 works written in Russian and 207 works in English. We propose a set of text representation models that reflect various linguistic phenomena,

in particular, morphological and syntactic ones. One distinctive feature of the proposed models is that they are interpretable. These models are used individually and in combination against a Doc2Vec baseline. For Russian, some of our models outperform Doc2Vec, but this does not happen in the case of English, for various reasons. However, the proposed models can also be used together with Doc2Vec, dramatically improving its performance: by 16.79% in the case of Russian and by 7.2% for English. Additionally, we experiment with two different methods for separating texts into blocks of K sentences (contiguous and bootstrapped) and performed parameter tuning of K . Finally, we conduct a feature importance analysis and show which linguistic markers of author style are the most pertinent for Russian, English and for both these languages. All code used in this work is made freely available to the community.

RUSSIAN LANGUAGE AND CORPUS DIVERSITY

Piperski A. Ch. (apiperski@gmail.com), Russian State University for the Humanities / National Research University Higher School of Economics, Moscow, Russia

This paper discusses the use of most widely-known Russian corpora, namely Russian National Corpus, ruTenTen, General Internet Corpus of Russian, and Araneum Russicum Maximum, for the theoretical study of Russian language. Based on a sample of papers from 2019, I demonstrate that scholars, especially theoretical linguists, tend to ignore the opportunities provided by a wide range of Web corpora, even though these resources are well-known to the NLP community. I present a selection of case studies to show that data from “non-classical” corpora can be used for studying various linguistic phenomena, such as: 1) variation in morphology and syntax; 2) word formation and lexical change; 3) construction grammar. I also claim that the underuse of non-classical corpora is partly due to the fact that they are (perceived as) not quite user-friendly.

“A TOT PEROVSKOJ NE DAL VSLAST POSPAT”: PROSODY AND GRAMMAR OF ANAPHORIC TOT THROUGH THE LENS OF CORPUS DATA

Podlesskaya V. I. (vi_podlesskaya@il-rgggu.ru), Russian State University for the Humanities, Moscow, Russia

Based on data from the Russian National Corpus and the General Internet-Corpus of Russian, the paper addresses syntactic, semantic and prosodic features of constructions with the demonstrative TOT used as an anaphor. These constructions have gained some attention in earlier studies [Paducheva 2016], [Berger, Weiss 1987], [Kibrik 2011], [Podlesskaya 2001], but their analysis (a) covered primarily their prototypical uses; and (b) was based on written data. The data from informal, esp. from spoken discourse show however that the actual use of these constructions may deviate considerably from the known prototype. The paper aims at bridging this gap. I claim (i) that the function of TOT is to temporarily promote a referent from a less privileged discourse status to a more privileged one; and (ii) that TOT can be analyzed on a par with switch reference devices in the languages where the latter are grammatically marked (e.g. on verb forms). The following parameters of TOT-constructions are discussed: syntactic and semantic roles of TOT and of its antecedent in their respective clauses, linear and structural distances between TOT and its antecedent, animacy of the maintained referent. Special attention is paid to the information structure of the TOT construction: I give structural and prosodic evidence that TOT never has a rhematic status. The revealed actual distribution of TOT (a) adds to our understanding of cross-linguistic variation of anaphoric functions of demonstratives; and, hopefully, (b) may contribute to further developing computational approaches to coreference and anaphora resolution for Russian, e.g. by improving datasets necessary for this task.

