ONLINE: ISSN 2075-7182

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2021)

Выпуск 20

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2021)

Issue 20

УДК 81.322(063)
ББК 81.1я431
К63

К63 **Компьютерная лингвистика** и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 20. Москва: РГГУ, 2021. С. I–764.

ISBN 978-5-7281-3032-1

Сборник включает 71 доклад международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2021», 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.

УДК 81.322(063)
ББК 81.1я431

ISBN 978-5-7281-3031-4
ISBN 978-5-7281-3032-1 (осн. том)

Предисловие

20-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 27-й международной конференции «Диалог». В 2021 году для публикации в ежегоднике редколлегией был отобран 71 доклад из 149, поданных на конференцию. Работы, представленные в сборнике, отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на конференции:

- компьютерные лингвистические ресурсы;
- компьютерный анализ документов (классификация, перевод, поиск, саммаризация, генерация, анализ тональности и аргументации и т. д.);
- глубокое обучение в компьютерной лингвистике (методики применения, содержательная лингвистическая интерпретация);
- компьютерный анализ Social Media;
- корпусная лингвистика и корпусометрия (методики создания, использования и оценки корпусов);
- компьютерная семантика (от аналитических до дистрибуционных моделей);
- лингвистические онтологии и автоматическое извлечение знаний;
- мультимодальные подходы к анализу языка (на стыке NLP и Computer Vision);
- мультимодальная лингвистика (включая лингвистический анализ речи);
- модели общения и диалоговые агенты;
- лингвистический анализ текста (морфология, синтаксис, семантика);
- компьютерная лексикография.

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» является подведение итогов технологических соревнований между разработчиками систем лингвистического анализа русскоязычных текстов — Dialogue Evaluation. В этом году состоялись 6 соревнований:

- RuSimpleSentEval (RSSE): задача упрощения тестов (text simplification);
- RuShiftEval: детектирование диахронических семантических сдвигов в русском языке;
- Low Resource ASR: распознавание речи для малоресурсных языков;
- Summarization: кластеризация, выбор и генерации заголовков для новостей;
- SemSketches: построение семантических скетчей;
- RuNormAS (Russian Normalization of Annotated Spans): приведение именованных сущностей к нормальной форме.

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 Редсовет отказался от печати сборника на бумаге. Все сборники размещаются на сайте конференции. С 2014 года данный сборник индексируются Scopus.

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

Организаторы

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

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

Международный программный комитет

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

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY, Россия
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Браславский Павел Исаакович	Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ им. М. В. Ломоносова
Захаров Леонид Михайлович	МГУ им. М. В. Ломоносова
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	Компания Yandex
Ляшевская Ольга Николаевна	Институт русского языка им. В. В. Виноградова РАН
Пиперски Александр Чедович	РГГУ
Толдова Светлана Юрьевна	НИУ «Высшая школа экономики»
Федорова Ольга Викторовна	МГУ им. М. В. Ломоносова
Шаров Сергей Александрович	Университет Лидса

Секретариат

Родионова Ольга Игоревна, *координатор оргкомитета*	Компания ABBYY, Россия
Ульянова Анна Вячеславовна, *секретарь оргкомитета*	РГГУ, Россия

Рецензенты

Августинова Таня
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Антонова Александра Александровна
Арефьев Николай Викторович
Архангельский Тимофей Александрович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Бенко Владимир
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Браславский Павел Исаакович
Бурцев Михаил Сергеевич
Васильев Виталий Геннадьевич
Гельбух Александр Феликсович
Гершман Анатолий
Губин Максим Вадимович
Гусев Илья Олегович
Диконов Вячеслав Григорьевич
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрушина Нина Роландовна
Жуковский Александр Евгеньевич
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Иванов Владимир Владимирович
Иомдин Леонид Лейбович
Ильвовский Дмитрий Алексеевич
Катинская Анисья Юрьевна
Кибрик Андрей Александрович
Клышинский Эдуард Станиславович
Клячко Елена Леонидовна
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Куратов Юрий Михайлович
Кутузов Андрей Борисович
Лапошина Антонина Николаевна
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Маккарти Диана
Малафеев Алексей Юрьевич
Малых Валентин Андреевич
Митрофанова Ольга Александровна
Наков Преслав
Никишина Ирина Юрьевна
Новицкий Валерий Игоревич
Переверзева Светлана Игоревна
Петрова Мария Владимировна
Пивоварова Лидия Михайловна
Пиперски Александр Чедович
Плунгян Владимир Александрович
Подлесская Вера Исааковна
Пономарева Мария Алексеевна
Рыгаев Иван Петрович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Смирнов Иван Валентинович
Смуров Иван Михайлович
Соловьев Валерий Дмитриевич
Татевосов Сергей Георгиевич
Толдова Светлана Юрьевна
Тутубалина Елена Викторовна
Усталов Дмитрий Алексеевич
Федорова Ольга Викторовна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шаров Сергей Александрович
Шелманов Артём Олегович
Юдина Мария Владимировна
Янко Татьяна Евгеньевна

Contents[1]

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

Matching semantic sketches to predicates in context using the BERT model

Aleksandrova Polina
NRU HSE
Moscow, Russia
paleksandrova37@gmail.com

Mokhova Anna
NRU HSE
Moscow, Russia
ann.mokhova@gmail.com

Nikolaenkova Maria
NRU HSE
Moscow, Russia
hublbublgog@gmail.com

Abstract

Modern language models have extensive information about the compatibility and meanings of various words. One of the ways to represent such lexical information, which is presented in the present study, is the construction of semantic sketches.

This paper presents a solution to the task of predicting a predicate from its most frequent actants and sirconstants using the application of the BERT neural network, which showed the best quality metrics in the Dialogue Evaluation SemSketches competition. The study analyzed several solutions approaching this task and ways to improve them based on the peculiarities of the architecture and the nature of data in terms of linguistics.

The results of testing the selected methods showed that the most successful tool for determining the semantic sketch of a predicate is the Conversational RuBERT model combined with the search for synonyms of the verbs sought in the training data.

Other promising ways to improve the quality of mapping the predicate to its semantic sketch include the use of contextualized embeddings to be able to take context into account, as well as fine-tuning of the models used.

Keywords: semantic sketches, lexical compatibility, language modeling.
DOI: 10.28995/2075-7182-2021-20-1-7

Соотнесение скетча с предикатом в контексте с использованием модели BERT

Александрова Полина
НИУ ВШЭ
Москва, Россия
paleksandrova37@gmail.com

Мохова Анна
НИУ ВШЭ
Москва, Россия
ann.mokhova@gmail.com

Николаенкова Мария
НИУ ВШЭ
Москва, Россия
hublbublgog@gmail.com

Аннотация

Современные языковые модели обладают широкой информацией о сочетаемости и значениях различных слов. Один из способов представлений таких лексических сведений, который представлен в настоящем исследовании, — конструирование семантических скетчей. В данной работе представлено решение задачи предсказания предиката по его наиболее частотным актантам и сирконстантам с помощью применения нейронной сети BERT, которое показало наилучшие метрики качества в рамках соревнования Dialogue Evaluation SemSketches. В ходе исследования было проанализировано несколько подходов, приближающих к решению этой задачи, а также способы их улучшения, основанные на особенностях архитектуры и природы данных с точки зрения лингвистики. Результаты тестирования выбранных методов показали, что наиболее успешным ин-

струментом для определения семантического скетча предиката является модель Conversational RuBERT в сочетании с поиском синонимов искомых глаголов в тренировочных данных. К другим перспективным способам улучшения качества сопоставления предиката с его семантическим скетчем можно отнести использование контекстуализированных эмбеддингов для возможности учитывать контекст, а также дообучение (fine-tuning) используемых моделей. гвистики.

Ключевые слова: семантические скетчи, лексическая сочетаемость, языковое моделирование.

1 Introduction

The concept of semantic sketch, which will be used in this paper, can be defined as follows: a semantic sketch is a representation of all the actants and sirconstants of a predicate, which are distributed into classes according to their semantic roles. Another definition, more often used when working with corpus methods, is the idea of representation "in the form of statistics of combinability of the analyzed word with syntactically related lexical units [2].

An example of a semantic sketch of the verb ‘играть’ is shown in Table 1:

	Sphere Special	Time	Agent	Locative	ContrAgent
0	в карты	в детстве	дети	на бирже	с детьми
1	в шахматы	на большой перемене	мальчишки	во дворе	с мальчишками
2	в футбол	по вечерам	пацаны	на бильярде	с читателем
3	в азартные игры	каждый день	игроки внизу	на компьютере	с собакой
4	в игры	допоздна	ребята	на площадке	с сыном
5	в прятки	в молодости	команда	на чужом поле	с ребятами

Таблица 1: Example of semantic scetch for verb ’играть’

The separation of semantic word sketches is widely applicable for lexical analysis of linguistic units and corpus representation of linguistic data. This approach was first proposed and implemented by Adam Kilgarriff within the SketchEngine project [8].

Currently, various methods of applying semantic sketches as a particular type of corpus rendition are becoming a subject of research in computational linguistics. This paper will address the problem of correlating a semantic sketch with a predicate in context, which has been put forward in the SemSketches competition.

2 Task Description

The task of semantic sketches prediction was introduced in the Dialogue Evaluation SemSketches competition. Baseline solution is presented in the work [7]. The contexts of use (sentences) for the most frequent Russian predicates, as well as a set of anonymized semantic sketches, were chosen as the initial data. Several such predicates are: выйти (go out), сидеть (sit),действовать (act), подумать (think), написать (write), воспринимать (perceive), завершать (complete), принять (accept), встречаться (meet), продать (sell), говорить (talk). Anonymized sketches are such sketches, for each of which information about the essential roles and their fillers is provided, but the predicate itself is hidden. The task comes down to matching each context with one of the anonymized sketches. In other words, a set of contexts for different predicates is given, and the selected predicate in the context must be mapped to a sketch. And several different contexts can correspond to the same sketch. The data for the training sample consisted of 2000 sentences and 20 sketches. The results of the models were tested on a benchmark sample of 44750 sentences and 895 sketches, from which, in turn, 4347 sentences and 100 sketches, dev.gold and manual dev.gold, were manually selected, respectively. The Bidirectional Encoder Representations from Transformers (BERT) model [1] was used to solve the problems of this study. BERT is a neural network from Google, created in 2018, which showed by a large margin state-of-the-art results

on a number of tasks. BERT can be used to create artificial intelligence programs to solve problems from various fields, in particular for natural language processing. RuBERT is a BERT model trained on the Russian-language part of Wikipedia and news data. Methods for adapting multilingual masked language models are presented in [5]. The use of this model has significantly improved the handling of Russian language data, and it will be used in some of the approaches described in the paper.

2.1 Baseline

This section will briefly present the solution of the organizers of the SemSketches competition, the results of which were used as a baseline. Determining the correspondences between a predicate and its semantic sketch was divided into several subtasks. First, syntactic parsing of sentences was applied to existing contexts, searching for the predicate and masking its direct dependents. In the next step, masked words were predicted using the RuBERT model. The resulting predicted sketches were compared with the reference variants, and the final accuracy quality metric was 0.1535.

3 Methods

3.1 Sketches

The solution [1] to the problem at hand is based on the idea of predicting a predicate for each sketch by generating templates from its data. The concept of masked language modeling, implemented in the BERT model used, is essential for template creation. During the neural network training, individual tokens were randomly masked in the input data, and the main task was to predict the token in place of the mask. This training procedure has a clear advantage in the described task over those models that learn to predict each next word based on the previous context because there is a possibility to predict a specific and any possible position in the sequence.

So, as a result of processing the sketch, the output is templates of two types:

1. MASK + role
   For each of the fillings of each role of the sketch the masked context was mapped to both left and right. The need to generate both [MASK] + role and role + [MASK] templates simultaneously is due to the fact that BERT, configured to process whole sentences, is very likely to predict punctuation marks in place of the mask in the right-hand position. In the final analysis, the predicted fillers on such patterns were treated separately.
   There is also a problem with predictions for templates with agentive roles and masks since in this case there are almost no verbs among the results: the model tends to construct sentences with nominal predicates. Therefore, it was decided to consider agentive roles separately.
2. Agent + [MASK] + role
   Each of the filler sets for roles was divided into two groups: fillers for agentic roles (these include 'Agent', 'Agent Metaphoric', 'Agent Route') and the rest. In the absence of agentic roles, we limited ourselves to the pronouns 'he' and 'they'.
   Thus, each of the role fillers other than agentic roles was matched with templates of the form Agent + [MASK] + role, where Agent is all the possible agentic fillers described above. The number of templates is not fixed - it depends on how many of the most frequent role fills are present in a particular sketch.

An example of several templates for the verb to 'собираться':
[MASK] в стаи
в стаи [MASK]

[1] The code is available on https://github.com/psaleksandrova/Matching-semantic-sketches-to-predicates-in-context-using-the-BERT-model

[MASK] по выходным
регулярно [MASK]
друзья [MASK] в последний раз
друзья [MASK] у кого-нибудь дома
публика [MASK] в 9 часов
военный совет [MASK] впервые
вся семья [MASK] в дорогу

3.2 RuBERT

After the templates were generated for the sketch, for each of them, a list of placeholders was predicted using RuBERT, which was also used in the organizers' solution.

After predicting the fillers in place of the mask, only verbs were selected from the resulting lists. The results obtained, as well as the predicates from the sentences, were lemmatized to find matches. Morphological analysis and lemmatization were implemented using the pymorphy2 library [4].

On the one hand, this is a necessary step to find identical verbs in predicates and sentences; on the other hand, an inaccurate definition of the initial form could disrupt the matching process (the initial form шило for the past tense verb masculine шить (sting)). Thus, each sketch was matched with a list of predicates predicted for each of the patterns. Next, the resulting list was ranked by frequency of predictions, and the most frequent predicate was selected as the final single predicate for the sketch, which was then matched to the sentence with the corresponding predicate. If there were several such sentences, the very first of them was chosen without any analysis, which entailed an unresolved polysemy, since the choice of the sketch sentence was determined by the verb alone and not by its context.

An example of a cross-section of the frequency-ranked list of predictions of the verb sketch ‘собираться’ in the Table 2.

быть	0.42058823529411765
прийти	0.21764705882352942
собрать	0.18529411764705883
ходить	0.17352941176470588
собираться	0.1470588235294117
ждать	0.1323529411764706
собраться	0.0941176470588235
играть	0.07941176470588235
войти	0.06764705882352941
ехать	0.06764705882352941

Таблица 2: Example of the frequency-ranked list of predictions of the verb sketch ‘собирать’

Accordingly, if there is a context for the predicate ’быть’ in the set of 2000 sentences, then that is what will be predicted for the given sketch.

3.3 Conversational RuBERT

As an improvement to the method proposed above, it was decided to use Conversational RuBERT. It is assumed that the model, which was trained on the texts of subtitles, blogs, social networks, etc. should better process various stable word combinations and, in general, better summarize the features of Internet data, which were just used to highlight the semantic sketches proposed in the competition. It is possible that the Wikipedia and news texts on which the standard RuBERT was trained do not in principle contain, or contain in small amounts insufficient for the model to "remember"some of the roles from the sketches or predicates in mind.

This approach also solves the problem of cases where the predicate predicted for the sketch was not found in any of the sentences. In the first iteration, the randomized selection was used in such cases as the answer, but the new method solved the problem by searching for verbs synonymous with the predicted predicate. A cosine distance comparison of vector word2vec representations was used to search by synonyms [3]. A static model from the RusVectores resource [6], trained on the texts of the National Corpus of the Russian language (NRU)[2], was used. Synonyms are those words of a similar part of speech whose embeddings are the least different (cosine distance is minimal) from the vector of the original verb.

4 Analysis

The work resulted in a best-effort error analysis, namely the reference predicate - predicted predicate pairs for each sketch were examined. In addition to explicit differences between verbs, which are challenging to explain linguistically, several explicit groups were identified for well-interpreted mismatches, which, with proper correction, can result in a correct prediction.

Among these discrepancies, there are cases where the verbs in the reference data and the prediction differ only in the feature of species, the presence of the category of return. Quite a few pairs with the same root morpheme and similar meaning but different prefixes. And also exciting cases are pairs of synonyms, both of which obviously fit the roles of a particular sketch, as well as antonyms. In fact, some roles may contain information about the sign of meaning itself (for example, the fact of winning in the example lose-win) but do not reflect its presence or absence.

In the table below, for the highlighted groups of inconsistencies, examples are given from the results of predictions compared to the reference predicate.

Characteristics of a mismatch	Example (benchmark - prediction
aspectual pairs	выходить - выйти останавливаться - остановиться вздохнуть - вздыхать бросить - бросать
reflexivity	менять-меняться катить - катиться завершить - окончиться схватить - схватиться
prefixal verbs	пожать-нажать просидеть-сидеть подействовать - действовать исполнить - выполнить
synonyms	вложить - вкладывать расстреливать-убивать отметить - обнаружить вскакивать - встать падать - упасть
antonyms	проиграть - выиграть прощаться - встретиться

Таблица 3: Example of semantic scetch for verb 'играть'

5 Discussion

As an improvement of the two main methods, some ideas were proposed which theoretically should have resulted in a quality gain, but unfortunately their implementation was not possible.

[2]https://ruscorpora.ru/new/

As noted earlier, when matching the predicted predicate for a sketch with the corresponding sentences, the context was not taken into account in any way, which gave rise to false matches given the polysemy of the predicate. As a solution to this problem, the idea of using embeddings that use information about context begs to be solved. Contextualized embeddings of each token or the whole sentence can be extracted from the BERT model: they contain information about the entire sequence. Usually, the latent states from the last layers of the neural network are used as embeddings. Accordingly, it is possible to present as a contextualized vector both verbs, potentially suitable for the mask places in the templates, and predicates in all sentences. And select the most appropriate predicate for the sketch by ranking by cosine closeness between vector representations. The reason for the inability to implement the method is the limited memory in the Google Colab service, which allowed to predict only 500 predicates out of 44750 available.

Fine-tuning of the model on specific data gives good results on a number of NLP problems. The volume of available sentences should be sufficient for the necessary tuning of weights. The assumption is that if we mask in contexts only the particular predicate in question, and BERT predicts it in the process of retraining, then in pattern prediction, more attention will be paid to verbs and from the required finite set. It would then be possible to implement one of the previously proposed methods on the pre-trained BERT model. The limited amount of computational resources is the factor that prevented the implementation of this concept.

6 Conclusion

As a result of the present work we have studied the nature of semantic sketches and possible approaches to predicate prediction based on its possible actant and syrconstant fillers. Table 4 presents the results of the evaluation experiments with both methods that were implemented in our study.