RUSSIAN NATURAL LANGUAGE GENERATION: CREATION OF A LANGUAGE MODELING DATASET AND EVALUATION WITH MODERN NEURAL ARCHITECTURES

Shaheen Z. (shaheen@itmo.ru)^{1,2}, **Wohlgemant G.** (gwohlg@itmo.ru)¹, **Zaity B.** (bassel.zaity@gmail.com)², **Mouromtsev D. I.** (mouromtsev@itmo.ru)¹, **Pak V. G.** (vadim.pak@cit.icc.spbstu.ru)²;

¹Faculty of Software Engineering and Computer Systems, ITMO University;

²Institute of Computer Science and Technology Peter the Great St. Petersburg Polytechnic University (SPbPU), St. Petersburg, Russia

Generating coherent, grammatically correct, and meaningful text is very challenging, however, it is crucial to many modern NLP systems. So far, research has mostly focused on English language, for other languages both standardized datasets, as well as experiments with state-of-the-art models, are rare. In this work, we i) provide a novel reference dataset for Russian language modeling, ii) experiment with popular modern methods for text generation, namely variational autoencoders, and generative adversarial networks, which we trained on the new dataset. We evaluate the generated text regarding metrics such as perplexity, grammatical correctness and lexical diversity.

LANGUAGE-SPECIFIC WORDS IN THE LIGHT OF TRANSLATION: THE RUSSIAN *TOSKA*

Shmelev A. D. (shmelev.alexei@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences; Moscow Pedagogical State University, Moscow, Russia

This paper presents a semantic analysis of the most language-specific Russian word for ‘sadness’, namely, *toska*. The analysis is based on the hypothesis that one may regard translation equivalents and paraphrases of a linguistic unit extracted from real translated texts as a source of information about its semantics. The appearance of language-specific words in translated texts may be even more useful for studying their semantics.

It turns out that *toska* is not all that rare in Russian translated texts. The study of the incentives that lead Russian translators to use the word *toska* often reveals important aspects of the semantics of this word. Stimuli for the appearance of *toska* in translations into Russian vary greatly. In general, when the original describes some bad feelings, the word *toska* appears if the original speaks of a subject’s unsatisfied desire, which desire may be vague and not well understood and it usually cannot be satisfied. In addition, the subject often feels lonely.

TAGGING AND PARSING OF MULTIDOMAIN COLLECTIONS

Sorokin A. A. (alexey.sorokin@list.ru), Moscow Institute of Physics and Technology, Neural Networks and Deep Learning Lab, Dolgoprudny, Russia,; Moscow State University, Faculty of Mathematics and Mechanics, Moscow, Russia, **Smurov I. M.** (ivan.smurov@abby.com), ABBYY, Moscow, Russia; , Moscow Institute of Physics and Technology, Computer Vision and Natural Language Processing Laboratory (ABBYY Lab), Dolgoprudny, Russia, **Kiryanov D. P.** (denkirjanov@gmail.com), SberDevices, Moscow, Russia

In this paper we describe our submission to GramEval2020 competition on morphological tagging, lemmatization and dependency parsing.

Our model uses biaffine attention over the BERT representations. The main feature of our work is the extensive usage of language model, tagger and parser fine-tuning on several distinct genres and the implementation of genre classifier. To deal with dataset idiosyncrasies we also extensively apply handwritten rules.

Our model took second place in the overall model performance scoring 90.8 aggregate measure over all 4 tasks.

VISUAL VS. AUDITORY PERCEPTION OF BULGARIAN STIMULI BY RUSSIAN NATIVE SPEAKERS

Stenger I. (ira.stenger@mx.uni-saarland.de),

Avgustinova T. (avgustinova@coli.uni-saarland.de), Saarland University, Saarbrücken, Germany

This study contributes to a better understanding of receptive multilingualism by determining similarities and differences in successful processing of written and spoken cognate words in an unknown but (closely) related language. We investigate two Slavic languages with regard to their mutual intelligibility. The current focus is on the recognition of isolated Bulgarian words by Russian native speakers in a cognate guessing task, considering both written and audio stimuli. The experimentally obtained intercomprehension scores show a generally high degree of intelligibility of Bulgarian cognates to Russian subjects, as well as processing difficulties in case of visual vs. auditory perception. In search of an explanation, we examine the linguistic factors that can contribute to various degrees of written and spoken word intelligibility. The intercomprehension scores obtained in the online word translation experiments are correlated with (i) the identical and mismatched correspondences on the orthographic and phonetic level, (ii) the word length of the stimuli, and (iii) the frequency of Russian cognates. Additionally we validate two measuring methods: the Levenshtein distance and the word adaptation surprisal as potential predictors of the word intelligibility in reading and oral intercomprehension.

AN EMPIRICAL INVESTIGATION OF LANGUAGE MODEL BASED REVERSE TURING TEST AS A TOOL FOR KNOWLEDGE AND SKILLS ASSESSMENT

Tarasov D. (dtarasov3@gmail.com), **Matveeva T.**, **Galiullina N.**, Meanotek, Kazan, Russia

Automating assessment of person's skills is an important area of study in artificial intelligence and natural language processing. In this work we conduct empirical study of a recently proposed Reverse Turing Test for Knowledge Assessment approach—a completely automated domain agnostic method of knowledge assessment that can operate completely without human assessor involvement. Our study involved 53 participants and three different knowledge domains. We conclude that this method can reliably differentiate between expertise levels and therefore can be a compelling alternative to human grading and multiple-choice tests in many domains.

SEMANTICS FOR *OB RATNO*: A RE-ENTRY INTO A DISCONTINUED STATE

Tatevosov S. G. (tatevosov@gmail.com), Lomonosov Moscow State University, Moscow,

Russia, **Kisseleva X. L.** (xkisseleva@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

This paper explores the meaning and distribution of *obratno*, one of the Russian repetitive and restitutive morphemes. We identify three essential characteristics of *obratno*: obligatoriness of the restitutive reading, narrow scope with respect of indefinites, and incompatibility with eventuality descriptions that entail a result state in the sense of [Kratzer 2000]. We argue that like garden-variety repetitive and restitutive morphemes (e.g., Russian *opjat'*), *obratno* denotes a partial identity function with a presupposition. Unlike such morphemes, however, the presuppositional content of *obratno* involves a return to the same state in which an entity had been before. We capture this characteristic relying on [Landman's 2008] notion of cross-temporal identity of eventualities and the derivative notion of a cross-temporal substate. This makes the repetitive reading of *obratno* unavailable, forces identity of the holders of a state, deriving the narrow scope effect, and guarantees that *obratno* is only compatible with target state descriptions.

PRETRAINING AND AUGMENTATION IN NAMED ENTITY RECOGNITION TASK FOR CYBERSECURITY DOMAIN IN RUSSIAN

Tikhomirov M. M. (tikhomirov.mm@gmail.com), **Loukachevitch N. V.** (louk_nat@mail.ru),

Sirotna A. Yu. (overnastuhed@yandex.ru), **Dobrov B. V.** (dobrov_bv@mail.ru), Lomonosov Moscow State University, Moscow, Russia

The paper presents the results of applying the BERT representation model in the named entity recognition task for the cybersecurity domain in Russian. Several variants of the model were investigated. The best results were obtained using the BERT model, trained on the target collection of information

security texts. This model achieved results, which were 15 percentage points of F1-macro measure greater than results of CRF, the best method in previous experiments for the same task and data. We also explored a new form of data augmentation for the task of named entity recognition.

COMBINED APPROACH TO HYPERNYM DETECTION FOR THESAURUS ENRICHMENT

Tikhomirov M. M. (tikhomirov.mm@gmail.com), Lomonosov Moscow State University, Moscow, Russia, **Loukachevitch N. V.** (louk_nat@mail.ru), Lomonosov Moscow State University, Moscow, Russia, Kazan Federal University, Kazan, Russia, **Parkhomenko E. A.** (parkat13@yandex.ru), Lomonosov Moscow State University, Moscow, Russia

This paper describes a combined approach to hypernym detection task. The approach combines the following techniques: distribution semantics, rule-based patterns, and modern neural networks (BERT). An important feature of our solution is that hypernyms are extracted only from a single text collection provided by the organizers. The described approach obtained the fourth result on the private nouns track. It was found out that the use of the rule-based patterns can significantly improve the results. Also, using the BERT model as an additional factor always helps to improve the performance.