Baseline	0.154
RuBERT	0.212
Conversational RuBERT	0.309

Таблица 4: Accuracy scores

We have carried out appropriate experiments on predicate prediction based on its semantic sketch and analyzed the results in terms of both the approach itself and the semantic nature of the data.

In the future, we plan to improve the implemented approach with an adjustment based on the semantic analysis of the current results; and, provided the computational resource problem is solved, to try to implement other methods we have proposed.

References

[1] Bert: Pre-training of deep bidirectional transformers for language understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.

[2] Detkova J. Novitskiy V. Petrova M. Selegey V. Differential semantic skethes for Russian Internet-corpora // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2020. — Vol. 17. — P. 20.

[3] Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // arXiv preprint arXiv:1310.4546. — 2013.

[4] Korobov Mikhail. Morphological analyzer and generator for Russian and Ukrainian languages // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2015. — P. 320–332.

[5] Kuratov Yuri, Arkhipov Mikhail. Adaptation of deep bidirectional multilingual transformers for russian language // arXiv preprint arXiv:1905.07213. — 2019.

[6] Kutuzov Andrey, Kuzmenko Elizaveta. WebVectors: a toolkit for building web interfaces for vector semantic models // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2016. — P. 155–161.

[7] SemSketches-2021: experimenting with the machine processing of the pilot semantic sketches corpus / Maria Ponomareva, Maria Petrova, Julia Detkova et al. // Proc Dialogue, Russian International Conference on Computational Linguistics. — Moscow, 2021.

[8] The Sketch Engine: ten years on / Adam Kilgarriff, Vít Baisa, Jan Bušta et al. // Lexicography. — 2014. — Vol. 1, no. 1. — P. 7–36.

Annotated Span Normalization as a Sequence Labelling Task

Daniil Anastasyev
Yandex
Moscow, Russia
dan-anastasev@yandex-team.ru

Abstract

In this paper, we describe a way to perform span normalization as a sequence labelling task. Our model predicts the modifications that should be applied to the span tokens to normalize them. This prediction is performed via sequence labelling, which means that each token is normalized independently. Despite the simplicity of the approach, we show that it can lead to the state-of-the-art results. We compare different pretraining schemas in application to this task. We show that the best quality can be achieved when the normalizer is trained on top of a BERT-based morpho-syntactic parser's representations. Moreover, we propose some additional features useful in the task and prove that auxiliary morpho-syntactic losses can help the model. Furthermore, we show that the model compares favourably with other contestant models of the RuNormAS competition.

Keywords: span normalization, lemmatization, pretrained language models, multitask learning
DOI: 10.28995/2075-7182-2021-20-8-15

Нормализация аннотированного спана как задача разметки последовательности

Анастасьев Д. Г.
Яндекс
Москва, Россия
dan-anastasev@yandex-team.ru

Аннотация

В этой статье мы описываем способ осуществления нормализации аннотированного спана как задачу разметки последовательности. Наша модель предсказывает модификацию, которую нужно произвести над токенами спана, чтобы получить его нормализованный вариант. Это предсказание осуществляется для каждого токена независимо. Данный подход является простым, однако способен давать высокое качество. Мы сравниваем различные схемы предобучения в контексте данной задачи и показываем, что наилучшие результаты могут быть достигнуты с применением модели предобученного морфо-синтаксического парсера на основе BERT. Мы предлагаем признаки, которые полезны в данной задаче, в дополнение к представлениям, получаемым из BERT. Итоговая модель показала одни из самых высоких результатов в рамках соревнования RuNormAS.

Ключевые слова: Нормализация аннотированного спана, лемматизация, предобученные языковые модели, многозадачное обучение.

1 Introduction

Annotated span normalization is the task of converting an annotated span of text to its normalized form. Such normalization is highly context-dependent. For instance, span "Александра Пушкина" will be normalized differently in the following contexts: "Александра Пушкина видели…"[1] (to "Александр Пушкин") and "Александра Пушкина видела…"[2] (to "Александра Пушкина").

[1] "Alexander Pushkin was seen..."
[2] "Alexandra Pushkina has seen..."

In a way, it is similar to lemmatization because you are required to perform disambiguation of a word to obtain its normal form. But in contrast to it, span normalization has to be syntactically correct: the syntactic dependencies between the span's words should be retained after the normalization.

The most common approach to perform it is based on a morpho-syntactic parser (e.g., it is achieved this way in Natasha library[3]). Each word of a span is converted to the correct form based on the syntactic relationships between the word and other words in the sentence. It is a suitable way to perform a span normalization without any specific train data available. However, it may be hard to obtain a normalization that would follow some external guidelines because you have to learn them first and implement them using some (maybe quite extensive) set of rules.

An alternative approach may be designed using language models or seq2seq framework as in [5]. In this case, a model is trained to predict the normalized sequence given the source sequence. It is quite straightforward to formulate the task this way and language models are known to lead to the state-of-the-art results in tasks such as machine translation. Another reason to utilize this approach is the fact that the model learns the normalization guidelines itself without additional supervision from your part.

In our approach, we consider something in between those two ways mentioned above. We perform the normalization as a sequence labelling task predicting a modification for each word in a span. The model is based on the word representations obtained from a morpho-syntactic parser. To this extent, our approach is similar to the first one. Nevertheless, we use only word embedding from the parser, not its final predictions. We assumed that such representations should contain all useful information that does not have to be converted to the parses. Our approach should also be faster than the one based on a language model because it predicts the normalization in a non-autoregressive manner.

2 Model

The model we used in our approach closely follows the model from [1] which has shown the state-of-the-art performance in morpho-syntactic parsing on the GramEval-2020 dataset [6].

The model consists of three main parts: a BERT-based embedder [2], an LSTM-based encoder and a simple classifier that predicts the modification which should be applied to the given word to obtain its normalized form. As a result, the model predicts the modifications for each word in each span in the sentence independently, which means that we can normalize all sentence spans in a single forward pass. Conversely, it may also lead to some inconsistencies between the normalized words.

Such a model would have been a simplified version of the model from [1], however, we added a few improvements to address the differences between the lemmatization and annotated span normalization tasks.

Figure 1 gives an overview of our model.

2.1 Span Features

The main difference between those tasks is the fact that the normalized form of a word depends not only on the disambiguated word itself but also on all the other words in the span. Let's consider the following sentence: "Правительство Российской Федерации рассмотрит…"[4]. The normalized form of the word "Российской"[5] depends on the span: it should be left unchanged when the first three words form the span, it is equal to "Российская" in the normalization of "Российской Федерации"[6] and it is normalized to "Российский" in a single word span.

To address such a difference, we used special features to mark those words that have to be normalized. The idea is based on BIO-notation in NER. We marked with the B-feature the first word of each span, while all the rest words of the span were marked with the I-feature and other words (not corresponding to any span) were marked with the O-feature. We trained three embeddings corresponding to those BIO features and concatenated them to the word embedding obtained from BERT.

[3] `https://natasha.github.io/`

[4] "The Government of the Russian Federation will consider…"

[5] "Russian"

[6] "The Russian Federation"

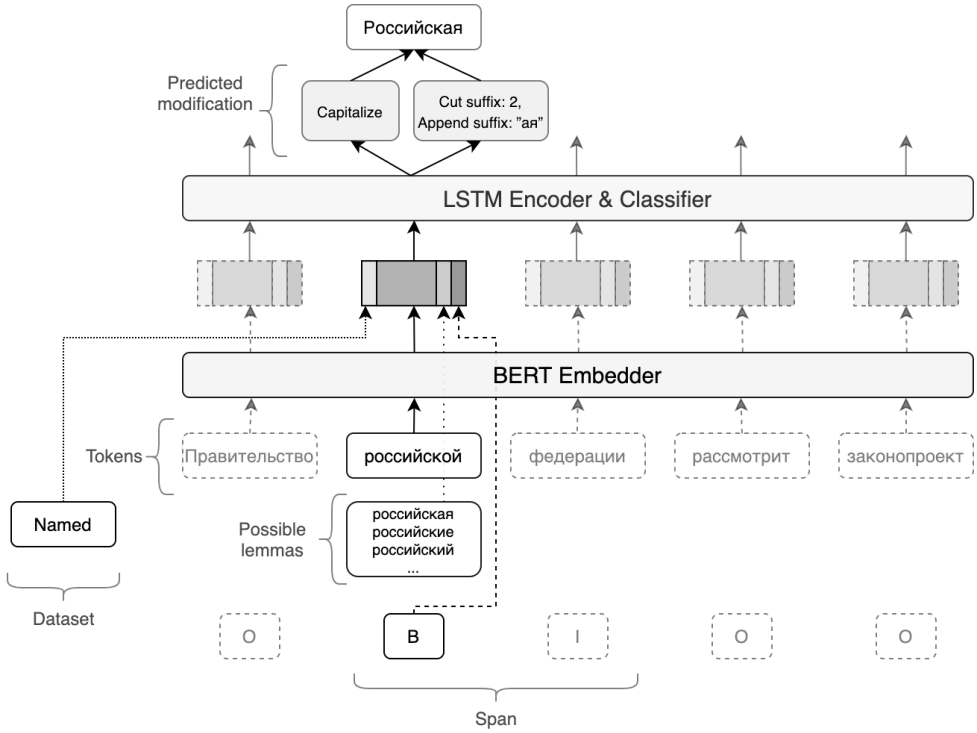


Figure 1: Schematic representation of our model

2.2 Prediction Factorization

We predicted the word normalization with a two-head classifier. One head was aimed to predict a set of modifications which should be applied to a word to obtain its normalized form, e.g.:

{”cut_prefix”: 0, cut_suffix”: 1, ”append_suffix”: ”ый”}

Another head predicted the word capitalization: lower- or uppercased, or capitalized. Such a process differs from the one in [1] in the way that the sets of spelling and capitalization modifications are now independent. We call it prediction factorization. Such factorization leads to a richer set of modifications because each spelling modification can be used together with each capitalization modification. It also should be easier to be learned by the model because of a significantly lower number of classes (approximately by the factor of three).

2.3 Lemma Features

We also conditioned the prediction on the set of spelling modifications allowed by a dictionary. We collected all the possible forms of each word and built a bag-of-spelling-modifications binary vector where 1s corresponded to indices of those modifications that can be obtained from the dictionary and 0s corresponded to all other modifications. This way, the model learned to prefer the allowed forms when it is possible but wasn’t constrained by them.

2.4 Auxiliary Morpho-Syntactic Loss

As we mentioned before, the span normalization can be performed based on a set of rules on top of morpho-syntactic parsing. We assumed that the model could benefit from an additional loss which will help it to learn morpho-syntactic relations better. For that purpose, we used a pretrained morpho-syntactic parser to predict the grammar values and syntactic relations in the train data. We trained our normalizer model to predict those (possibly dirty) values. We show that such auxiliary loss leads to some improvement of the normalization quality.

2.5 Pretrained Models

Pretrained models' usage is known to lead to great improvement in the model quality (e.g., [2]). We considered the RuBERT model and RuBERT fine-tuned for morpho-syntactic parsing in our work. The latter model should be more aware of morpho-syntactic relationships, so we hypothesized that it may suit our task better.

We also considered two modes of training. In the first variant, we trained only the LSTM-based encoder and classifier, while in the second one, we also fine-tuned the RuBERT's parameters.

However, it should be noted that the first approach on top of a morpho-syntactic parser is the best one in terms of its applicability: such a model may be used for normalization as well as parsing with the slowest part of computations (BERT representations obtaining) performed only once.

2.6 Ensembling

We also considered an ensemble of few models trained with a similar set of parameters. To select the prediction from such an ensemble, we used a majority vote with additional filtering. Such filtering worked in the following way: whenever a known word existed among the predicted normalization, we removed all unknown options from the voting. After the filtering, we selected the most popular predicted normalization.

3 Data

We trained our model on the data from the RuNormAS competition. We used random 90% of the data for training and the rest for validation purposes.

The data consisted of two subsets: generic spans and named entities. In the former subset, the model was required to normalize an arbitrary span, while in the latter, it should have to normalize only named entities. Consequently, generic spans generally were longer (in the number of tokens) and the named spans contained much more infrequent words (which would have been hard to find in a general-purpose dictionary).

	Generic	Named
Span count	54360	39575
Token count	136357	70997
Avg span length	2.5	1.8
Unknown token count	1376	5776
Unknown token rate	1%	8%

Table 1: Dataset statistics

Table 1 provides the train dataset details. We called *unknown* those tokens that were not available in the pymorphy2 dictionary (from [4]).

The main issue was to convert the data to the sequence-labelling-friendly format. To achieve it, we aligned the original texts with the corresponding normalized sequences and removed those which could not be aligned correctly.

The chief source of misalignments was a different number of tokens in the source and target sequences. Usually, it was a result of an additional spelling correction performed during the normalization of the training data. E.g., in some cases, the original text contained two words "анти террористической"[7] that should have been normalized to a single word "антитеррористический". Unfortunately, our model was not designed to fix such misspellings, so we did not have any option but to filter such sequences out of the training dataset. It led to the loss of about 2% of the training data.

We trained every model on all available data and did not try any specialized model for the generic or named subsets. However, we noted a slightly different format of normalization between the generic and named subtasks. E.g., a simple postprocessing rule that selects an infinitive form instead of a participle or

[7] "antiterrorist"

gerund improved the quality on the generic subtask but decreased it on the named one. To give the model an ability to learn such rules itself, we tried to concatenate a dataset embedding to each word embedding.

4 Experiments

4.1 Experimental Setup

As it was mentioned before, we highly reused the model from [1] in our setup. The whole experimental setup was also based on the provided code[8]. We utilized allennlp [3] and transformers [7] libraries for our implementation. The parameters of the BERT-based embedder and LSTM-encoder were the same. We used 8-dimensional span and dataset embedding features. We used the trainable_bert checkpoint to initialize the morpho-syntactic parser.

The code used for those experiments was published in a separate branch[9].

4.2 Embedders' Comparison

Table 2 summarizes the results achieved by different combinations of embedders, their finetuning modes and the applied auxiliary losses. *Syntax-BERT* models are the models initialized from a morpho-syntactic parser's checkpoint. *Multitask* models are the models trained with the auxiliary losses, while single corresponds to the models trained only to normalize words. *Frozen* or *trainable* relates to the BERT parameters' finetuning.

We also implemented a baseline based on the syntax parser predictions (*Syntax Baseline*). We normalized span tokens using the Natasha library implementation mentioned above, but we utilized syntax predictions from the same morpho-syntactic parser as in the *Syntax-BERT* models. Natasha predicts lowercased normalization so we used the source capitalization for the normalized tokens to achieve better quality. There still is some room for improvement with a capitalization predictor: we could have achieved 91.27% quality on the *Named* subset, if capitalization had not been considered.

	Syntax Baseline	RuBERT				Syntax-BERT			
		Frozen		Trainable		Frozen		Trainable	
		Single	Multitask	Single	Multitask	Single	Multitask	Single	Multitask
Generic	94,52%	96,81%	96,98%	96,87%	97,43%	97,30%	**97,74%**	97,10%	97,72%
Named	90,23%	96,20%	96,05%	97,47%	97,64%	97,74%	97,58%	97,41%	**97,77%**

Table 2: Embedders' comparison

Our baseline model shows the worst results which proves that good normalization is hard to obtain without additional handwritten rules or special model training.

Clearly, the model benefits from morpho-syntactic pretraining. It is especially noticeable in the case of frozen models comparison. As we mentioned before, we considered the frozen Syntax-BERT model to be the most useful in downstream applications because it can perform lots of tasks at the same time. However, the results show that such a model is also good from the normalization quality point of view.

Surprisingly, in the single-task training mode, the frozen Syntax-BERT is better than the trainable model. We can assume that the model slightly overfits when all parameters are trained. However, this is not the case in the multitask training. The additional losses prevent the model from forgetting useful morpho-syntactic relations learned by the parser.

RuBERT-based models work significantly better in the trainable setup. It may be explained by the fact that BERT's representations should be adjusted for the task (BERT-based morpho-syntactic parser also performed poorer without proper finetuning). However, a multitask trainable model is almost on par with the Syntax-BERT models. It suggests that the model also managed to learn some useful morpho-syntactic information from the automatic markup.

To assess this hypothesis, we compared the qualities of the multitask models on the GramEval-2020 dataset. The results are presented in Table 3.

[8] https://github.com/DanAnastasyev/GramEval2020

[9] https://github.com/DanAnastasyev/GramEval2020/tree/span_normalization

Model	POS	Feats	LAS
Frozen Syntax-BERT	**96,20%**	**96,40%**	**84,60%**
Trainable Syntax-BERT	95,55%	96,09%	83,70%
Trainable RuBERT	94,76%	95,81%	81,56%

Table 3: Comparison of multitask models on the GramEval-2020 test data

It is expected that the frozen model initialized from the parser trained on the GramEval dataset should achieve the highest scores. However, the results of other models are also promising. As anticipated, the trainable variant of the same model doesn't forget much because of the auxiliary losses. Moreover, the trainable RuBERT-based model learns morpho-syntactic parsing reasonably well without been exposed to a significantly larger and cleaner original dataset.

4.3 Model Features' Ablation

Table 4 shows how different features of the proposed model affected the quality. The Final model is based on Syntax-BERT fine-tuned in the multitask mode. Each subsequent row shows how the model would have performed without one feature.

	Generic	Named
Final model	97,72%	97,77%
- span embedding	84,70%	95,68%
- lemmatization factorization	97,48%	97,54%
- dataset embedding	97,28%	97,46%
- possible lemma embedding	97,10%	97,48%

Table 4: Model features' ablation

Just as expected, the span embeddings have a key role in our model. The model is unable to understand the span boundaries and performs simple lemmatization without them. It especially affects the model's quality on the generic subset where the spans tend to be longer.

Lack of any other feature also reduces the model quality, although much less substantially.

4.4 Comparison on RuNormAS

Finally, we can compare the model quality with other contestants of the RuNormAS competition. Our team was called *qbic* and we ended up in second place in the generic track and first place in the named entity normalization track. The full results are presented in Table 5.

Team	Generic	Named
ksmith	**98,01%**	98,12%
qbic	97,91%	**98,15%**
eindenbom	97,58%	97,92%
king_menin	96,45%	95,75%
baseline	77,32%	88,81%
fateev.da	77,30%	88,97%
shkunkov.a	0,00%	76,80%

Table 5: Comparison with other contestants

We used ensembling of five models to achieve better quality. The ensemble outperformed a single model by 0,2% on the generic subset and by 0,4% on the named subset.