DISCOURSE FEATURES OF BLOGS IN SUBCORPUS OF RUSSIAN RU-RSTREEBANK

Toldova S. (toldova@yandex.ru), NRU Higher School of Economics, Moscow, Russia; **Davydova T.** (tdadidik@gmail.com), **Kobozeva M.** (kobozeva@isa.ru), **Pisarevskaya D.** (dinabr@gmail.com), FRC CSC RAS, Moscow, Russia

The paper presents a corpus study of the discourse features in the corpus of blogs. It is based on the data of Ru-RSTreebank annotated within the framework of the Rhetorical Structure theory [Mann, Thompson 1988]. The Ru-RSTreebank represents genres of news and popular science, scientific papers, and blogs texts. Blog subcorpus contains such topics as travelling, cosmetics, sports and health, psychology, IT and tech and some others. Blogs texts constitute a specific genre as they combine properties of written and spoken discourse. The purpose of the paper is to investigate discourse features of blogs in comparison with other genres. We analyze the variation in rhetoric relations distribution among genres, and single out the differences in discourse connectives usage. Furthermore, we check the distribution of other discourse features reported in different studies for spoken discourse and for social media in the Ru-RSTreebank blogs subcorpus. The general frequency analysis and the experiments on RandomForest classifier application to genre recognition have shown that the most important rhetoric relations specific to blogs are Evaluation and Contrast, that there is a tendency to use shorter discourse units and not to express the discourse relations overtly via subordinative conjunctions.

DISTRIBUTIONAL MODELS AND AUXILIARY METHODS FOR DETERMINING THE HYPERNYMS OF WORDS IN RUSSIAN

Yadrintsev V. V. (vvyadrincev@gmail.com), Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences; Peoples Friendship University of Russia (RUDN University), Moscow, Russia, **Ryzhova A. A.** (Anastasiia.Ryzhova@skoltech.ru), Skolkovo Institute of Science and Technology, Moscow, Russia, **Sochenkov I. V.** (sochenkov@isa.ru), Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, Moscow, Russia

This paper describes our participation in the first shared task on Automatic Taxonomy Construction for the Russian language RUSSE’2020. The goal of this task is the following: input words (neologisms that are not yet included in the taxonomy) need to be associated with the appropriate hypernyms from an existing taxonomy. For example, for the input word “duck”, it is expected that participants will provide a list of its ten hypernyms-synsets to which the word can most likely be attributed, such as “animal,” “bird” and so on. An input word can refer to one, two, or more “parents” at the same time.

In this article we are trying to answer the following question: what results can be achieved using only “raw” vectors from distributional models without additional training? The article presents the results for several pre-trained models that are based on fastText, Elmo, and BERT algorithms. Also, an out-of-vocabulary analysis was performed for the models under consideration. Taking into account all public scores from the leaderboards, we showed the results corresponding to the following places in the ranking: the 3rd place on public nouns, the 2nd on private nouns, the 4th on public verbs, and the 4th on private verbs.

RUSSIAN ADVERB *DAVNO* ‘LONG AGO, FOR A LONG TIME’ REVISITED FROM A CORPUS PERSPECTIVE

Yanko T. E. (tanya_yanko@list.ru), Institute of Linguistics; Pushkin State Russian Language Institute, Moscow, Russia

During the last twenty years, the Russian adverb *davno* ‘long ago, for a long time’ was widely discussed in literature. It was recognized that the unique parameter of *davno* is its inability to be the theme of a sentence. Moreover, if *davno* functions in the context of aspectual forms relating to the past it can only be the rheme. In the context of the aspectual verbal forms relating to the past but preserving the connection with the moment of speech, *davno* can be either the rheme proper, or a component of the rheme. A classic example of an aspectual verb form referring to the past is the general factual meaning of the imperfective aspect. At present, the spoken data corpora can shed light on the communicative structure analysis, since the prosodic structure of the sound speech provides a straightforward access to the communicative structure. Novel parameters of *davno* are as follows. 1) Whereas *davno* is traditionally recognized as a word of rhematic polarity it can nevertheless function as a component of the theme in the context of attributive clauses and constructions (*Davno soglasovannyj visit dolzhen byl sostojatsja v aprele* ‘A visit planned long ago would take place in April’). 2) The general factual meaning of the imperfective aspect, contrary to what was assumed before, is not an absolute prerequisite for *davno* to function as the rheme. The spoken corpus showed that in the context of negation and in the context of the verbs of speech, the general factual allows for *davno* to function as a component of the rheme but not the rheme proper (*Ja davno tebjja ne videl* ‘I have not been seeing you for a long time’; *My davno govorili, chto nasha zadacha — eto borjba s terrorismom* ‘We have been insisting for a long time that our main goal is the struggle against terrorism’). 3) A specific type of questions with the initial *davno* (as well as with other adverbs with the meaning of a considerable quantity like *chasto* ‘often’, *mnogo* ‘much’, and *daleko* ‘far away’) is singled out. Such questions cannot be unambiguously classified either as *yes-no*-questions or as *wh*-questions (*I davno vy zdesj stoite?* ‘And how long are you staying here?’). A description of unique prosody of such questions is given. 4) In the context of discourse continuity, *davno* acquires the rising prosody which is in fact uncharacteristic of a word, which is unable to be the theme (*Xotel eto sdelat’ davno, no teperj sdelaju točno* ‘I wished to do it long ago, but now I will do it for sure’). The rising tone is accounted for by the meaning of continuity, which has the same prosody as the theme. 5) In constructions *kogda-to davno* ‘once upon a time’, *ochenj davno* ‘very long ago’, *davno-davno* ‘very long ago’, *davnym-davno* ‘very long ago’, *dovoljno davno* ‘quite long ago’, *ne tak davno* ‘not so long ago’ *davno* loses its rhematic polarity.

The parameters of *davno* are exemplified by spoken fragments taken from the Multimodal corpus of the Russian National corpus, and the minor working collection of the Russian speech recordings specifically set up for this investigation. The software program Praat was used in the process of analyzing the sound data.

THE RUSSIAN *KAK BY*: SEMANTICS, PRAGMATICS, AND DIACHRONY

Zalizniak **Anna A.** (anna.zalizniak@gmail.com), Institute of Linguistics of the RAS; Institute of Informatics Problems of the FRC CSC RAS, Moscow, Russia

The article considers the semantics of the Russian word *kak by*. It demonstrates that there are three main types of use of this word that are relevant for the modern Russian language: 1) as an approximation indicator, i.e. the marker of an approximative, indirect or metaphorical use of the linguistic unit it introduces (cf. *léd na reke sluzhil kak by mostom* ‘ice on the river served as a kind of bridge’; *on kak by veduschij specialist v dannoj oblasti* ‘he is sort of leading specialist in this field’); 2) as an indicator of epistemic indefiniteness (cf. *infljatsii kak by net* ‘there is

<kak by> no inflation’); 3) as an illocutionary operator (“illocutionary mitigator”), mitigating the illocutionary force of the assertive speech act (cf. *Ja kak by ispolnitel’nyj direktor kompanii* ‘I am <kak by> the chief executive officer of the company’, uttered by the actual CEO of the company). We suggest that the initial meaning of *kak by* is that of a marker of descriptive indefiniteness (in an outdated use after the verbs of fuzzy perception), which has served as a source for both the approximation meaning, which is the main function of this word in contemporary Russian and that of epistemic indefiniteness. In its function as an “illocutionary mitigator” that emerged at the very end of the 20th century in the course of pragmaticalisation, the word *kak by* belongs to the class of discourse markers that ensure the success of a communicative act.

The study was based on the Russian National Corpus (www.ruscorpora.ru), including its oral and parallel subcorpora.