5 Error Analysis

We performed some error analysis of the final ensemble. We identified the following types of errors:

- *Space* error is an error that could have been fixed by correct spacing, e.g., ”В.Филев” was predicted instead of the expected ”В. Филев”. Our algorithm used the original spaces for the normalization, so it was not able to correct such cases. This type of error was ignored in the generic track.
- *Punctuation* error occurred because of the mismatch of the source and normalized texts punctuation, e.g., ”дети - сироты” (with a hyphen) instead of ”дети – сироты” (with en-dash).
- *Capitalization* error is a result of incorrectly predicted capitalization. It is common in some multi-word expressions with ambiguity, e.g., in the normalization ”институт исследований Южной Азии” predicted instead of ”Институт исследований Южной Азии” the word ”институт” may be capitalized or not depending on the context. This type of error was also ignored in the generic track.
- *Alignment* error may be a result of the incorrect word count (e.g., ”будет увеличиваться” instead of ”увеличиваться”) or incorrect word ordering (e.g., ”Юрий Мирошник” instead of ”Мирошник Юрий”). Such errors clearly could not be fixed by our model.
- *Word form* errors. We split them into four categories depending on whether the expected normalization could have been found in the pymorphy2 dictionary (*Known-*/*Unknown-*) and whether the predicted normalization could have been found in the dictionary (*-Known*/*-Unknown*). Most of the errors are in the Known-Known category, which means that the model was not able to select a correct word form of a common word.

We believe that *space*, *punctuation* and *alignment* errors are minor and may be ignored.

The qualitative analysis can be found in Table 6.

	Generic		Named	
Error	Single model	Ensemble	Single model	Ensemble
Space	-	-	11 (4.30%)	11 (5.19%)
Punctuation	15 (7.77%)	17 (9.60%)	25 (9.77%)	22 (10.38%)
Alignment	32 (16.58%)	30 (16.95%)	20 (7.81%)	14 (6.60%)
Capitalization	-	-	39 (15.23%)	38 (17.92%)
Known-Known	136 (70.47%)	121 (68.36%)	121 (47.27%)	95 (44.81%)
Known-Unknown	4 (2.07%)	3 (1.69%)	7 (2.73%)	9 (4.25%)
Unknown-Known	3 (1.55%)	3 (1.69%)	5 (1.95%)	6 (2.83%)
Unknown-Unknown	3 (1.55%)	3 (1.69%)	39 (15.23%)	28 (13.21%)

Table 6: Error analysis of multitask trainable Syntax-BERT model and its ensemble

It is clear that ensembling helps to select a better word form and reduces errors count this way. It shows the instability in the training process because a significant number of errors was fixed by the majority vote over the predictions of similar models.

6 Conclusion

We proposed an alternative way to perform span normalization and showed that it works reasonably well on the RuNormAS dataset achieving the best or almost the best results there. We believe that our method should be considered favourably when enough training data is available for normalizer because it is simpler than language model training or normalization rules designing in such a scenario. The inference speed is bound by the speed of the syntax parser’s encoder, which should be practicable in most applications.

References

[1] Anastasyev D. G. Exploring Pretrained Models for Joint Morpho-Syntactic Parsing of Russian // Proceedings of the International Conference ”Dialog 2020”. — Moscow, Russia, 2020. — P. 1–12.

[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // 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). — Minneapolis, Minnesota, 2019. — P. 4171–4186.

[3] Gardner Matt et al. AllenNLP: A Deep Semantic Natural Language Processing Platform // Proceedings of Workshop for NLP Open Source Software (NLP-OSS). — Melbourne, Australia, 2018. — P. 1–6.

[4] Korobov Mikhail. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. — 2015. — P. 320–332.

[5] Korzun V.A. Proper Names Normalization without Semantic Parsing // http://www.dialog-21.ru/media/4671/доклад_корзун.pdf. — 2019.

[6] Lyashevskaya O. et al. GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing // Proceedings of the International Conference ”Dialog 2020”. — Moscow, Russia, 2020. — P. 553–569.

[7] Wolf Thomas et al. Transformers: State-of-the-Art Natural Language Processing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — Online, 2020. — P. 38–45.

DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model

Nikolay Arefyev[◇△▽]
Moscow, Russia
nick.arefyev@gmail.com

Maksim Fedoseev[◇]
Moscow, Russia
maxim.fedoseev13@gmail.com

Vitaly Protasov[○]
Moscow, Russia
vitaly.protasov@skoltech.ru

Daniil Homskiy[◇]
Moscow, Russia
homdanil123@gmail.com

Adis Davletov[◇▷]
Moscow, Russia
dev.davletov@gmail.com

Alexander Panchenko[○]
Moscow, Russia
A.Panchenko@skoltech.ru

[◇]Lomonosov Moscow State University
[△]Samsung Research Center Russia
[▽]HSE University
[○]Skolkovo Institute of Science and Technology
[▷]RANEPA

Abstract

In this paper, we describe our solution of the Lexical Semantic Change Detection (LSCD) problem. It is based on a Word-in-Context (WiC) model detecting whether two occurrences of a particular word carry the same meaning. We propose and compare several WiC architectures and training schemes, and also different ways to convert WiC predictions into final word scores estimating the degree of semantic change.

We participated in the RuShiftEval LSCD competition for the Russian language, where our model achieved 2nd best result during the competition. During post-evaluation experiments we improved the WiC model and managed to outperform the best system. An important part of this paper is detailed error analysis where we study the discrepancies between WiC predictions and human annotations and their effect on the LSCD results.

Keywords: semantic change detection, XLM-R, contextualized embeddings
DOI: 10.28995/2075-7182-2021-20-16-30

ГлубокоеНедоразумение: какие значения тяжело различать контекстуализированным моделям значений слов

Арефьев Николай[◇△▽]
Москва, Россия
nick.arefyev@gmail.com

Федосеев Максим[◇]
Москва, Россия
maxim.fedoseev13@gmail.com

Протасов Виталий[○]
Москва, Россия
vitaly.protasov@skoltech.ru

Хомский Даниил[◇]
Москва, Россия
homdanil123@gmail.com

Давлетов Адис[◇▷]
Москва, Россия
dev.davletov@gmail.com

Панченко Александр[○]
Москва, Россия
A.Panchenko@skoltech.ru

[◇]Московский Государственный Университет им. М. В. Ломоносова
[△]Московский Исследовательский Центр Самсунг
[▽]Национальный исследовательский университет «Высшая школа экономики»
[○]Сколковский Институт Науки и Технологий
[▷]РАНХиГС

Аннотация

В этой статье мы описываем наше решение проблемы обнаружения семантических сдвигов значений слов (LSCD). Оно основано на модели Word-in-Context (WiC), которая по двум вхождениям одного слова определяет, имеют ли эти вхождения одно значение. Мы предлагаем и сравниваем несколько архитектур и схем обучения WiC, а также различные способы преобразования предсказаний WiC в итоговые оценки для слов, отражающие степень их семантического сдвига.

Мы участвовали в соревновании RuShiftEval по обнаружению семантических сдвигов значений слов русского языка, где наше решение заняло 2ое место. После завершения соревнования мы улучшили нашу модель WiC и смогли превзойти лучшую систему. Важной частью этой статьи является подробный анализ ошибок, в котором мы изучаем расхождения между прогнозами WiC и аннотациями людей, а также влияние этих расхождений на результаты LSCD.

Ключевые слова: обнаружение семантических изменений, XLM-R, контекстуализированные вектора слов

1 Introduction

The task of Lexical Semantic Change Detection (LSCD) is to determine how senses of a particular word changed between two time periods. The change of word senses in time is a rather complex phenomenon. Thus, several formal settings exist for this task with different annotation schemes and quality metrics, each having its own pros and cons. We developed our system for the RuShiftEval competition [5], where the main goal was to rank the given test words similarly to the ranking by their gold COMPARE scores [11]. To obtain the gold COMPARE scores during the construction of the dataset, the annotators were presented several dozen sentence pairs, each representing two occurrences of the same word, one sampled from an old corpus and another from a new corpus. The annotators estimated the similarity of those word occurrences in meaning on a 1-4 scale, where larger values corresponded to higher degree of similarity. This is commonly known as the Word-in-Context (WiC) task [7]. To obtain gold COMPARE score for a particular word, the annotations of sentence pairs containing this word were averaged.

We decided to follow the same word scoring procedure, but replaced human annotators with a WiC model to solve the task. Our system achieved the second-best result during the competition. After the competition, we managed to outperform the winner by improving the architecture and the training procedure of our WiC model.

Our main contributions are the following.

- An approach to the RuShiftEval LSCD task employing a WiC model is proposed. Our system implementing this approach achieved the 2nd best result in the competition and outperformed the winner after the competition.
- We proposed different architectures and training schemes of a WiC model and compared their performance in the LSCD task.
- A detailed error analysis is performed, showing which distinctions between word senses a WiC model annotates differently compared to human annotators and how it effects the final LSCD results.

2 Related work

RuShiftEval [5] is the first LSCD shared task for the Russian language. Its data annotation scheme and metrics follow those proposed as a part of Diachronic Usage Relatedness (DURel) dataset for German [11]. In [9] the RuSemShift dataset is introduced, which was proposed as the training and the development set for RuShiftEval. In our work, we employed the COMPARE scores from that dataset, which estimate the similarity in meaning between two time periods both for words and sentence pairs. We left out other potentially useful information about the variability of meaning inside each time period. An alternative dataset annotation procedure and metrics were proposed in SemEval-2020 Task 1 [10], where the authors tried to account for the appearance or disappearance of relatively rare word senses, which the COMPARE metric is not sensitive to. They basically clustered word occurrences corresponding to their senses using human annotators instead of an automated clustering system. Based on this gold standard clustering, two subtasks were proposed. The first subtask required binary classification determining if the set of word senses has changed. The second subtask required ranking words according to the change in sense frequencies.

In [9] several LSCD models are compared for the Russian language. One of them is ELMo [6], which is a recurrent neural network trained as a language model on texts from the Russian National Corpus [1] (RNC) and used to build contextualized embeddings of target words. Then they compute the cosine similarity

[1] https://ruscorpora.ru

or the Jensen-Shannon divergence to receive the semantic change score. The best quality among the models considered by them according to the COMPARE metric is 0.403 for the first part of RuSemShift and 0.541 for the second part. Since we have split the RuSemShift dataset into training and development parts, our results are not directly comparable, though the experiments in Section 4 suggest that our system significantly improves those results for RuSemShift.

Next we describe best LSCD methods proposed for the SemEval-2020 Task 1 [10].

UG Student Intern team[10] achieved the 1st place in subtask 2 (ranking). They use word2vec SGNS word embeddings with the Orthogonal Procrustes alignment and compute Euclidean distance instead of cosine distance to evaluate the semantic change. Unfortunately, this team did not publish a system description paper.

Jiaxin & Jinan team [12] achieved the 3rd place in subtask 1 (binary classification) and the 2nd place in subtask 2. They use Temporal Referencing [3] to solve the alignment problem. Instead of training two models and then aligning, they train one model, in which the postfix "_new" is added to the target words from the examples of the new time period, and the postfix "_old" is added to the words from the examples of the old time period. To get a threshold for the classification task, they fit a Gamma distribution by the cosine distance (Gamma Quantile Threshold method). Word embeddings were extracted from fine-tuned BERT or SGNS.

UWB team [8] achieved the 1st place in subtask 1 and the 4th place in subtask 2. They use Canonical Correlation Analysis (CCA) and modification of the Orthogonal Transformation from VecMap for linear transformation to move from the source space (first time period) to the target space (second time period). For word embeddings, they also employ SGNS. After a linear transformation, they calculate the cosine similarity. They proposed different ways to find an optimal threshold based on averaging cosine similarities for each word.

3 Semantic change detection method

To estimate the COMPARE score of a particular target word during the dataset construction, the pairs of sentences were sampled from two time periods. For each pair of sentences, each annotator specified a number from 1 to 4, where 1 stands for unrelated word meanings, and 4 stands for identical meanings. Those annotations were averaged across annotators first, and then across all sentence pairs containing the target word. To approximate this process of computing word scores, our system employs a Word-in-Context model, which solves the same task as the annotators. The input data for this model consists of a target word and two sentences in which the word appears. The model determines whether the target word is used in the same sense or different senses.

First, we have built and trained a WiC model. Then for each test word we retrieved sentences containing this word, and constructed pairs of sentences belonging to different time periods. The scores for sentence pairs were obtained from the WiC model and aggregated into the final word scores.

3.1 Construction of WiC sentence pairs

For each test word, we retrieved the examples from the diachronic subset of the RNC corpus[2] (RNC). This corpus consists of three parts: *Soviet, Pre-Soviet, Post-Soviet*. Since the corpus contains only plain texts, to find examples for a particular word in all forms we used Rulemma lemmatizer[3].

Next, we sampled 100 sentences (or all sentences, if there were fewer) from each time period and constructed sentence pairs for each pair of periods. The examples were sampled from a uniform distribution. For each word, we removed 25% of the longest sentences, 25% of the shortest sentences, and all sentences where the target word was the first or the last word. This was done based on the intuition that for optimal WiC performance, the context shall be long enough, but not very long for faster processing, and there shall exist some preceding and succeeding words for the target word, which often provide the most informative context. In appendix B additional experiments with removal of short and long examples are described.

[2] https://ruscorpora.ru/new/en/corpora-usage.html

[3] https://github.com/Koziev/rulemma

3.2 Scoring sentence pairs

3.2.1 WiC model architecture

As a backbone for our WiC model we employed the XLM-R masked language model [2], which was pre-trained on about 2TB of texts in 100 languages. This enables us using WiC training data in different languages, which improves the overall performance. To score each sentence pair we feed it to XLM-R in the standard format:

```
<s>sentence1</s>sentence2</s>
```

After that we calculate the contextualized embeddings for the target word in each sentence by averaging the outputs of the last transformer layer corresponding to all of its subwords (*mean* pooling). Additionally we experimented with the $first$ pooling when the output on the first subword is taken.

Then we aggregate two target word embeddings from two sentences using one of the following options:

1. **concat**: (x, y), the concatenation of two embeddings;
2. **comb_dmn**: $(x-y, \bar{x} \circ \bar{y})$, the concatenation of the difference (**d**) of non-normalized and component-wise product (**m**) of normalized (**n**) embeddings;
3. **dist_l1**: $\|x - y\|_1$, L1-distance between the embeddings, also known as the Manhattan distance;
4. **dist_l1ndotn**: $(\|\bar{x} - \bar{y}\|_1, \langle \bar{x}, \bar{y} \rangle)$, the concatenation of L1-distance and the dot product (**dot**) of the normalized (**n**) embeddings. The second feature is essentially the cosine similarity.

While the first two options produce high-dimensional vectors, the other two result in a vector with one or two components only. The obtained vector is passed through a classification head, which has a dense layer of size hs and $tanh$ activation, followed by a linear layer, or only a linear layer (denoted as $hs = 0$). For the first two aggregation options, we always used a hidden layer of the size $hs = 1024$ (the size of the embeddings in large XLM-R). For the distance-based inputs we found that a linear head outperformed non-linear one. We inserted batch normalization [4] before the first layer, which proved to be especially efficient for L1-distance inputs.

Based on the intuition that the similarity between two-word occurrences does not depend on the order of sentences in a sentence pair, we employed training time and test time augmentation. For each example, we create an additional one by swapping two sentences. Thus, two scores were obtained for each example. For training, we always use that augmentation since, from our preliminary experiments, it helps for some architectures and never hurts. For inference we either take the first score (i.e. disable test time augmentation), or average them.

3.2.2 WiC training

We also look at different ways to train the model using MCL-WiC[4] and RuSemShift [9] datasets. **MCL-WiC** consists of an English training set (8000 examples), multilingual development sets with both sentences in one of the following languages: English, French, Russian, Arabic, Chinese (1000 ex. each), and test sets with cross-lingual (one sentence in English, the second in another language) and multilingual parts (1000 ex. each). **RuSemShift** consists of two pairs of periods: there are pairs of sentences from pre-Soviet and Soviet periods in the first part, and Soviet and post-Soviet in the second part. We have split each part into a train and a development subsets, ensuring there is no intersection between the target words in those subsets (lexical split).

The weights of the pre-trained XLM-R large model were used for initialization, and then we train model on the following datasets or their combinations.

1. **MCL-WiC** training set consists of the original MCL-WiC training set in English, 70% of each non-English development set (2800 ex.) and all trial sets (72 ex.). The development set is the rest 30% of non-English development sets (1200 ex.) and the full English development set (1000 ex.).
2. **MCL-WiC en-en** training set is the original MCL-WiC training set. It is used to estimate the performance of a model trained only on English WiC data.
3. **MCL-WiC ru-ru** means that we train only on data in Russian, including 70% of the development set and the whole test and trial sets (1708 ex. in total).

[4] https://github.com/SapienzaNLP/mcl-wic

4. **RuSemShift** training set is combined from both training sets from our split, 3898 examples in total. The development set consists of two parts - one for pre-Soviet and Soviet periods, another for Soviet and post-Soviet periods.

The training process consists of one or two steps. Each step employs its own training set and loss function. All training schemes are enumerated in Table 1. Cross entropy and mean squared error losses are denoted as CE and MSE accordingly.

To employ all training data we have in Russian simultaneously, we have developed a new loss function $MSE+$, which can handle both binary targets from MCL-WiC and real-valued targets from RuSemShift. For the real-valued targets, it is equivalent to MSE loss, while for binary targets, it penalizes the predictions using MSE loss only when they are outside (1,2) interval for negative examples or outside (3,4) interval for positive. Since the labels are binary, we only know whether the meaning is similar or different, but do not know the exact degree of similarity, thus any prediction from the appropriate intervals is suitable. Another option is binarizing RuSemShift and using CE loss. Examples with scores not less than 3 were treated as positive. For negative examples we set the threshold of 2 during the competition and 3 in the following experiments.

Train#1	Loss#1	Train#2	Loss#2
MCL-WiC	CE	-	-
MCL-WiC en-en	CE	-	-
MCL-WiC ru-ru	CE	-	-
RuSemShift	MSE	-	-
MCL-WiC	CE	MCL-WiC ru-ru	CE
MCL-WiC	CE	RuSemShift	MSE or CE
MCL-WiC	CE	RuSemShift + MCL-WiC ru-ru	MSE+ or CE

Table 1: Training schemes. We take XLM-R pre-trained as a MLM, fine-tune it first on Train#1 with Loss#1, and then optionally on Train#2 with Loss#2.

3.3 Scoring words

Mean. The simplest method to compute the final score for a particular word and a pair of periods is to calculate the mean of scores for all corresponding sentence pairs. across all pairs of sentences containing the target word for a pair of periods.

Isotonic regression (IsoReg). Depending on the loss function, the predicted scores for sentence pairs may not be in the same range or distribution as the human scores. This may result in incorrect word raking after simple averaging. We try to make sentence pair scores more similar to human scores by fitting isotonic regression [1]. We feed the predicted score for a sentence pair to the isotonic regression and get the modified score, which is then averaged across sentence pairs. Isotonic regression is trained on sentence pairs from the training subset of RuSemShift (3989 training examples).