ZERO FORMS IN MORPHOLOGICAL PARADIGMS: THE VERB “BE” IN RUSSIAN

Zimmerling A. V. (fagraey64@hotmail.com), Pushkin state Russian language institute; Institute of Linguistics, Russian Academy of science; Moscow pedagogical state university, Moscow, Russia

This paper offers a corpus analysis of the Russian verb *быть* ‘be’ which has an abnormal present tense paradigm including a zero form $\emptyset^{\text{BE.PRES}}$ and overt forms *есть*^{BE.PRES} and *суть*^{BE.PRES} which do not discriminate person and number and are distributed syntactically. I discuss different approaches to the grammar of *быть* and argue that Apresjan’s model which recognizes $\emptyset^{\text{BE.PRES}}$, *есть*^{BE.PRES} and *суть*^{BE.PRES} as parts of one and the same lemma is superior to alternative models splitting *быть* into two lemmas representing copula vs content verb ‘be’. The peripheral status of overt present BE-forms compared with $\emptyset^{\text{BE.PRES}}$ in the Russian National Corpus is confirmed by three measures: 1) dispersion of texts where a BE-form occurs; 2) uneven coverage in different persons and numbers; 3) ratio of copular uses vs content verb uses. 1–2 person present tense BE-forms attested in RNC are internal borrowings from Old Russian and Old Church Slavonic, while *есть*^{BE.PRES} and *суть*^{BE.PRES} are inherited 3rd person elements which take over 1–2 person uses. The historical 3Pl *суть* is redundant in a system, where a more frequent 3rd person form *есть* is licensed in the plural: it survives by a minority of speakers either as an optional 3Pl copula in formal discourse or as an emphatic copula in oral discourse. The form *есть*^{BE.PRES} occurs in all persons and numbers both as content verb and as copula but is underrepresented as 3Pl copula: this gap is filled by $\emptyset^{\text{BE.PRES}}$. The frequency of the zero copula $\emptyset^{\text{BE.PRES}}$ can be measured in corpora without syntactic annotation on the basis of systemic proportion between present vs past tense uses of *быть* and on the basis of approximation samples for contexts where overt copulas alternate with $\emptyset^{\text{BE.PRES}}$.

THE PERCEPTION OF ROBOTS EMOTIONAL GESTURES AND SPEECH BY CHILDREN SOLVING A SPATIAL PUZZLE

Zinina A. A. (zinina_aa@nrcki.ru), **Zaidelman L. Y.** (zaydelman_ly@nrcki.ru), **Kotov A. A.** (kotov_aa@nrcki.ru), **Arinkin N. A.** (arinkin_na@nrcki.ru), Kurhcatov Institute; Russian State University for the Humanities, Moscow, Russia

The emotional behavior of a companion robot is important for human-robot interaction in the situation of training tasks. We examined the influence of emotional gestures and emotional speech of the robot on its perception by primary school students (N=52, male, female, mean age 9.8) in the situation of joint solution of the spatial Tangram puzzle. It was shown that emotional gestures make a significant contribution to the attractiveness of the robot for the child. It was also found that test subjects prefer the robot with emotional gestures and speech over the robot with neutral gesture and speech behavior. The study also analyzed the communicative behavior of children, identified typical communicative signs that are typical for interaction start with the robot, for monitoring the game and for difficult situations. We described typical mistakes that children make when assembling a puzzle together with the robot.