Linear regression (LinReg). Instead of using simple averaging of scores for sentence pairs, we can use a trainable function to predict word scores. Input features are the mean and quartiles of scores for all sentence pairs of a particular word and pair of periods. We trained this model on words from the training subset of RuSemShift (69 training words). The features can be calculated both on sentence pairs from RuSemShift, or sampled sentence pairs (**LinReg_s**).

4 Experiments and results

To evaluate our WiC model and select its hyperparameters, we employed two types of metrics. Spearman correlation between model scores and gold scores for sentence pairs from RuSemShift (***sentSpear***) shows how well the model solves WiC task. Spearman correlation between the final word scores and the gold values of the COMPARE metric (***wordSpear***) estimates the final performance on LSCD task.

Both metrics are calculated on each of two development sets (***dev1*** for pre-Soviet – Soviet, ***dev2*** for Soviet – post-Soviet) and three test sets (***p12*** for pre-Soviet – Soviet, ***p23*** for Soviet – post-Soviet, ***p13*** for pre-Soviet – post-Soviet). For majority of experiments we show averaged dev and a test metrics. In the experiments with WiC model architecture and training, for evaluation we employed the same sentence pairs that were annotated by humans, otherwise we could not calculate *sentSpear*. However, for the final results in Table 2 and word scoring experiments in Section 4.3, 100 sampled pairs for each word and pair of periods were used instead. Additionally, when training on MCL-WiC we employ accuracy on the English dev set (***en-acc***), or an average accuracy over non-English dev sets (***nen-acc***) for early stopping.

Method/Team	**Avg**	**p12**	**p23**	**p13**
Best results of other teams				
GlossReader (1st best result)	**0.802**	0.781	**0.803**	**0.822**
vanyatko (3rd best result)	0.720	0.678	0.746	0.737
Our submissions: team *DeepMistake* (2nd best result)				
first+concat on MCL_{CE}^{en-acc} →$\text{RSS}_{MSE}^{dev2-sentSpear}$ (***M1***), LinReg	**0.791**	**0.798**	0.773	0.803
M1, Mean	0.789	0.794	0.773	0.799
M1, IsoReg	0.789	0.793	0.775	0.798
p12, p13: *M2*; *p23*: first+concat on $\text{MCL}_{CE}^{nen-acc}$ →RSS+ruMCL$_{CE}^{dev1-sentSpear}$, IsoReg	0.785	0.773	**0.802**	0.780
mean+dist_l1ndotn-hs300 on $\text{MCL}_{CE}^{nen-acc}$ → RSS+ruMCL$_{MSE+}^{dev1-sentSpear}$ (***M2***), Mean	0.780	0.773	0.786	0.780
LinReg on *M1* + *M2* + *M3*	0.780	0.756	0.772	**0.811**
p12, p13: mean+dist_l1 on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{MSE}^{dev2-sentSpear}$ (***M3***), Mean *p23*: first+concat on $\text{MCL}_{CE}^{nen-acc}$ →RSS+ruMCL$_{CE}^{dev2-sentSpear}$, Mean	0.779	0.749	0.801	0.788
p12, p13: *M2*; *p23*: max+concat on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{MSE}^{dev2-sentSpear}$, Mean	0.778	0.779	0.775	0.779
p12, p13: *M3*, LinReg* *p23*: mean+comb_dmn on MCL_{CE}^{en-acc} →$\text{RSS}_{MSE}^{dev2-sentSpear}$, LinReg	0.757	0.750	0.732	0.788
Our best models with ablation analysis				
mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{MSE}^{dev2-sentSpear}$, Mean	**0.823**	**0.825**	**0.821**	**0.823**
mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →RSS+ruMCL$_{MSE+}^{dev2-sentSpear}$, Mean	0.803	0.800	0.798	0.811
mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$, Mean	0.776	0.777	0.778	0.772
mean+concat on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{MSE}^{dev2-sentSpear}$, Mean	0.768	0.760	0.759	0.784
mean+concat on $\text{MCL}_{CE}^{nen-acc}$ →RSS+ruMCL$_{MSE+}^{dev2-sentSpear}$, Mean	0.791	0.790	0.786	0.797

Table 2: Evaluation and post-evaluation results. MCL is MCL-WiC, RSS is RuSemShift. *LinReg** denotes LinReg on two features only - the mean and the median. M1, M2, M3 abbreviate duplicated WiC model specifications. *LinReg on M1+M2+M3* denotes LinReg on features from all those models. **concat**: (x, y), **comb_dmn**: $(x - y, \bar{x} \circ \bar{y})$, **dist_l1**: $\|x - y\|_1$, **dist_l1ndotn**: $(\|\bar{x} - \bar{y}\|_1, \langle \bar{x}, \bar{y} \rangle)$

4.1 Submissions and post-competition improvements

During the evaluation phase, we made 10 submissions. Their results with the best results of other teams are shown in Table 2. The first 2 arguments of the method name indicate which pooling and aggregation of the two target word embeddings were used. Then the training scheme is specified with subscript and superscript specifying loss function and early stopping metric. The word scoring method is appended after the comma when it differs from the default mean over sentence pairs. In some submissions, for different pairs of time periods we used predictions of different models.

After the competition, we analyzed various aggregation methods of the target word embeddings and training options and managed to achieve better results than the winning submission for all pairs of time periods. The best model uses *dist_l1ndotn* without hidden layer (*hs0*) and is trained on MCL-WiC first, then on RuSemShift with MSE loss. From ablations we notice that the average quality is reduced by 5 points when only the first training step is left. For *dist_l1ndotn* is better to train on RuSemShift with MSE loss than on RuSemShift+ruMCL-WiC with $MSE+$ loss, but for *concat* vice versa.

In appendix A we additionally compare the performance of our best model with human performance.

4.2 WiC architecture and training scheme

Embeddings aggregation. To compare different methods of aggregation of target word embeddings, we trained several WiC models on MCL-WiC, and then optionally fine-tuned them on RuSemShift with MSE loss. For early stopping we employed *nen-acc* on the first dataset and *dev2-sentSpear* on the second. We used mean pooling for subwords and also Mean aggregation of scores for sentence pairs.

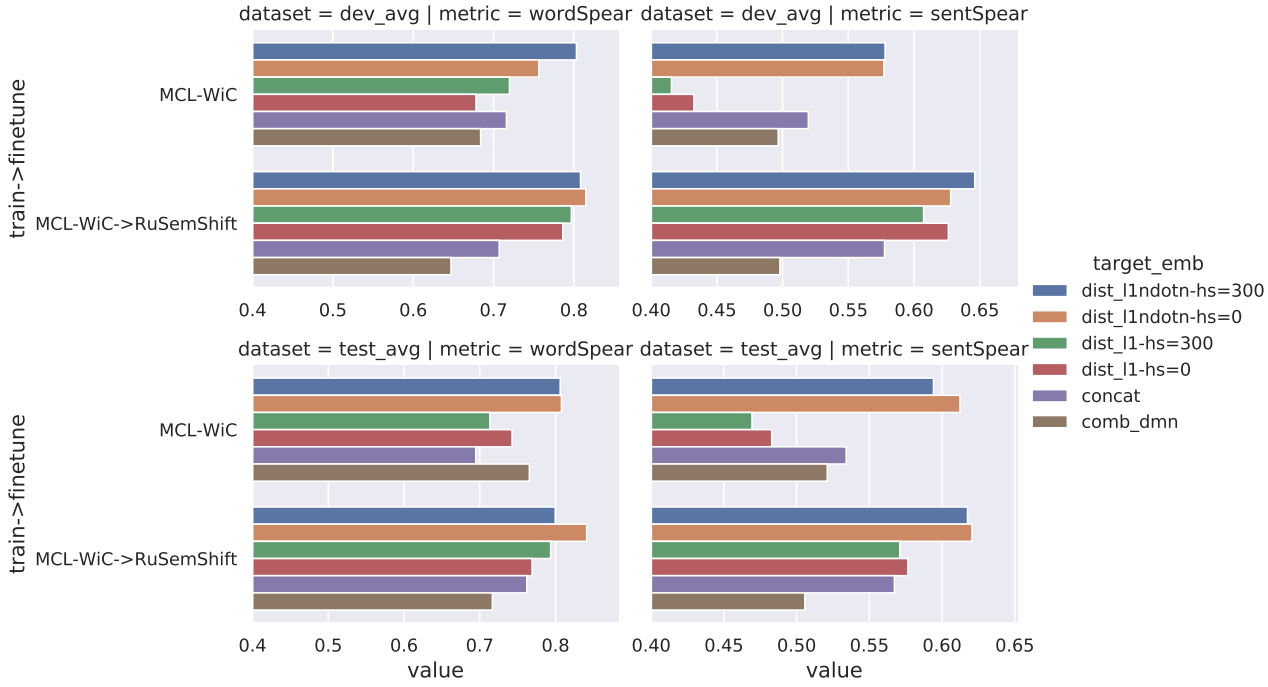


Figure 1: Comparison of the target embedding aggregations. Values are sentSpear (left) and wordSpear (right) on dev (up) and test (down).

Figure 1 shows that the best methods of combining two embeddings of the target word are $dist\_l1ndotn$ either with dense layer size $hs = 300$ or without dense layer. Generally, training on RuSemShift after training on MCL-WiC improves performance a bit or at least does not hurt. And for the two-step training, the basic one-dimensional $dist\_l1$ works better than high dimensional concatenation or $comb\_dmn$. Moreover, concatenation always works better than $comb\_dmn$ for two-step training, but when training only on MCL-WiC $comb\_dmn$ gives higher *test-wordSpear*.

WiC model training. Figure 2 compares different training schemes. All compared models employ the $mean$ subword pooling, the concatenation of target word embeddings and *dev2-sentSpear* for early stopping. Training schemes are specified in the following format.

- In the case of one-step training: ”training dataset” (loss function of the training).
- In the case of two-step training: ”training dataset #1” -> ”training dataset #2” (loss function of the second training). The loss function of the first step is *CE* by default.

Evidently, two-step training procedure employing both the large multilingual MCL-WiC dataset and the task-specific RuSemShift dataset significantly boosts the performance compared to single-step training on any of the datasets alone. Using the combination of RuSemShift and the part of MCL-WiC in Russian with the proposed *MSE+* loss for the second training step generally gives the best overall performance, except for the *dev-wordSpear*, which is a little better for another scheme presumably due to the metric variance. Using RuSemShift on the second training step shows much better performance than employing the Russian part of MCL-WiC for the same purpose. For the single-step training schemes, we observe a large difference between dev and test performance for model trained on only English or Russian parts of MCL-WiC. The model trained on RuSemShift with MSE loss ranks sentences much worse than the one trained on MCL-WiC, but their results of the final word ranking are comparable. Surprisingly,

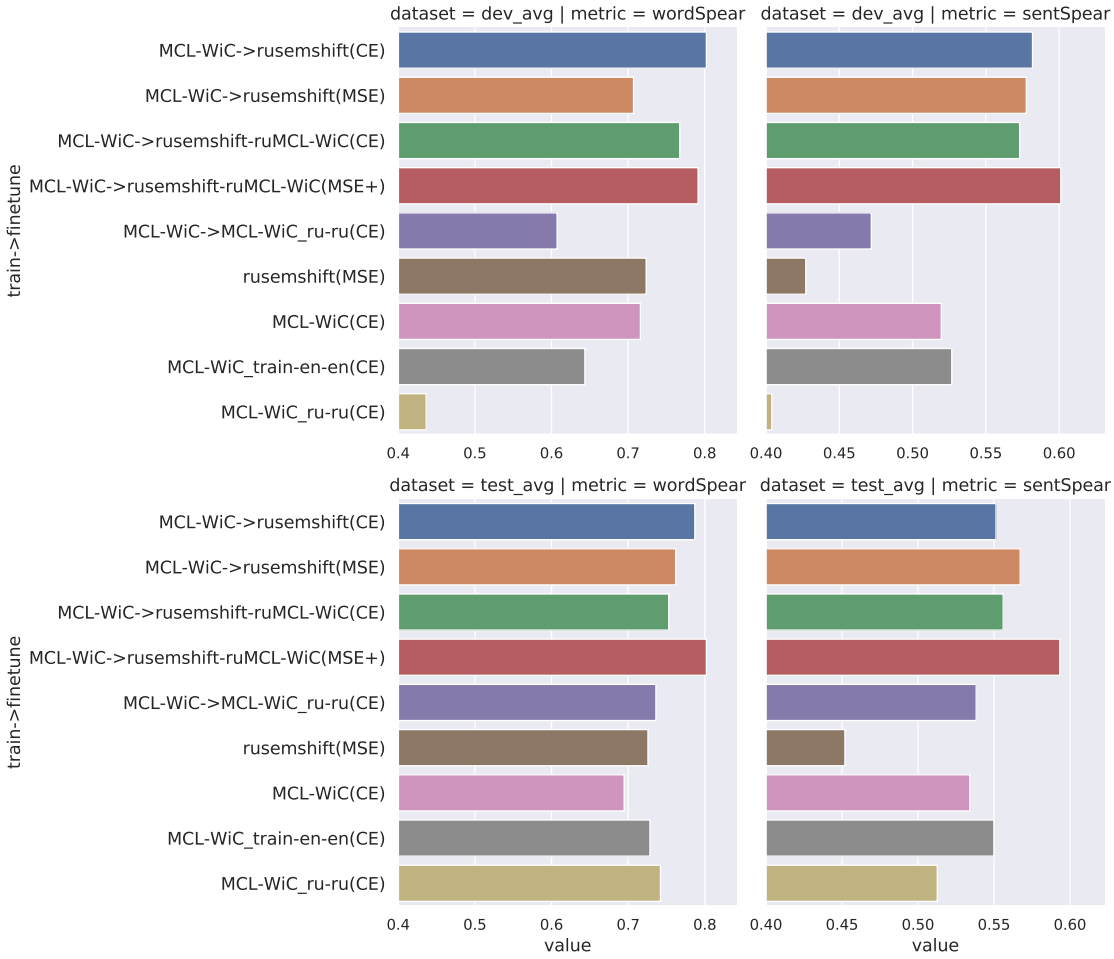


Figure 2: Comparison of the WiC model training. Aggregation of the target word embeddings is *concatenation*.

one can train the model on the English part of MCL-WiC only and still obtain a relatively good LSCD results comparable to the 3rd best team in the competition. This shows strong zero-shot cross-lingual transfer capabilities of the underlying XLM-R model. Training on the Russian part of MCL-WiC alone gives mixed results presumably due to much smaller size of this part.

4.3 WiC scores aggregation for word scoring

Figure 3 shows the comparison of different methods of obtaining word scores from sentence pairs scores. The quality does not depend on the method of word scoring as much, as on the WiC model architecture or training scheme, but the linear regression gives consistently the best or nearly the best results.

Finally, we estimated how the quality of the final word ranking depends on the number of sentence pairs sampled for each word. We sampled each number of sentence pairs 30 times and calculated the mean and the standard deviation of the target *wordSpear* metric. Figure 4 shows the results for test words. As we expected, the target metric improves rapidly with the number of sampled sentences, and also its standard deviation decreases. Even for 80 samples std is 0.8 point, suggesting that different sampled pairs result in 2-3 point difference in the target metrics. This figure also suggests that if only several dozen of sentence pairs were annotated for each word during the construction of a LSCD dataset, the difference between methods of 5-10 points may be due to chance. The green dashed line shows the results when we run our system on the same sentence pairs that were annotated by humans. Unsurprisingly, this results in better

estimate of the gold word scores. However, the difference becomes small as the number of sampled pairs approaches one hundred.

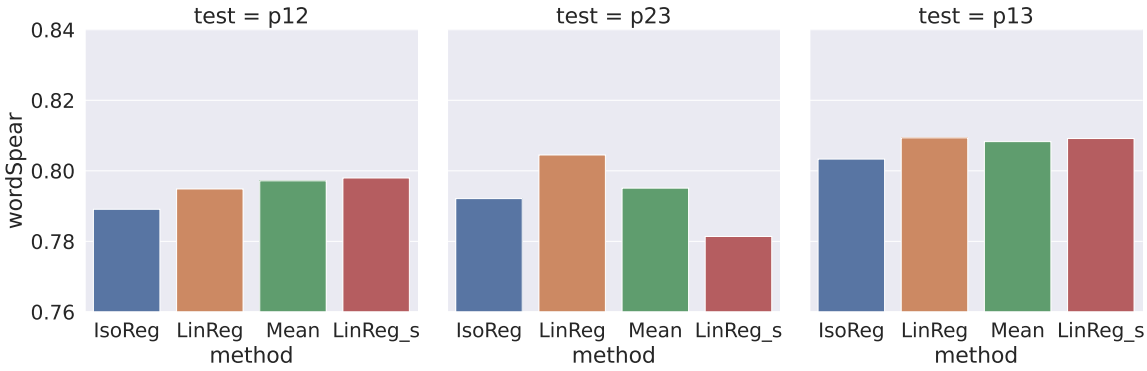


Figure 3: Word scoring methods. Results for the predictions on the sampled test pairs. Model: dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →RSS+ruMCL$_{MSE+}^{dev2-sentSpear}$

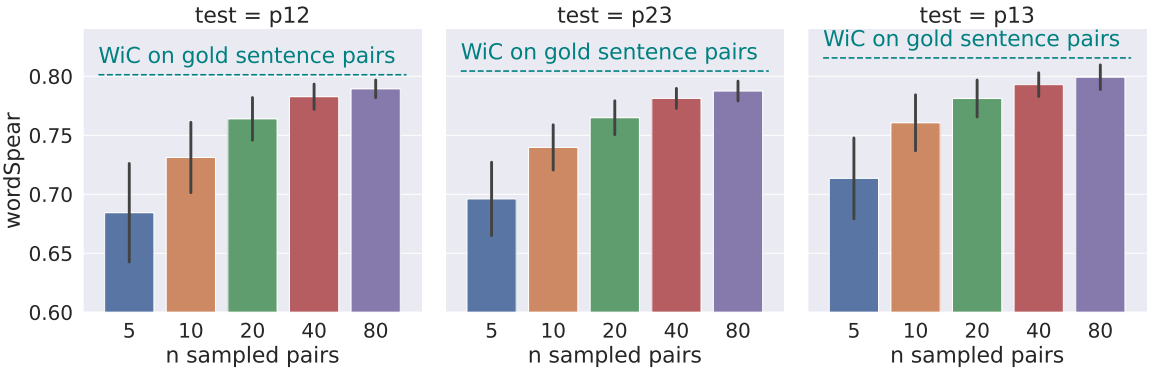


Figure 4: Dependence on the number of sampled pairs. Error bars show one standard deviation. Model: mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →RSS+ruMCL$_{MSE+}^{dev2-sentSpear}$

5 Error analysis

This section is devoted to getting some insights into the types of errors, their relative frequencies and reasons. We used the test set consisting of 99 unique target words and sentence pairs for them with human annotations, which was provided by the organizers after the competition. We employed the WiC model with first subword pooling and concatenation of target embeddings trained on $\text{MCL}_{CE}^{nen-acc}$ →RSS+ruMCL$_{CE}^{dev1-sentSpear}$ with *mean* word scoring. This model achieved one of the best results among our submissions.