Авторский указатель

Августинова Т.	684	Иншакова Е. С.	387
Акжигитова Г. Р.	348	Инькова О. Ю.	310, 372
Алексеева С. В.	147	Иомдин Б. Л.	400
Анастасьев Д. Г.	1	Иомдин Л. Л.	400
Арефьев Н. В.	14, 188	Кабанов А. В.	14
Аринкин Н. А.	811	Кирьянов Д. П.	671
Артемова Е. Л.	417	Киселева К. Л.	708
Архипов М. Ю.	270	Коновалов В. П.	445
Ахмадеева И. Р.	432	Кононенко И. С.	432
Баранов А. Н.	48	Корзун В. А.	467
Батура Т. В.	417	Кортаев Н. А.	455
Беликов В.	62	Котов А. А.	811
Блеканов И. С.	77	Кувшинова Т.	517
Блинова О. В.	77	Куратов Ю. М.	445
Большина А. С.	119	Кустова Г. И.	485
Бочаров В. В.	93	Кутузов А.	501
Бурцев М. С.	445	Левонтина И. Б.	529
Валюков А.	570	Логачева В.	579
Власова Н. А.	554	Лукашевич Н. В.
Воронцов К. В.	256, 284	119, 541, 579, 724, 736
Галиуллина Н.	696	Ляшевская О. Н.	241, 554
Галицкий Б. А.	361	Малафеев А.	606
Геймбух Г. А.	348	Малых В.	570
Гончаров А. А.	310	Матвеева Т.	696
Горбова Е. В.	321	Мироненко Ф. Д.	300
Гордеев Д. И.	188, 348	Михайлов В.	501
Гуляев П. А.	445	Модина В. В.	77
Давлетов А. А.	188, 348	Мордашова Д. Д.	241
Дале Д. С.	177	Никишина И.	579
Деткова Ю.	211	Новицкий В.	211
Добров Б. В.	724	Оленикова А. В.	596
Добровольский Д. О.	48	Панченко А.	579
Дурандин О.	606	Пархоменко Е. А.	736
Дьячков В. В.	229	Петрова М.	211
Евсеев Д. А.	270	Пимонова Е.	606
Еремеев М. А.	256	Пиперски А. Ч.	615
Жукова В. А.	241	Плешак П. С.	229
Зайдельман Л. Ю.	811	Подлеская В. И.	455, 628
Зализняк Анна А.	784	Рахилина Е. В.	241
Зизов В. С.	14	Рей А. И.	188, 348
Зинина А. А.	811	Родина Ю.	501
Иванин В. А.	417	Русначенко Н. Л.	541
Иванов В. В.	417	Рыжова А. А.	763
Ильвовский Д. А.	361	Садекова Т. Р.	284

Саркисян В. В.	417	Тутубалина Е. В.	417
Селегей В.	62, 211	Федорова О. В.	455, 596
Селегей Д.	62	Федосеев М. В.	14
Сидорова Е. А.	432	Фельдман Д. Г.	284
Сизов В. Г.	387	Феногорова А. С.	300
Сиротина А. Ю.	724	Филипецкая Д. В.	300
Слюсарь Н. А.	147	Фомин В.	501
Смирнова К. В.	455	Хомченкова И. А.	229
Смуrow И. М.	417, 671	Циммерлинг А. В.	796
Сорокин А. А.	445, 671	Чернова Д. А.	147
Соченков И. В.	763	Чернявский Д.	570
Стойнова Н. М.	229	Чуйкова О. Ю.	160
Табишева А. О.	300	Шаврина Т. О.	554
Тарасов Д.	696	Шаландар Г, де	93
Тарасов Н. А.	77	Шмелев А. Д.	658
Татевосов С. Г.	708	Штенгер И.	684
Тихомиров М. М.	724, 736	Эндресен А. А.	241
Тихонова М. И.	300	Ядринцев В. В.	763
Трофимов И. В.	554	Янко Т. Е.	773

Author Index

Akhmadeeva I. R.	432	Fomin V.	500
Akzhigitova G. R.	348	Frolova T. I.	106
Alexeeva S. V.	147	Galitsky B. A.	361
Anastasyev D. G.	1	Galiullina N.	696
Arefyev N. V.	13, 187	Geymbukh G. A.	348
Arinkin N. A.	811	Goncharov A. A.	310
Arkhipov M. Yu.	270	Gorbova E. V.	321
Artemova E. L.	416	Gordeev D. I.	187, 348
Avgustinova T.	684	Gulyaev P. A.	445
Badryzlova Yu. G.	33	Ilvovsky D. A.	361
Bakhanova M.	198	Inkova O. Yu.	310, 372
Baranov A. N.	49	Inshakova E. S.	387
Batura T. V.	416	Iomdin B. L.	401
Belikov V.	63	Iomdin L. L.	106, 401
Blekanov I. S.	76	Ivanin V. A.	416
Blinova O. V.	76	Ivanov V. V.	416
Bocharov V. V.	93	Kabanov A. V.	13
Boguslavsky I. M.	106	Khomchenkova I. A.	228
Bolshina A. S.	119	Kirianov D. P.	670
Budennaya E. V.	133	Kisseleva X. L.	708
Burtsev M. S.	445	Kobozeva M.	747
Chalendar G., de	93	Kononenko I. S.	432
Cherniavskii D.	570	Konovalov V. P.	445
Chernova D. A.	147	Korotaev N. A.	454
Chuikova O. Iu.	160	Korzun V. A.	467
Dale D. S.	177	Kotov A. A.	811
Davletov A. A.	187, 348	Kunilovskaya M.	474
Davydova T.	747	Kuratov Y. M.	445
Derbanosov R.	198	Kustova G. I.	485
Detkova J.	211	Kutuzov A.	474, 500
Dikonov V. G.	106	Kuvshinova T.	517
Dobrov B. V.	724	Lazursky A. V.	106
Dobrovol'skij D. O.	49	Levontina I. B.	529
Durandin O.	606	Logacheva V.	579
Dyachkov V. V.	228	Loukachevitch N. V.
Endresen A. A.	241	119, 541, 579, 724, 736
Eremeev M. A.	256	Lyashevskaya O. N.	241, 553
Evdokimova A. A.	133	Malafeev A.	606
Evseev D. A.	270	Malykh V.	570
Fedorova O. V.	454, 596	Matveeva T.	696
Fedoseev M. V.	13	Mikhailov V.	500
Feldman D. G.	283	Mironenko F. D.	299
Fenogenova A. S.	299	Modina V. V.	76
Filipetskaya D. V.	299	Mouromtsev D. I.	644

Mordashova D. D.	241	Slioussar N. A.	147
Nikishina I.	579	Smirnova K. V.	454
Nikolaeva Ju. V.	133	Smurov I. M.	416, 670
Novitskiy V.	211	Sochenkov I. V.	762
Olenikova A. V.	596	Sorokin A. A.	445, 670
Pak V. G.	644	Stenger I.	684
Panchenko A.	579	Stoynova N. M.	228
Parkhomenko E. A.	736	Sukhova N. V.	133
Petrova M.	211	Tabisheva A. O.	299
Pimonova E.	606	Tarasov D.	696
Piperski A. Ch.	615	Tarasov N. A.	76
Pisarevskaya D.	747	Tatevosov S. G.	708
Pleshak P. S.	228	Tikhomirov M. M.	724, 736
Plum A.	474	Tikhonova M. I.	299
Podlesskaya V. I.	454, 628	Timoshenko S. P.	106
Rakhilina E. V.	241	Toldova S.	747
Rey A. I.	187, 348	Trofimov I. V.	553
Rodina J.	500	Tutubalina E. V.	416
Rusnachenko N. L.	541	Valukov A.	570
Rygaev I. P.	106	Vlasova N. A.	553
Ryzhova A. A.	762	Vorontsov K. V.	256, 283
Sadekova T. R.	283	Wohlgenannt G.	644
Sarkisyan V. V.	416	Yadrintsev V. V.	762
Selegey D.	63	Yanko T. E.	773
Selegey V.	63, 211	Zaidelman L. Y.	811
Shaheen Z.	644	Zaity B.	644
Shavrina T. O.	553	Zalizniak Anna A.	784
Shmelev A. D.	658	Zhukova V. A.	241
Sidorova E. A.	432	Zimmerling A. V.	795
Siroтина A. Yu.	724	Zinina A. A.	811
Sizov V. G.	387		

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 19 (26). 2020

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06