We define $\Delta Rank$ as the difference between the rank of a word predicted by our model and the gold rank. Words with a high $|\Delta Rank|$ value are considered serious errors, we decided to focus on the words with $|\Delta Rank| \geq 25$. This resulted in 24 words from *p12*, 18 words from *p23*, and 13 words from *p13*. Many of those words are incorrectly ranked in several pairs of time periods, thus, there are 27 unique words that we analyzed in total. They are shown in Figure 5.

5.1 Classification of WiC model mistakes

To understand the reasons of incorrect ranking of words under consideration, for each of them we have selected 5-7 annotated sentence pairs with the highest difference between gold annotations and the predicted WiC scores (the difference was 1.5 at least). We obtained 171 pairs of sentences in total, and annotated them according to the error types described below.

Figure 5: Words with large difference between the predicted and the gold ranks at least for one pair of epochs.

Table 3 shows the results of error analysis and examples. More examples can be found in appendix C. We identified four typical reasons (error types) of the high disagreement between WiC model predictions and human annotations.

1. *Model can not find the difference*. The model incorrectly classifies two word occurrences as having the same meaning.
2. *Model sees wrong difference*. The model incorrectly classifies two word occurrences as having different meanings.
3. *Model seems to be right*. These are pairs of sentences, that in our opinion were correctly classified by the model, but incorrectly annotated by one or more annotators.
4. *Ambiguity*. From the context we could not understand whether two word occurrences have the same meaning.

The most frequent error type (39% of all analyzed examples) is *Model can not find the difference*. This strongly effects the final ranking of words like *тачка* (car / cart), *увольнение* (dismissal / vacation), *дядька* (servant tutor / uncle / mister) that obtained or lost the first sense, which the model cannot distinguish from others.

Error types *Model seems to be right* and *Model sees wrong difference* have almost equal frequency (23.2% and 22.8%). In the sentences of the first type, the target word usually has two senses that are sim-

Table 3: Types of errors, their proportions and examples.

Type	Gold	Model	Pair of sentences
Model can not find the difference. **39%**	1 1 1	4	Пазульский был приговорен в Одессе к 12 годам, 100 плетям и трем годам прикования к **тачке**. Они с Верой выпорхнули из дверей **тачки** и поплыли в невесомости спортивного зала.
Model seems to be right. **23.2%**	1 3 3	1	”Если гомеопаты не обходятся без прививок, то чего же нам стесняться!” – так утешают себя **маньяки** прививок ... Только преследований **маньяка** мне в таком состоянии не хватало
Model sees wrong difference. **22.8%**	4 4 4	1	На руках она с усилием тащила Федьку, прижав его поперек **живота**, чем он нисколько не смущался. Боли в **животе** не то стали слабее, не то он к ним привык
Ambiguity. **15%**	1 4 3	4	Призыв 1902 года в 1924 году дал Красной армии 4.700 партийцев; при **увольнении** же этого возраста в 1926 году в запас Красная армия дала стране 19.439 партийцев. **Увольнение** производится в порядке очередности.

ilar to some degree. Often there is disagreement between annotators in such cases. Since there were only three annotators, even one incorrect annotation significantly effected the resulting mean human score for a sentence pair. For instance, Table 3 contains sentences for the word *маньяк* (maniac), which has a direct meaning (a mentally ill man) and a figurative meaning (a person obsessed with a passionate attraction to something). The model correctly distinguished these senses, but two out of three annotators decided that those senses are very similar. For sentence pairs of the type *Ambiguity* there is large disagreement between annotators, hence, the aggregated gold annotation is almost random. For instance, in Table 3 the word *увольнение* (dismissal / vacation) in the second sentence can express any of its meanings.

This error analysis suggests that about 60% of the analyzed sentences are actually incorrect predictions, while the rest are hard cases where human annotators disagree with each other. Such cases can be easily found by high deviation between annotations, and it may be beneficial involving additional annotators to resolve disagreement or to filter unclear examples.

Another technical issue we found were incorrectly tagged examples in the test set. The organizers of the competition published the test set with annotated sentences. It consists of three files, each of them contains approximately 3000 pairs of sentences containing some target word highlighted by special tags **, **. However, in a significant proportion of sentences (Table 4) the target word was not tagged, which was a problem for our WiC model. Finally, we fixed the largest part of incorrect examples, the rest consisted of sentences with abbreviated target words, for example: *апостол* $\rightarrow$ *ап.*, *век* $\rightarrow$ *в.* Our fixes are merged into the published version of the dataset, hopefully, making the dataset better.

Table 4: Incorrectly tagged examples.

Test set	Total examples	Bad examples	Fixed examples
p12	2965	236	214
p23	2967	221	191
p13	2969	249	207

6 Conclusion

We have proposed an approach to Semantic Change Detection employing a Word-in-Context model and found that it has strong performance achieving the 2nd best result among other competing approaches, which can be further improved to outperform the 1st best result by improving WiC architecture and training procedure. Regarding the architecture, we have found that a simple linear head on top of concatenated L1 distance and dot product between contextualized XLM-R embeddings provides better performance than more common alternatives like embedding concatenation and non-linear classification heads. For

the training procedure, using training on a large multilingual WiC dataset first and fine-tuning on a task-specific RuSemShift data later results in the best overall performance.

We performed a detailed error analysis of sentences, where disagreement between model and annotators was the highest. To understand why the model fails, we annotated 171 pairs of sentences and revealed four different types of low model results. Such mistakes included examples when the model correctly distinguished the difference or similarity of senses, when annotators were wrong, and vice versa.

References

[1] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An Empirical Distribution Function for Sampling with Incomplete Information. *The Annals of Mathematical Statistics*, 26(4):641 – 647, 1955.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[3] Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[5] Andrey Kutuzov and Lidia Pivovarova. Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*, 2021.

[6] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[7] Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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)*, pages 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Ondřej Pražák, Pavel Přibáň, Stephen Taylor, and Jakub Sido. UWB at SemEval-2020 task 1: Lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 246–254, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[9] Julia Rodina and Andrey Kutuzov. RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[10] Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[11] Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[12] Jinan Zhou and Jiaxin Li. TemporalTeller at SemEval-2020 task 1: Unsupervised lexical semantic change detection with temporal referencing. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation,* pages 222–231, Barcelona (online), December 2020. International Committee for Computational Linguistics.

A Comparison with human quality

It is interesting to compare the performance of our best model with human performance. There are three columns containing annotations of each sentence pair by different annotators in RuShiftEval [5] and five columns in RuSemShift [9]. To estimate human performance, we excluded examples with one or more annotations absent or having zero (undecided) values. For examples with full set of annotations we calculated *wordSpear* and *sentSpear* between each column and the mean of other columns (*annotN VS mean_wo_N*). However, this method of human performance estimation shall be taken with a grain of salt. It is totally correct only if each human annotated the whole dataset, which may not be true for RuSemShift and RuShiftEval datasets that were annotated using crowdsourcing. We suppose, that each of three or five annotators specified in the datasets really correspond to several humans. In this case we are probably overestimating human performance of word ranking (*wordSpear*), because after averaging across all sentence pairs individual misconceptions about senses of a particular word resulting in incorrect annotations of some sentence pairs with this word by one human will be partially compensated by annotations of other sentence pairs with the same word obtained from other humans. To get better estimates of human performance for datasets annotated with crowdsourcing, additional identifiers of humans who annotated each example are required, which are not available for these two datasets. However, we still believe that our estimates may be useful to some degree, even if they are just an upper bound. For comparison in table 7 we provide estimates of human performance obtained by the same procedure for the DURel [11] LSCD dataset in German, which has a structure similar to RuShiftEval and RuSemShift. This dataset was fully annotated by each of five annotators, thus, our procedure shall estimate human performance correctly for DURel.

method	wordSpear		sentSpear	
	dev1	dev2	dev1	dev2
annot1 VS mean_wo_1	0.941	**0.964**	0.516	0.587
annot2 VS mean_wo_2	0.940	0.940	0.527	0.545
annot3 VS mean_wo_3	0.943	0.951	0.547	0.545
annot4 VS mean_wo_4	0.939	0.941	0.545	0.558
annot5 VS mean_wo_5	**0.961**	0.923	0.524	0.566
mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc} \rightarrow RSS_{MSE}^{dev2-sentSpear}$	0.819	0.811	**0.599**	**0.657**

Table 5: Comparison of our best model with human quality on RuSemShift development set.

method	wordSpear			sentSpear		
	p12	p23	p13	p12	p23	p13
annot1 VS mean_wo_1	0.932	0.953	**0.959**	0.579	0.621	0.628
annot2 VS mean_wo_2	0.920	0.957	0.948	0.578	0.616	0.605
annot3 VS mean_wo_3	**0.936**	**0.958**	**0.959**	**0.597**	0.626	0.616
mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc} \rightarrow RSS_{MSE}^{dev2-sentSpear}$	0.825	0.821	0.823	0.596	**0.634**	**0.631**

Table 6: Comparison of our best model with human quality on RuShiftEval.

method	wordSpear	sentSpear
annot1 VS mean_wo_1	0.875	0.670
annot2 VS mean_wo_2	0.883	0.698
annot3 VS mean_wo_3	0.883	0.678
annot4 VS mean_wo_4	0.939	0.745
annot5 VS mean_wo_5	0.943	0.717

Table 7: Human quality on DURel.

Tables 5 and 6 show that our best model performs ranking of sentence pairs similarly or better than humans. However, word ranking results are significantly worse than those of humans. We suppose that such discrepancy may be due to different biases in mistakes made by humans and our model while scoring sentence pairs. Probably, humans underestimate and overestimate similarity between word occurrences with similar probabilities, so their mistakes are less biased and cancel each other when the average of scores for sentence pairs is calculated to produce word scores. In contrast, as we observed from the error analysis our model in principal does not see differences between some senses. Thus, it consistently overestimates similarity between occurrences of those senses and averaging does not help.

Comparing our estimates of human performance for the Russian datasets and DURel, we observe that humans have better agreement in ranking of sentence pairs (*sentSpear*) on DURel, especially annotators 4 and 5, who were student with a background in historical linguistics. However, our estimates of human performance for word ranking (*wordSpear*) on the Russian datasets is similar or higher than the performance of those two best annotators of DURel. This inderectly supports our hypothesis about overestimation of human performance on the Russian datasets. However, more information about annotators from the crowdsourcing platform is required to draw reliable conclusions.

B Removing short and long sentences

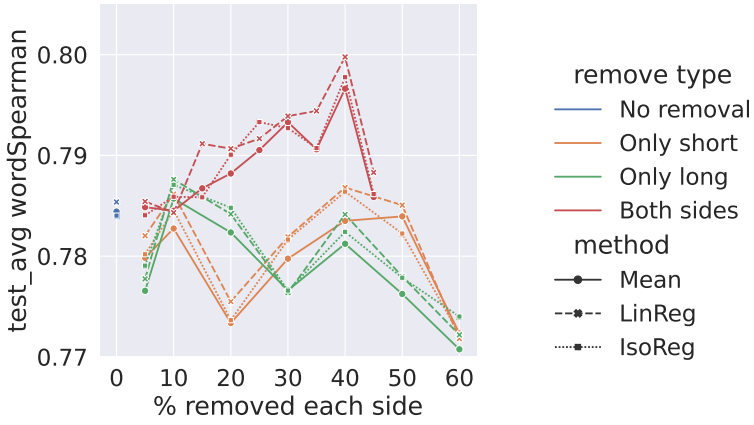


Figure 6: Dependency on percent of removed short and long sentences. Model: first+concat on $\text{MCL}^{en-acc}_{CE} \rightarrow \text{RSS}^{dev2-sentSpear}_{MSE}$

During the competition for each target word we removed 25% of the longest sentences and 25% of the shortest sentences before sampling examples for this word. This was based on our intuition that the WiC model can perform worse on very short or long examples, while thresholds were set arbitrarily. After the competition we decided to study whether removing only shortest or only longest examples is better, and also selecting better threshold. Figure 6 shows the dependence of the competition metric (*wordSpear* on RuShiftEval averaged over three pairs of time periods) on the percent of longest or shortest sentences removed. For symmetric removal the specified percent is removed from each side, resulting in two times more sentences removed. Surprisingly, we observe that removing only shortest or only longest sentences for each word help little. Removing both shortest and longest sentences consistently improve results. The best results are obtained when 40% of the shortest and 40% of the longest sentences are removed for each word, and examples are sampled from only 20% of sentences of medium length.

C More examples of mistakes

Table 8: Samples for mistakes *Model seems to be right* and *Model can not find the difference*

Type	Gold	Model	Pairs of sentences
Model seems to be right. **23.2%**	4 4 3 1 2 2 2 3 4	1 4 1	1) Он, как видно из его стихотворений, взялся за дело поэта по призванию; он сильно сочувствует вопросам своего времени, страдает всеми недугами **века**, болезненно мучится несовершенствами общества и сгорает нетщетно жаждою споспешествовать его совершенствованию и торжеству на земле истины, любви и братства. Во второй половине прошлого **века** рытье колодцев было заменено бурением скважин. 2) Унять было невозможно, по крайней мере в ту минуту, и – вдруг окончательная катастрофа как бомба разразилась над собранием и треснула среди его: третий чтец, тот **маньяк**, который все махал кулаком за кулисами, вдруг выбежал на сцену. Не надо было быть балетным **маньяком**, чтобы понять, что балериной эта особа никогда не будет. 3) Это жертва не моя, а всего **уклада** жизни! Там, верно, рукомойники в сенях, пахнет кухней, мокрые дрова возле печек – убогий, нелюбимый мною, дачный зимний **уклад**.
Model can not find the difference. **39%**	2 1 2 1 1 1 1 2 1 1 1 1 1 4 3	4 4 4 4 4	1) Скоро прибыли к нему **братья** его, Андрей и Борис, с их многочисленною дружиною: не было ни упреков, ни извинений, ни условий; единокровные обнялися с видом искренней любви, чтобы вместе служить отечеству и христианству. Он говорил ей: ”Ты, **брат**”. 2) Злополучный иеромонах был вытащен из **огня** со слабыми признаками жизни. Ночью немцы обрушили на наше расположение массированный артиллерийский **огонь**. 3) В годовщину свадьбы буду выставлять на балконе огненные **цифры**. Сергей Глазьев с **цифрами** в руках наглядно доказал, что идет хищническая добыча нефти и газа, главная цель которой – сверхприбыль любой ценой. 4) Вадима Петровича начинало брать раздражение и на бывшего своего **дядьку**. – Что за ярмарка – Так День Нэзалэжности! – отвечает **дядька** в национальном гуцульском костюме, с приклеенными усами. 5) Команда, за исключением вахтенных, ушла в **увольнение**, в город. Первым плохим признаком стал запущенный в прессу слух, что сразу же после **увольнения** Примакова Рапота сам подал в отставку, а на его место уже подбирается новая кандидатура.

Table 9: Samples of mistakes *Model sees wrong difference* and *Ambiguity*

Type	Gold	Model	Pairs of sentences
Model sees wrong difference. **22.8%**	4 4 4 4 4 4 4 3 4	1 1 1	1) Тут в **кармане** тысяча рублев положена. Вместо птиц он приносил домой целые **карманы** камней и сваливал их под навесом в ящик. 2) Когда Их Величества повернули к той стороне **стены**, которая ведет к Спасским воротам, на площади уже стояла тысячная толпа, приветствовавшая Царя и Царицу восторженными кликами ”ура” и бросаньем в воздух шапок. В конце этого двора у **стены** поставлены бочки, на которые наложены доски. 3) Однажды Петр застал сына в сарае, мальчик пытался пристроить к старому корыту колесо **тачки**. Таня и Освальд не решались сойти с тропы, чтобы не наступить на змею, не потревожить эльфа с крошечной **тачкой**, не поднять из логова сказочного волка.
Ambiguity. **15%**	4 3 4 4 1 4 2 3 1	1 4 4	1) Итак, чтобы покончить с этим предметом, прошу Вас, друг мой, никогда не опасаться задеть мою авторскую **амбицию**. Разговор, конечно, бессмысленный и бесполезный, но наведший все на те же размышления: о Церкви, о Православии, о той мелочной и даже злобной каше интриг, самолюбий, **амбиций**, эгоцентризмов, в которой приходится в Церкви жить. 2) Ребята, с **тачками** туда!.. Скотник, насквозь напитанный запахом навозной жижи, тот самый, что не довез своей **тачки**, а пошел смеяться с мужиками, когда мужики еще добродушно покуривали на крыльцах людской – вертел в руках терракотовую копию химеры: – Ванька, гляди-ка! 3) Понемногу мы привыкали к своему домику и своему новому **укладу**. В странах Европы всегда происходило соприкосновение, а нередко и противоборство, с местной традицией, куда входили и эстетические представления заказчиков, и бытовой **уклад**, и местные ремесленные корпорации.

An Interpretable Approach to Lexical Semantic Change Detection with Lexical Substitution

Arefyev N. V.
Lomonosov Moscow State University
Samsung R&D Institute Russia
HSE University
Moscow, Russia
nick.arefyev@gmail.com

Bykov D. A.
Lomonosov Moscow State University
Moscow, Russia
dima13051998@gmail.com

Abstract

In this paper we propose a new Word Sense Induction (WSI) method and apply it to construct a solution for the RuShiftEval shared task on Lexical Semantic Change Detection (LSCD) for the Russian language. Our WSI algorithm based on lexical substitution achieves state-of-the-art performance for the Russian language on the RUSSE-2018 dataset. However, our LSCD system based on it has shown poor performance in the shared task. We have studied mathematical properties of the COMPARE score employed in the task for measuring the degree of semantic change, as well as the discrepancies between this score and our WSI predictions. We have found that our method can detect those aspects of semantic change, which the COMPARE metric is not sensitive to, such as appearance or disappearance of a rare word sense. An important property of our method is its interpretability, which we exploit to perform the detailed error analysis.

Keywords: Lexical Substitution, Lexical Semantic Change Detection, Word Sense Induction
DOI: 10.28995/2075-7182-2021-20-31-46

Интерпретируемый подход к обнаружению семантических сдвигов с помощью лексических подстановок

Арефьев Н. В.◇ △ ▽
Москва, Россия
nick.arefyev@gmail.com

Быков Д. А. ◇
Москва, Россия
dima13051998@gmail.com

Московский государственный университет имени М. В. Ломоносова ◇
Московский Исследовательский Центр Самсунг △
Национальный исследовательский университет «Высшая школа экономики» ▽

Аннотация

В данной статье мы предлагаем новый метод решения задачи выделения значений слов (WSI) и строим на его основе решение задачи обнаружения семантических сдвигов (LSCD). Наш алгоритм выделения значений слов превосходит по качеству предыдущие методы на датасете RUSSE-2018 для русского язика. Однако наша система решения задачи LSCD показывает низкие результаты на RuShiftEval. Мы изучили математические свойства метрики COMPARE, используемой в данной задаче для оценки степени семантического сдвига, а также расхождения между данной метрикой и предсказаниями нашего метода. Мы обнаружили, что наш метод может определять такие семантические сдвиги, к которым данная метрика не чувствительна, например, появление или исчезновение редкого значения. Другим отличительным свойством нашего метода является интерпретируемость, которую мы использовали при анализе ошибок.

Ключевые слова: Задача обнаружения семантических сдвигов, лексические подстановки, выделение значений слов.

1 Introduction

Lexical semantic change detection (LSCD) is a problem of detecting changes in word meaning over time. Semantic change is a complex phenomenon which is hard to define or formalize. This results in different data annotation schemes, a bunch of complementary metrics, and also different LSCD systems sensitive

to different aspects of the whole phenomenon. The idea behind our approach is first to discover word senses using a Word Sense Induction (WSI) method based on clustering of lexical substitutes. Then we can study the discovered senses and find out if any new sense appeared or an old one disappeared between two time periods. Unlike popular approaches to LSCD that build a single vector for each word in each time period [11], our approach is much more interpretable because it can display examples of each discovered sense, or label corresponding clusters with the words related to those senses. Following this approach, we developed a system for the RuShiftEval [7] competition on LSCD for the Russian language.

Our WSI model is based on the method proposed in [2] for English, which generates lexical substitutes for all occurrences of a particular target word with BERT [6], and then clusters bag-of-word representations of those substitutes. Lexical substitutes are usually different for target word occurrences with different senses and similar for those with the same sense. For instance, the word *mouse* will receive substitutes like *keyboard, monitor* in the phrase *connect the mouse to your PC*, while in the phrase *laboratory mouse* substitutes like *frog, rabbit, monkey* will be generated. We make the base method multilingual by replacing English BERT [6] model with XLM-R [5] model trained on texts in 100 different languages. However, simply replacing BERT with XLM-R to generate substitutes does not work well for the Russian language because the base method generates only single subword substitutes, which are either functional words or word pieces. This results in poor inventory of meaning-bearing substitutes and very poor WSI performance. Thus, we propose a technique to generate multi-subword substitutes, which boosts WSI performance for the Russian language (and likely other languages too, though we did not test it yet). Additionally, we propose and experiment with different Hearst-like patterns specific for the Russian language, and also their combinations. Generating multi-subword substitutes and selecting the best combination of Hearst-like patterns results in new SOTA on the bts-rnc dataset from RUSSE-2018 WSI competition for the Russian language.

During the RuShiftEval competition period we did not manage to achieve good values of the COMPARE metric used in the competition, and also could not outperform our strong but less interpretable baseline based on the orthogonal Procrustes alignment of SGNS vectors, which was among three best performing methods in the SemEval-2020 Task 1 LSCD competition [11], though that competition employed different annotation scheme and metrics. In the post-competition period we performed a mathematical analysis of the COMPARE metric and discovered that unlike our WSI-based method, it is not sensitive to the appearance or disappearance of rare senses, but rather estimate the change in the relative frequency of the most frequent sense. We show that for those words that have gained or lost one of their senses the COMPARE metric can be either higher or lower than for those words that did not undergo any change, which raises a question about the applicability of this metric for LSCD.

Based on our analysis of the COMPARE metric, after the competition we decided to try the same substitution-based distance metric employed for clustering in our WSI method, but skipping the clustering itself to adapt to the properties of the COMPARE metric. This adaptation gave large improvement of the competition metric and outperformed the strong baseline. Finally, we performed error analysis of our WSI-based method. It has revealed that there are actually cases when our method discovers the appearance of a new rare sense, which the COMPARE metric is not sensitive to.

2 Related work

RuShiftEval [7] is the first LSCD shared task for the Russian language. There are three time periods in the RuShiftEval competition (Soviet, pre-Soviet, and post-Soviet). For the evaluation, they were grouped into three pairs of periods. The participants were asked to estimate the semantic shift of each test word between all pairs of periods by providing 3 scores for each word. The data annotation scheme and the evaluation metric follow those proposed as a part of the Diachronic Usage Relatedness (DURel) dataset for German [12]. The performance of the participating systems was evaluated by calculating the Spearman correlation between the predicted word scores and the values of the COMPARE [12] metric. To calculate the COMPARE metric during the dataset annotation, for each word and each pair of time periods several dozens sentence pairs were randomly sampled. In these pairs the first sentence was sampled from the first period and the second sentence from the second period. Each sentence pair was annotated

by humans using 1-4 scale where larger values corresponded to more similar senses of the target word. Finally, annotations of sentence pairs were averaged to produce word scores. The COMPARE metric is known to confuse polysemy and meaning change [12], in this paper we describe other interesting properties of this metric.

An alternative approach to annotation and evaluation was proposed in SemEval-2020 Task 1 [11], where the authors tried to account for the appearance or disappearance of relatively rare word senses, which the COMPARE metric is not sensitive to. They basically clustered word occurrences based on human judgements about similarity of these occurrences by meaning. The transitivity of the same-sense relation is exploited to improve annotation efficiency. Based on the obtained gold standard clustering corresponding to word senses, two subtasks were proposed. The first subtask required binary classification determining if the set of senses of a particular word has changed. The second subtask required ranking words according to the change in relative frequencies of their senses.

Our solution is based on a substitution-based approach to WSI [1, 2, 4] which exploits lexical substitutes to distinguish word senses. Following [2] we used dynamic patterns and develop specific patterns for the Russian language. We developed a lexical substitution model based on the XLM-R multilingual masked language model [5], which is trained on texts in 100 languages. As a baseline we used the LSCD method [10] based on orthogonal Procrustes alignment of SGNS word embeddings [8]. Similar methods based on SGNS embeddings and different alignment techniques were the best performing methods in SemEval-2020 Task 1 [11].

3 Methods of WSI and LSCD

Our LSCD system retrieves examples for the target word from each time period and performs WSI based on lexical substitution for all those examples. The result of WSI is a clustering of word occurrences. Our solution submitted to the competition analyzed the number of examples from each time period in each cluster and made the predictions based on those numbers. The clusters can be labeled with distinctive substitutes. Both these labels and the examples in each cluster can be examined, which makes the proposed method interpretable.

After the competition, we developed a simpler LSCD method inspired by the revealed properties of the evaluation metric. This method employs the same measure of similarity between lexical substitutes that was used for clustering in the WSI algorithm, but skips clustering itself. This significantly improves the evaluation metric, but makes the predictions less intuitive and interpretable.

3.1 Lexical Substitution and Word Sense Induction

Our WSI method employs lexical substitutes generated using the XLM-R masked language model (MLM) [5] to represent word meaning manifested in a particular context. We tried several approaches to generate substitutes. One of them is replacing the target word in the given text fragment with a special token <mask> and asking the MLM to recover it. This results in contextually plausible words, but often they are not related to the target word and do not describe its sense in any way. To solve this problem we used dynamic patterns proposed in [1], which are Hearts-like patterns applied to the inputs of the MLM. For instance, we can replace the target word T with "`T and <mask>`" or "`<mask> and T`", resulting in words that can stand in a co-ordinated row with the target, mostly co-hyponyms of the target word. We compared a number of such patterns and discovered that the best performance is achieved by two symmetric patterns "`T (а также <mask>)`" and "`<mask> (а также T)`" (literally translated as "`T (and also <mask>)`" and "`<mask> (and also T)`"), which were both selected for our final solution.

However, the XLM-R vocabulary contains only functional and very frequent Russian words, most content words are split into several subwords. We found that single-subword substitutes alone result in poor performance of WSI for the Russian language and extended the base approach by generating multi-subword substitutes. To achieve this, we insert several <mask> tokens in the pattern ("`T and <mask><mask><mask>`", for instance). Then we generate substitutes from left to right using beam search, i.e. we take K most probable subwords predicted by XLM-R for the first `<mask>` token, for each

of them predict K most probable continuations resulting in K^2 sequences of two subwords filling two <mask> tokens, and from them leave only K most probable sequences, etc. Unlike traditional language models, a masked language model is not an autoregressive model. Thus, the described procedure is not a mathematically sound way to deal with the distribution learnt by the model, though it gives empirically good results. Some examples are presented in table 1. For our final solution we generated substitutes consisting of 1 to 3 subwords.

Next, in order to represent the lexical meaning of the target word and not its grammatical form, the generated substitutes are lemmatized. For each lemma the sum of the predicted probabilities of all of its forms and across two symmetric patterns with different number of masks is calculated. Finally, K substitutes with the highest sum of probabilities are taken for each example. Following [2], we build TF-IDF bag-of-words vectors of the generated substitutes and clustered these vectors using agglomerative clustering with cosine distance and complete linkage. For WSI experiments in section 4.1, the number of clusters is either selected for each word individually using the silhouette score similarly to [4], or is selected the same for all words based on training set metrics. For LSCD we iterated over several values to obtain different clusterings (see section 3.2). Finally, for the qualitative analysis in section 4.4 only the silhouette score was used.

cluster ID	sentence	top 10 substitutes
0	дополнительных вертикально расположенных **винта (а также <mask><mask>)**. С целью обеспечения устойчивости вертолета плоскости вращения несущих винтов были немного наклонены внутрь, этим достигался эффект, аналогичный эффекту поперечного V крыла	**крыла**: 2e-2 **колеса**: 9e-3 **двигателя**: 9e-3 опоры: 8e-3 **хвоста**: 7e-3 пружины: 6e-3 **других элементов**: 6e-3 **привода**: 5e-3 диска: 5e-3 двух дополнительных: 5e-3
0	настоящее время разворачивается его серийное производство. Основными особенностями машины является отсутствие хвостового **винта (а также <mask><mask>)** и модульная компоновка. Преимущества соосной схемы обеспечили вертолету следующие потребительские свойства: – высокую	**двигателя**: 5e-2 **крыла**: 1e-2 **привода**: 1e-2 **колеса**: 1e-2 амортизатора: 1e-2 **хвоста**: 8e-3 кузова: 8e-3 его отсутствие: 7e-3 тормозов: 6e-3 **других элементов**: 6e-3
1	после десяти часов и садились закусывать. Одни ужинали, другие играли в скромные **винт (а также <mask><mask>)** и преферанс, третьи проигрывались в «железку» и штрафами покрывали огромные расходы Кружка.	покер: 1e-2 **рулетку**: 9e-3 **в карты**: 9e-3 карточные: 7e-3 лото: 5e-3 тайм: 5e-3 спиннинг: 4e-3 **теннис**: 4e-3 экспресс: 3e-3 шашки: 3e-3
1	к своему сослуживцу Шешковскому, у которого каждый день собирались чиновники играть в **винт (а также <mask><mask>)** и пить холодное пиво. «Своею нерешительностью я напоминаю Гамлета. – думал Лаевский	**в карты**: 8e-2 шахматы: 8e-2 мяч: 2e-2 **рулетку**: 2e-2 курить: 1e-2 стрелять: 9e-3 **теннис**: 8e-3 хоккей: 8e-3 шутить: 7e-3 другие игры: 6e-3

Table 1: An example of inputs to the XLM-R model, generated substitutes and clusters assigned.

3.2 Semantic Change Detection

Our WSI method can be directly used to decide that a new sense appeared or an old disappeared between two time periods if there are clusters containing only examples from one of the periods. However, for the RuShiftEval competition the words had to be ranked according to some scores reflecting the strength of semantic change. To calculate the score of semantic change for a particular word and two time periods, we sampled A sentences containing this word from each time period and clustered all $2A$ sentences several times using our WSI method with different number of substitutes and different number of clusters [1]. Then from the clustered sentences we randomly generated B pairs of sentences with the first sentence from an old period, and the second from a new one. Each sentence was sampled from the corresponding A sentences with replacement. Then the final score was calculated as the average across all clusterings of the indicators that both sentences fell into the same cluster. The intuition behind this method is the following.

[1] Our best submitted solution iterated over [50,60,70,80,90,100,200,300] substitutes and [2,3] clusters.

If two word occurrences are identical in meaning, they are likely to be grouped together in many different clusterings. However, if the meanings are different, they can be grouped together by chance or mistake, but hopefully not too often. For the best submitted solution we set $A = 60$ and $B = 120$.

3.3 Cluster labeling

During WSI the clustering algorithm groups together those target word occurrences that have similar substitutes. Thus, we can find those substitutes that are specific to each cluster. Following [2], we calculate the Pointwise Mutual Information (PMI) for a given substitute and a given cluster. To label each cluster, we select substitutes having the maximum PMI value with that cluster. Despite many noisy substitutes appearing in such labels, they are still useful as compact representations of clusters. Thus, we employ these labels for the qualitative analysis.

3.4 Cosine Similarity Averaging

After the competition we developed a new algorithm using the same distances between substitute vectors, but skipping the clustering procedure. For each word we sampled $A = 80$ examples from each period. For each pair of periods and each word we generated $B = 120$ pairs of sentences in which the first sentence was from an earlier period and the second sentence was from a later period. Then, for each pair of sentences we calculated the cosine distance between substitute vectors. Finally, for each target word the average distance was calculated, producing the predicted word scores.

3.5 SGNS+OP+CD

As a baseline in our study we used SGNS vectors with orthogonal Procrustes alignment and cosine distance proposed in [10]. We got the best results with the following hyperparameters: windowsize=5, k=5, t=0.001, dimensions=300, minCount=0, iters=5.

4 Results and Discussion

4.1 Experiments with Word Sense Induction

Table 2 compares our WSI method to the previous best results during and after RUSSE-2018 competition on bts-rnc dataset [9] using the official metric of that competition, which is the Adjusted Rand Index (ARI), and official evaluation scripts. All experiments and hyperparameter selection were performed on the training set, while the final results in Table 2 are reported on the public and the private parts of the test set.

model	public test ARI	private test ARI
RUSSE-2018 results		
AdaGram baseline [9]	0.262	0.213
RUSSE-2018 best result [9]	0.351	0.338
RUSSE-2018 2nd best result [3]	0.281	0.281
other published results		
RuBERT semantic fingerprint [13]	0.21$^{\triangle}$	
bayes-comb-silnc [4]	0.502	0.451
bayes-comb-fixnc [4]	0.464	0.438
our results		
andalso-2subwords-silnc	0.525	0.507
andalso-2subwords-fixnc	**0.564**	**0.573**

Table 2: Comparison with the published results on the RUSSE-2018 bts-rnc dataset. Selecting the number of clusters for each word using silhouette score (silnc) or as a hyperparameter on train (fixnc). $^{\triangle}$The subset of bts-rnc used to evaluate this method is not specified in [13].

To find optimal hyperparameters of the lexical substitution model, we employed maxARI metric introduced in [4], which is the maximum possible ARI achieved by agglomerative clustering with different number of clusters. First, we compare Hearst-like patterns that potentially can give good substitutes for the Russian language. In Figure 1, we consider patterns with only one `<mask>` token, which is to the left from the target word. We selected the best performing patterns `"<mask> и т"` (`"<mask>`

and T"), "<mask> или T" ("<mask> or T"), "<mask> (а также T)" ("<mask> (and also T)"), "<mask> (в том числе T)" ("<mask> (including T)").

For the selected patterns we considered two possible orders of the target word and the <mask> token. After the competition we also experimented with symmetric combinations obtained by multiplying probability distributions over possible substitutes. This is equivalent to leaving only substitutes that are probable in both patterns. Figure 1 shows that this symmetric combinations consistently improve the results for all patterns, especially when multi-subword substitutes are generated. Interestingly, putting <mask> to the right of the target, like in "T and <mask>", works bad when generating multi-subword substitutes because multi-word expressions such as *other varieties*, *in other places* are often generated, which are not very related to any particular word senses. Also we notice that generating two-subword substitutes significantly improve the performance for symmetric combinations, while generating three-subword substitutes seems to be an overkill.

Finally, Table 2 compares our WSI method using two-subword substitutes and symmetric combination of "X (а также X)" ("X (and also X)") to the best previous results on the bts-rnc dataset, where it achieves the new state-of-the-art performance. The application of dynamic patterns sometimes produces ungrammatical sentences. In Appendix D we provide some preliminary analysis of this problem.

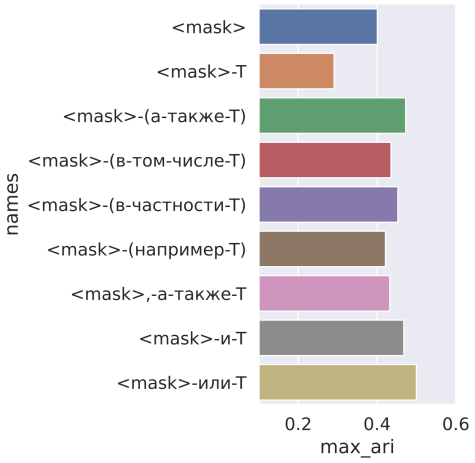


Figure 1: Comparison of different patterns with a single <mask> token.

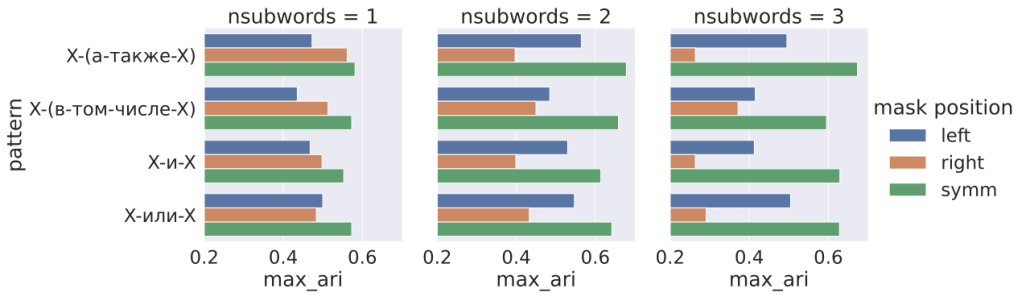


Figure 2: Comparison of patterns with different number and positions of <mask> tokens.

4.2 LSCD Results for RuShiftEval

In the RuShiftEval the systems were ranked by the Spearman correlation between their scores and the values of the COMPARE metric for each word. This metric is the average of annotator predictions for a set of sentence pairs, where the first sentence refers to an earlier time period, and the second to a later one. Unfortunately, during the evaluation period we did not have the cosine averaging approach yet. The results in Table 3 show that the cosine averaging method gives word scores that are much better correlated

with the COMPARE metric compared to the submitted WSI-based method and SGNS+OP+CD baseline. If submitted, the cosine averaging method could give us 4th place among all teams.

methods	pre-Soviet:Soviet	Soviet:post-Soviet	pre-Soviet:post-Soviet	average
best competition results				
GlossReader	0.781	0.803	0.822	0.802
DeepMistake	0.798	0.773	0.803	0.791
vanyatko	0.678	0.746	0.737	0.720
our submissions				
WSI based approach	0.274	0.202	0.307	0.261
SGNS + OP + CD	0.2690	0.2240	0.4230	0.3050
our post-evaluation results				
Cosine Similarity Averaging	0.671	0.677	0.658	0.66

Table 3: LSCD results on RuShiftEval, Spearman's rank correlation.

4.3 The COMPARE metric analysis

In this section we describe the properties of the COMPARE metric explaining what kind of semantic changes does it reflect and what kind of changes it is not sensitive to. To simplify our analysis, we will discuss only cases when a word has unrelated senses and each pair of its occurrences receive annotations of 1 (entirely different senses) or 4 (same sense).

We start with a simple example. Suppose there is a word (a homonym) having two unrelated senses in the old corpus with probabilities of 0.9 for sense1 and 0.1 for sense2. First, suppose the least frequent sense disappeared in the new corpus and only the most frequent is left. Then the second sentence sampled from the new corpus always has the first sense of the target word. Thus, we sample a pair of sentences with the same word meaning and receive the score 4 from annotators with probability of 0.9, while sampling unrelated word meanings and obtaining the score 1 has the probability of 0.1 (Table 4 shows all possible outcomes and their probabilities for this and the following cases). Thus, the expected value of the COMPARE metric is $4 * 0.9 + 1 * 0.1 = 3.7$. Now suppose that the most frequent sense disappeared. The same calculations result in the COMPARE metric of $4 * 0.1 + 1 * 0.9 = 1.3$. Finally, suppose nothing changed, then the COMPARE metric will be $4 * (0.81 + 0.01) + 1 * (0.09 + 0.09) = 3.46$. Thus, if a sense has disappeared (or appeared, as the calculations are symmetric), the COMPARE metric can be both near its maximum of 4 or its minimum of 1 depending on the sense frequencies. And when nothing changed, the result is in between. Thus, the metric does not reflect appearance or disappearance of word senses, but rather frequency distribution of those senses.

outcome			**outcome probability**		
sentence1	**sentence2**	**annotation**	*only sense1 left*	*only sense2 left*	*nothing changed*
sense1	sense1	4	0.9	0	0.81
sense2	sense2	4	0	0.1	0.01
sense1	sense2	1	0	0.9	0.09
sense2	sense1	1	0.1	0	0.09
			COMPARE=3.7	**COMPARE=1.3**	**COMPARE=3.46**

Table 4: Possible outcomes, their probabilities and corresponding annotations for our two sense example.

Now let us study the metric in a more general case. Suppose there are several unrelated senses of a word with probabilities $p = (p_1, ..., p_n)$ in the old corpus and $q = (q_1, .., q_n)$ in the new one. For convenience, let us order senses such that $p_1 > p_2 > ... > p_n$. The COMPARE metric is the average over annotations for sampled sentences. Thus, it is an estimate of the expected value of annotation for a pair of sentences sampled from uniform distributions over the old and the new corpus, which is the sum of probabilities of sampling the same sense[2] or different senses[3] of the target word from two corpora multiplied by the corresponding annotations (4 or 1):

[2] which is the probabilities of sampling i-th sense from both corpora summed over all senses

[3] it can be simply calculated as the probability of NOT sampling the same sense

$$\underset{s_1 \sim p, s_2 \sim q}{E} annot(s_1, s_2) = 4 * \sum_i p_i q_i + 1 * (1 - \sum_i p_i q_i) = 3 * \sum_i p_i q_i + 1$$

Since q_i are probabilities, this sum can be treated as a weighted average of p_i with weights q_i. This weighted average achieves maximum when $q_1 = 1$ (only the most frequent sense left) and minimum when $q_n = 1$ (only the least frequent sense survived). When nothing has changed, it is somewhere in between.

This analysis suggests that a method that predicts high semantic change scores for those words which obtained or lost a rare sense will have low Spearman correlation with the COMPARE metric simply due to the shown properties. At the same time, a method that is sensitive only to the change in the frequency of the most frequent sense may be considered as a good LSCD method by this metric.

4.4 Qualitative and error analysis

In this section we study the discrepancies between the results of our WSI algorithm and the COMPARE metric values. For each word we sampled 80 examples from the pre-Soviet period and 80 from the post-Soviet period from the Russian National Corpus. From the test set we selected two groups of words, including 50 words with COMPARE > 3 into the first group and 12 words with COMPARE < 2 into the second group. According to the COMPARE metric, the second group shall contain words with strong semantic shift while the first group shall not.

We applied our WSI algorithm to all of these words. We predicted that a word has acquired a new sense according to our algorithm if there was a cluster containing at most two examples from the pre-Soviet period and at least 4 examples from the post-Soviet period (such bounds compensate for mistakes of the clustering algorithm). From the first group (high COMPARE) we took 10 words that acquired a new sense according to our predictions, while from the second group (low COMPARE) we took 10 words which presumably did not acquired or lost any senses[4] (Table 5). Those words have the largest discrepancy between our predictions and the values of the COMPARE metric in the gold standard.

first group			**second group**		
word	**COMPARE**	**discrepancy type**	**word**	**COMPARE**	**discrepancy type**
бригада	3.08	correct new sense (OK)	дух	1.88	a lot of senses (OK)
жесть	3.41	incorrect new sense	наложение	1.78	a lot of senses (OK)
обоснование	3.58	incorrect new sense	полоса	1.41	all senses in one cluster
ранец	3.38	incorrect new sense	роспись	1.57	changed MFS (OK)
сверстник	3.82	incorrect new sense	ссылка	1.93	changed MFS (OK)
список	3.05	incorrect new sense	тачка	1.89	all senses in one cluster
стол	3.25	incorrect new sense	хрен	1.6	changed MFS (OK)
тупик	3.14	incorrect new sense	центр	1.87	a lot of senses (OK)
увольнение	3.32	senses are combined	ядро	1.47	changed MFS (OK)
углеводород	3.2	correct new sense (OK)	ясли	1.9	a lot of senses (OK)

Table 5: Discrepancies between predictions of our method and the COMPARE metric. Discrepancies that are not errors of our method are marked with (OK).

Appendix A shows clusters for the words from the first group, which obtained a new sense according to our predictions. The first type of discrepancies (**correct new sense**), are those words, that really obtained a new sense at least in our sampled subset of RNC, but were not detected by the COMPARE metric. There are two words, *бригада (military brigade / team of workers)* and *углеводород (hydrocarbon as a class of organic chemical compounds / oil and natural gas as economic resources)*, for which our predictions seem to be correct. In Appendix C random examples containing these words from two time periods and the assigned clusters are shown. For the first word *бригада* WSI found the most frequent sense associated with the army, which was overwhelming in the pre-Soviet corpus, and the second sense

[4]Sense frequencies still might have changed, it can be easily detected from cluster sizes if required. However, it is an open question whether LSCD methods shall detect only words that acquired or lost some sense, or words with relative frequencies of senses changed also. To make our qualitative analysis more simple and clear, we selected the first option to retrieve words for analysis.

(a team of workers), which was found in the post-Soviet corpus only and is still relatively rare compared to the first one. For the second word, WSI found the correct new sense associated with oil and natural gas as economic resources, and an existing one associated with chemical compounds. Almost all examples for the ”oil and natural gas” sense are from the post-Soviet corpus, the value of COMPARE metric are relatively high probably due to the existing sense still dominating. The second type of discrepancies (**incorrect new sense**) are the words *жесть (tinplate), ранец (satchel), стол (table), обоснование (justification), тупик (deadlock), список (list), сверстник (coeval)*. Our method divided some senses into more than one cluster. Thus, the prediction is wrong. The third type (**senses are combined**) consists of one word *увольнение*, the senses of ”vacation” and ”dismissal” were merged into a single cluster, but a few examples of the sense ”dismissal” were put into a separate cluster.

In the second group (Appendix B), there are words with low COMPARE metric for which our method has not found any new or lost senses. We found three types of such words. The first type (**all senses in one cluster**) consists of the words *полоса (line / time interval / forest belt), тачка (car / cart)*, for which our method mistakenly found only one frequent sense not counting outliers. For the rest of the words, the clustering results seem to be correct. Moreover, despite no senses were obtained or lost, the observed distribution of sense frequencies explains low values of COMPARE metric. For the words *ядро (nucleus / cannonball), ссылка (deportation / reference), хрен (horseradish / man), роспись (list / painting / signature)*, the most frequent sense changed (**changed MFS**). The words *наложение (imposition / overlay), дух (spirit / ambience / ghost), центр (downtown / middle), ясли (manger / nursery / crib)* are highly polysemous, thus, receive low COMPARE values, which is a known problem for COMPARE metric [12].

The results of our error analysis are shown in Table 5. We have found 10 words out of 20, for which our method correctly predicted that the set of word senses has changed or has not changed. The discrepancy with the COMPARE metric for these words can be explained by the fact that this metric reflects the change in sense frequencies rather than the appearance or disappearance of senses, and also gives low scores to highly polysemous words. Indeed, according to the analysis of the clusters built by our WSI algorithm there was a significant change in relative frequency of the most frequent senses of several words that have not obtained or lost senses. Also we have found highly polysemous words with low values of the COMPARE metric. For two words with relatively high COMPARE values the appearance of a new sense was correctly predicted, but this sense has not become the most frequent sense, which explains the discrepancy. The most frequent error of our method is splitting a single sense into multiple clusters, which we will try to overcome in the future work.

5 Conclusion

We proposed a WSI method achieving new state-of-the-art for the Russian language on the bts-rnc dataset. Also we proposed two approaches for solving LSCD task. The first approach is based on our WSI method. This approach has an important advantage of being interpretable. One can look at the clusters that contain sentences from only one time period and understand which sense appeared or disappeared, or if this is just a mistake of the algorithm. Our second method is less interpretable, but it achieves higher results according to the competition metric.

References

[1] Asaf Amrami and Yoav Goldberg. Word Sense Induction with Neural biLM and Symmetric Patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, 2018.

[2] Asaf Amrami and Yoav Goldberg. Towards better substitution-based word sense induction. *arXiv e-prints*, pages arXiv–1905, 2019.

[3] Nikolay Arefyev, Pavel Ermolaev, and Alexander Panchenko. How much does a word weigh? weighting word embeddings for word sense induction. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, pages 68–84, Moscow, Russia, 2018. RSUH.

[4] Nikolay Arefyev, Boris Sheludko, and Tatiana Aleksashina. Combining Neural Language Models for Word Sense Induction. In *Analysis of Images, Social Networks and Texts*, page 105–121. Springer International Publishing, 2019.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019.

[7] Andrey Kutuzov and Lidia Pivovarova. RuShiftEval: a shared task on semantic shift detection for Russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*, 2021.

[8] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[9] Alexander Panchenko, Anastasia Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. 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"*, pages 547–564, Moscow, Russia, 2018. RSUH.

[10] Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, 2019.

[11] Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[12] Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[13] Aleksandr Slapoguzov, Konstantin Malyuga, and Evgenij Tsopa. Word Sense Induction for Russian Texts Using BERT. In *Conference of Open Innovations Association, FRUCT*, number 28, pages 621–627. FRUCT Oy, 2021.

Appendix A Words with high COMPARE and a new sense predicted

In Figure 3 we show 10 words from RuShiftEval dataset with COMPARE > 3, for which our method predicted that a new sense appeared in the post-Soviet time period, i.e. returned a cluster containing at least 4 examples from the post-Soviet period and at most 2 examples from the pre-Soviet period.

Words		Cluster names	OLD COUNT	NEW COUNT
бригада	0	2, армия, батальон, батарея, войско, два, дивизия, дивизя	79	61
	1	служба, машина, они, человек, полиция, организация, я, весь	1	19
жесть	0	страх, я, лицо, глаз, форма, силой, красота, кровь	13	22
	1	бетон, бронза, глина, железо, из, картон, лист, масло	65	53
	2	любовь, красота, сила	2	5
обоснование	0	построение, изложение, практика, толкование, понимание, выражение	30	21
	1	создание, разработка, осуществление, проектирование, использование, реализация	11	18
	2	аргумент, факт, в, другой, причина, решение, аргументация, доказательство	6	17
	3	проектирование, расчёт, разработка, другой, содержание, решение	0	9
	4	признание, основа, оправдание, подтверждение, смысл, установление	29	12
	5	результат, другой, основа, выражение, содержание, основание	4	3
ранец	0	вода, в, друг, другой, я, дом, она, предмет	2	7
	1	автомат, без, велосипед, весь, вещий, глаз, голова, деньга	66	72
сверстник	0	враг, коллега, мой, не, отец, соратник, ученик, родные	78	56
	1	самый, взрослый, старший, с, школьник, весь, семья, ребёнок	2	24
список	0	письмо, документ, текст, дать, предложение, книга, лист, фамилия	41	17
	1	это, память, остальной, в, лист, журнал, рейтинг, резюме	13	22
	2	остальной, рейтинг, состав, они, по, программа, этот, другой	1	4
	3	описание, по, номер, название, перечень, некоторый, портрет, фото	24	35
	4	программа, он, свой, реестр, число, текст, в, отчёт	1	2
стол	0	блюдо, стакан, к, другой, ребёнок, в, он, стол	2	8
	1	внутри, голова, дверь, книга, комната, кровать, кухня, мебель	77	68
	2	другой, на, и, к, в, перед, зал, кресло	1	4
тупик	0	дом, угол, дорога, конец, ряд, один, в, обратно	14	27
	1	паника, затруднение, сомнение, ужас, шок, точка, хаос, наоборот	61	22
	2	-, ., 3, </s>, автомат, велосипед, ключ, пистолет	1	14
	3	кризис, опасность, проблема, положение, он, ситуация, колея, только	4	17
увольнение	0	арест, болезнь, выход, должность, заключение, заявление, измена, изменение	79	75
	1	в, на, работа, отпуск, освободить, после, я, он	1	5
углеводород	0	водород, воздух, кислота, окисел, он, пара, раствор, с	78	42
	1	ресурс, сырьё, природный, нефтяной, минеральный, топливо, и, нефтепродукт	2	38

Figure 3: Words with high COMPARE that acquired a new sense according to our predictions

Appendix B Words with low COMPARE and no new senses predicted

In Figure 4 we show 10 words from RuShiftEval dataset with COMPARE < 2, for which our method did not find any new senses in the post-Soviet period compared to the pre-Soviet one, i.e. there is no clusters with at least 4 examples from the post-Soviet period and at most 2 examples from the pre-Soviet period.

Figure 4: Words with low COMPARE and no new senses predicted.

Appendix C Examples of sentences for the words *бригада* and *углеводород*

Tables 6, 7 provide examples of sentences for two words with high COMPARE metric for which our method found a new sense. Each table contains 10 randomly sampled sentences from the pre-Soviet and post-Soviet time periods. From these examples we conclude that our algorithm correctly detected that a new sense appeared at least in the sampled examples. This sense is surely related to the old sense, but in our opinion it is not equal to it.

cluster ID	sentence
Pre-Soviet	
0	Нигде у нас нет резерва , и государь с **бригадою** пехоты стоит почти на аванпостах в средине пустого пространства (верст с 50 до 60) , разделяющего войска тырновские (Николай Николаевич) от войск наследника .
0	Батальоны образцовые , но командир **бригады** генерал-майор Эллис так прирос к гвардейским порядкам , что не может свыкнуться с мыслью , что он не в Красном Селе : за два часа перед смотром и накануне весьма продолжительно производил репетиции прохождением церемониальным маршем с музыкою и хоровыми солдатскими ответами на начальнические приветствия .
0	На правом берегу Днепра действовала и II-ая конная армия под командой бывшего войскового старшины Миронова , в составе 2-ой , 16-ой и 21-ой кавалерийских дивизий и особой конной **бригады** .
1	В Скорцару и Янке : 7-я и 8-я легкая батареи 15-й артиллерийской **бригады**.
0	В поле не потеряется и возложенную на него задачу выполнит с успехом ; к подчиненным строго требователен , беспристрастен и справедлив ... ” и так много , и потом заключение : “ Достоин быть командиром неотдельной **бригады** ” .
0	На Кавказе , поступив в Куринский полк , я застал Барятинского уже полковником , командовавшим одним из батальонов Кабардинского полка , который вместе с Куринским составлял **бригаду** , и хотя полки эти стояли в разных местностях , но их частям , кроме экспедиций , нередко приходилось встречаться в крепости Грозной , где была штаб-квартира командиров бригады и дивизии , причем последний был в то же время и начальником левого фланга Кавказского корпуса , и когда вместо генерала Фрейтаг [а] прибыл в Грозную начальником левого фланга генерал Нестеров , то , как знакомый ранее во Владикавказе с его семейством , я посещал его каждый раз по приходе в Грозную колонн Куринского полка из крепости Воздвиженской , причем мне приходилось встречаться в доме Нестерова и с князем Барятинским .
0	Перед фронтом генерала Бабиева были обнаружены все части II-ой конной армии Миронова (2-ая , 16-ая и 21-ая кавалерийские дивизии и отдельная кавалерийская **бригада**) .
0	Декабря 4 я назначен был командовать бригадою резервных рот 2 артиллерийской дивизии , и 23 артиллерийской бригады , я принял бригаду генваря 1829 г . 1829 год .
0	Командиры полков второй **бригады** так же , как и сам Вульферт , редко бывали с нами , потому что они шли сзади , на один переход , и являлись к Скобелеву , только когда догоняли нас на дневках .
0	Подлинный подписал : Начальник Главного штаба , артиллерии генерал-майор Ермолов Приказ по 1-й Западной армии 88 25 августа 1812 года Главная квартира В позиции при Бородине По воле главнокомандующего : Артиллерии генерал-майору Костенецкому поручается в командование **бригада** состоящая из Либавского и Софийского полков .
Post-Soviet	
0	В годы службы полковника в **бригаде** подчиненные не раз докладывали ему о том , что “ солдатское радио ” сообщало о потайных лазах в фундаменте гарнизонного клуба , через которые неустановленные военнослужащие якобы ходили в “ самоволку ” .
0	Короче , пока Витек пил в обкомовской гостинице , пытаясь забыть девку , сбежавшую из профессорской семьи на строительство Камского гиганта , старшеклассницу , изуродованную передовой **бригадой** маляров коммунистического труда , онемевшую от боли , когда ее сунули руками в известь и держали так , потому что медленно работала , и бригаду из-за нее лишили премии и переходящего из рук в руки , как крановщица Кларка , вымпела .
0	Приезжала следственная **бригада** .
0	В ВВС и ПВО включены также три авиабазы , зенитная ракетная **бригада** , другие части и подразделения .
1	В тот год огромный их дом , похожий на бастион , был точно бы взят приступом и покорился орде нагловатых , спешащих строительных **бригад** .
1	**Бригада** Усмана выезжала отныне только на подмогу обывателям при серьезных “ разборках ” , “ стрелках ” или “ наездах ” .
0	3-й (Румелийский) корпус : Пехота : 7 линейных полков (21 батальон) , 7 стрелковых батальонов , Боснийская **бригада** (6 батальонов) , Греческий волонтерский пограничный полк (3 батальона) , Боснийский волонтерский пограничный полк (4 батальона) , Никшичский албанский волонтерский батальон , Герцеговинский пограничный батальон .
1	За животными ухаживает целая **бригада** женщин , одна из которых – жена Майка , Джанин .
0	Не исключено , что в соседской квартире **бригада** выполняла отдельные работы .
1	Англичане специальную **бригаду** в Забайкалье прислать собираются , чтобы все тщательно подсчитать и внести читинский “ Локомотив ” в Книгу рекордов Гиннесса .

Table 6: Examples of sentences with the word *бригада*

cluster ID	sentence
Pre-Soviet	
0	– По изложении взгляда на изомерию ароматических **углеводородов** , там поясняется , “ что и в формуле бензола нет необходимости допущения реального существования остатков ” (стр . 23) .
0	Из 8 г диметилангеликалактона получено около 2 г перегона , состоявшего из тиофенового **углеводорода** и тиофенового фенола .
0	Теперь я хотел бы особенно подчеркнуть , что такое различие единиц сродства атома углерода даже не во всех случаях необходимо для объяснения изомерии предельных **углеводородов** , если последняя действительно имеет место .
0	Лучшему выходу препятствует образование буро-красной смолы и **углеводородов** , которые всегда происходят во время реакции йодистого бутила на цианистую соль , и также образование бутиламина при действии соляной кислоты на продукт .
0	В кометных хвостах , которые , как известно , состоят преимущественно из газообразных **углеводородов** , мы имеем дело с отдельными молекулами , радиус которых r108 см , а плотность d10 , как показал F. Exner ; в этом случае , однако , наша формула (5) неприменима во всей строгости , так как отдельные молекулы не суть абсолютно черные тела и радиус их мал сравнительно с длиной волны падающего на них света ; поэтому мы можем только утверждать , что отталкивание хвостов , во много раз превышающее притяжение их , притом различное для различных веществ хвоста и обратно пропорциональное квадрату расстояния от Солнца , не противоречит нашей формуле .
0	Грэм нашел в воздухе двух копей в Нью-Кестле 82,5-94,5 % легких **углеводородов** , 4,5-16,5 % азота и 1,0-1,3 % кислорода .
0	В первый раз было взято на 2 г **углеводорода** 5 г марганцовокислого калия , 8 г едкого натра и 1000 мл воды .
0	Такое превращение имеет место для эфилена и для бутилена де Люиня ; сюда же принадлежат и описанные отношения бутилена , открытого мною , но есть , однако , пример , что из **углеводорода** получается алкоголь , не тожественный , а изомерный с первоначальным ; таковы отношения обыкновенного амильного алкоголя , амилена и амиленгидрата .
0	В обоих случаях мы должны искать причину изомерии кислот в изомерии аллиленов или двуобромленных пропиленов ; но если принять для пропилена химическое строение , выраженное формулой , то углеродные атомы этого **углеводорода** не находятся все в одинаковых химических условиях , как это имеет место для углеродных атомов этилена .
0	Для превращения этого **углеводорода** в триметилкарбинол можно , однакоже , как мною указано уже и прежде , употребить серную кислоту , разведенную до некоторой степени водой (около 5 ч. кислоты на 1 ч. воды) .
Post-Soviet	
0	Нефтедобыча на водосборной площади р. Вах , служащей водоисточником г. Нижневартовска , стала причиной повышения предельно допустимого уровня содержания нефтяных **углеводородов** и тяжелых металлов в речной и питьевой воде [20] .
1	Суркова и А.А . Трофимука (1994) более 85 % запасов нефти в бассейне связаны с баженовско-неокомской нефтяной генерационно-аккумуляционной системой , характеризующейся превосходным по потенциалу источником **углеводородов** – баженовской кремнисто-глинистой толщей верхнеюрского возраста и морскими песчаниками неокомского возраста с высокими коллекторскими свойствами .
0	На второй стадии осуществляется очистка ВМС от остатков эмульсий жидких **углеводородов** в водной фазе .
1	По прогнозу Хьюджеса , пик разработки сланцевых **углеводородов** придется на 2017 год , после чего начнется падение , в результате которого за два года добыча упадет до уровня 2012-го .
1	В результате в России сейчас практически нет даже сколько-нибудь устоявшихся данных по запасам и ресурсам таких **углеводородов** , которые не вызывали бы сомнений и споров у экспертов (а такие цифры есть , например , по США) .
0	Так , проведенными ранее работами в различных климатических зонах было показано , что при температуре , близкой к оптимальной , утилизация **углеводородов** достигала 90-100 % всего за несколько дней (table) .
1	В условиях низких цен на **углеводороды** инструменты выживания ищут все .
0	В промышленных выбросах содержатся смертельно ядовитые вещества : окись углерода , двуокись азота , **углеводороды** .
0	В качестве первичного акта формирования залежи рассматривалось спонтанное газовыделение из растворов в области их закритического метастабильного насыщения низкокипящими **углеводородами** (прежде всего , метаном) .
1	– Судя по всему , Азербайджан обречен быть нашим главным конкурентом на рынке **углеводородов** .

Table 7: Examples of sentences with the word *углеводород*

Appendix D Ungrammatical sentences after application of dynamic patterns

After application of dynamic patterns some sentences may become ungrammatical. This may affect the quality of generated substitutes. In this section we describe some preliminary study of this problem. We consider the disagreement in number between a verb and its subject, which is one type of grammatical errors that can be the result of application of dynamic patterns. We selected 3 real examples from the

RUSSE-2018 bts-rnc dataset and also made up one simple sentence ourselves. Each example has a singular verb and a singular subject. After one of the dynamic patterns is applied to the singular subject, it becomes plural and the agreement is broken. In tables 8, 9, 10 we show 10 most probable substitutes (after combining substitutes consisting of 1, 2 and 3 subwords) generated for this broken sentence and for the same sentence fixed manually.

Table 8 shows examples for the pattern "`Т и <mask>`". For our made-up example and the first real example the substitutes are syntactically plausible words, but mostly not related to the target word. After fixing the agreement more co-hyponyms appear among substitutes, which is better for representing the meaning of the target word. For the last two examples there are no significant changes in the substitutes after fixing the agreement. For the pattern "`Т или <mask>`" (table 9) substitutes differ less for correct and incorrect versions of sentences and seem to describe the meaning equally well. The pattern "`<mask> (а также Т)`" (Table 10) seems to be the most robust to incorrect agreement.

The inspected examples show that in some cases substitutes generated for a grammatical and ungrammatical versions of the same sentence can be significantly different. It seems that this effect depends on sentence complexity and dynamic pattern used. We plan to explore this effect in more details in our future work.

sentence	substitutes
Мальчик и <mask> ел шоколадное мороженное.	не: 0.51 сам: 0.12 никогда не: 0.08 с удовольствием: 0.06 вовсе не: 0.04 так: 0.03 правда: 0.03 больше не: 0.02 не только: 0.02 в детстве: 0.02
Мальчик и <mask> ели шоколадное мороженное.	девочка: 0.33 девушка: 0.26 его друзья: 0.23 мама: 0.10 женщина: 0.04 мать: 0.04 сестра: 0.03 его мама вместе: 0.03 она: 0.03 мальчик: 0.02
Пантюхин в Склифе сейчас. Он выползти на улицу успел, а на Золоткова **балка и <mask>** обрушилась. Эх, душой компании парень был! 28-летний Геннадий так и не	вовсе: 0.53 вовсе не: 0.09 крыша: 0.09 не: 0.07 так: 0.04 сама: 0.04 гора: 0.04 на него: 0.03 она: 0.02 без того: 0.02
Пантюхин в Склифе сейчас. Он выползти на улицу успел, а на Золоткова **балка и <mask>** обрушились. Эх, душой компании парень был! 28-летний Геннадий так и не	крыша: 0.13 гора: 0.12 лес: 0.06 стена: 0.05 дом: 0.05 все: 0.04 дерево: 0.03 город: 0.03 снег: 0.02 стены: 0.02
дворце в Летнем саду. Впоследствии ее передали в Академию наук. Первая книжная **лавка и <mask>** была открыта в Гостином дворе на Петербургской стороне. Здесь продавали печатные указы	магазин: 0.40 библиотека: 0.25 аптека: 0.07 книгарня: 0.05 книжный магазин: 0.05 печатница: 0.05 рынок: 0.03 магазин книг: 0.02 столовая: 0.02 библиотеки: 0.02
дворце в Летнем саду. Впоследствии ее передали в Академию наук. Первая книжная **лавка и <mask>** были открыты в Гостином дворе на Петербургской стороне. Здесь продавали печатные указы	магазин: 0.52 библиотека: 0.16 книжный магазин: 0.12 аптека: 0.10 рынок: 0.06 печатница: 0.05 первый магазин: 0.03 библиотеки: 0.03 магазины: 0.03 книгарня: 0.03
, нападения, грабежи Самара, 30, VIII. В селе Быковом Остроге ограблена казенная винная **лавка и <mask>** . Похищено казенных денег 756 р., 12 ведер водки и собственных денег торговца	магазин: 0.44 торговец: 0.18 аптека: 0.06 склад: 0.05 торговый дом: 0.05 банк: 0.04 ресторан: 0.03 дом: 0.03 другой магазин: 0.02 другое имущество: 0.01
, нападения, грабежи Самара, 30, VIII. В селе Быковом Остроге ограблены казенная винная **лавка и <mask>**. Похищено казенных денег 756 р., 12 ведер водки и собственных денег торговца	магазин: 0.45 торговец: 0.15 торговый дом: 0.09 аптека: 0.07 дом: 0.06 банк : 0.05 другой магазин: 0.05 склад: 0.04 ресторан: 0.02 почта: 0.02

Table 8: Examples with pattern "`Т и <mask>`"

sentence	substitutes
Мальчик или <mask> ел шоколадное мороженное.	девочка, который: 0.19 девочка не: 0.14 мальчик: 0.13 мальчика: 0.07 не: 0.07 девушка: 0.05 он: 0.04 женщина: 0.04 как: 0.03 я: 0.03
Мальчик или <mask> ели шоколадное мороженное.	мы: 0.35 девочка,: 0.13 девочка, которые: 0.09 Мы: 0.09 вы: 0.08 мальчики: 0.04 девушки: 0.03 Вы: 0.03 дети: 0.03 они: 0.02
Пантюхин в Склифе сейчас. Он выползти на улицу успел, а на Золоткова **балка или <mask>** обрушилась. Эх, душой компании парень был! 28-летний Геннадий так и не	гора: 0.22 что: 0.06 что - то: 0.05 крыша: 0.05 дерево: 0.03 же: 0.03 дом: 0.02 камень: 0.02 как там: 0.02 стена: 0.02
Пантюхин в Склифе сейчас. Он выползти на улицу успел, а на Золоткова **балка или <mask>** обрушились. Эх, душой компании парень был! 28-летний Геннадий так и не	гора: 0.15 стена: 0.13 дом: 0.08 крыша: 0.06 мост: 0.04 дерево: 0.03 стены: 0.03 лес: 0.02 камни: 0.02 камень: 0.02
дворце в Летнем саду. Впоследствии ее передали в Академию наук. Первая книжная **лавка или <mask>** была открыта в Гостином дворе на Петербургской стороне. Здесь продавали печатные указы	книжный магазин: 0.13 книжница: 0.12 рынок: 0.08 книгарня: 0.05 аптека: 0.04 салон: 0.02 базар: 0.02 книгария: 0.02 Библиотека: 0.01 библиотеко: 0.01
дворце в Летнем саду. Впоследствии ее передали в Академию наук. Первая книжная **лавка или <mask>** были открыты в Гостином дворе на Петербургской стороне. Здесь продавали печатные указы	книжный магазин: 0.18 библиотека: 0.15 лавки: 0.08 книжница: 0.05 рынок: 0.04 аптека: 0.04 книгарня: 0.04 библиотеки: 0.03 Библиотека: 0.02 первый магазин: 0.01
, нападения, грабежи Самара, 30, VIII. В селе Быковом Остроге ограблена казенная винная **лавка или <mask>**. Похищено казенных денег 756 р., 12 ведер водки и собственных денег торговца	магазин: 0.29 склад: 0.08 дом: 0.03 заведение: 0.02 аптека: 0.02 ресторан: 0.02 торговец: 0.02 фабрика: 0.02 бар: 0.02 лавки: 0.01
, нападения, грабежи Самара, 30, VIII. В селе Быковом Остроге ограблены казенная винная **лавка или <mask>** . Похищено казенных денег 756 р., 12 ведер водки и собственных денег торговца	магазин: 0.31 склад: 0.05 дом: 0.05 аптека: 0.04 торговец: 0.03 ресторан: 0.03 торговый дом: 0.03 банк: 0.02 другой магазин: 0.02 магазин водки: 0.01

Table 9: Examples with pattern "`Т или <mask>`"

sentence	substitutes
Мальчик (а также <mask>) ел шоколадное мороженное.	мама: 0.18 его родители: 0.18 я: 0.08 папа: 0.07 девушка: 0.06 родители: 0.06 его бабушка: 0.05 отец: 0.04 мать: 0.03 мальчик: 0.03
Мальчик (а также <mask>) ели шоколадное мороженное.	его родители: 0.31 мама: 0.16 родители: 0.14 девушка: 0.09 его бабушка: 0.06 мать: 0.03 мальчик: 0.03 дети: 0.03 папа: 0.03 я: 0.02
Пантюхин в Склифе сейчас. Он выползти на улицу успел, а на Золоткова **балка (а также <mask>)** обрушилась. Эх, душой компании парень был! 28-летний Геннадий так и не	на него: 0.11 она: 0.04 машина: 0.03 дом: 0.03 крыша: 0.02 он: 0.02 я: 0.02 дверь: 0.02 дома: 0.02 его: 0.02
Пантюхин в Склифе сейчас. Он выползти на улицу успел, а на Золоткова **балка (а также <mask>)** обрушились. Эх, душой компании парень был! 28-летний Геннадий так и не	крыша: 0.04 другие: 0.03 дома: 0.03 люди: 0.03 на него: 0.03 дом: 0.03 остальные: 0.02 город: 0.02 она: 0.02 он: 0.02
дворце в Летнем саду. Впоследствии ее передали в Академию наук. Первая книжная **лавка (а также <mask>)** была открыта в Гостином дворе на Петербургской стороне. Здесь продавали печатные указы	библиотека: 0.35 магазин: 0.19 аптека: 0.15 рынок: 0.06 книжный магазин: 0.05 печатная: 0.03 книгарня: 0.03 книжница: 0.03 музей: 0.02 Библиотека: 0.02
дворце в Летнем саду. Впоследствии ее передали в Академию наук. Первая книжная **лавка (а также <mask>)** были открыты в Гостином дворе на Петербургской стороне. Здесь продавали печатные указы	библиотека: 0.25 аптека: 0.20 магазин: 0.12 другие: 0.07 рынок: 0.03 книжный магазин: 0.03 музей: 0.03 магазины: 0.03 другая: 0.02 библиотеки: 0.02
, нападения, грабежи Самара, 30, VIII. В селе Быковом Остроге ограблена казенная винная **лавка (а также <mask>)** . Похищено казенных денег 756 р., 12 ведер водки и собственных денег торговца	магазин: 0.26 банк: 0.07 дом: 0.05 аптека: 0.05 другой магазин: 0.04 другие: 0.03 две другие: 0.03 почта: 0.03 банки: 0.03 деньги: 0.02
, нападения, грабежи Самара, 30, VIII. В селе Быковом Остроге ограблены казенная винная **лавка (а также <mask>)** . Похищено казенных денег 756 р., 12 ведер водки и собственных денег торговца	магазин: 0.26 дом: 0.07 банк: 0.07 другой магазин: 0.06 аптека: 0.05 другие: 0.04 две другие: 0.04 банки: 0.04 почта: 0.03 торговец: 0.02

Table 10: Examples with pattern "`Т (а также <mask>)`"

Near-duplicate handwritten document detection without text recognition

Oleg Bakhteev
Antiplagiat, Dorodnicyn CC FRS CSC RAS
Moscow, Russia
bakhteev@ap-team.ru

Rita Kuznetsova
MIPT
Moscow, Russia
rita.kuznetsova@phystech.edu

Andrey Khazov
Antiplagiat
Moscow, Russia
khazov@ap-team.ru

Aleksandr Ogaltsov
Antiplagiat
Moscow, Russia
ogaltsov@ap-team.ru

Kamil Safin
MIPT
Moscow, Russia
kamil.safin@phystech.edu

Tatyana Gorlenko
Antiplagiat
Moscow, Russia
gorlenko@ap-team.ru

Marina Suvorova
Antiplagiat
Moscow, Russia
suvorova@ap-team.ru

Andrey Ivahnenko
Antiplagiat
Moscow, Russia
ivahnenko@ap-team.ru

Pavel Botov
Antiplagiat
Moscow, Russia
botov@ap-team.ru

Yury Chekhovich
Antiplagiat, Dorodnicyn CC FRS CSC RAS
Moscow, Russia
chehovich@ap-team.ru

Vadim Mottl
Dorodnicyn CC FRS CSC RAS
Moscow, Russia

Abstract

The paper presents a novel method for near-duplicate detection in handwritten document collections of school essays. A large amount of online resources with available academic essays currently makes it possible to cheat and reuse them during high school final exams. Despite the importance of the problem, at the moment there is no automatic method for near-duplicate detection for handwritten documents, such as school essays. The school essay is represented as a sequence of scanned images of handwritten essay text. Despite advances in recognition of handwritten printed text, the use of these methods for the current task is a challenge. The proposed method of near-duplicate detection does not require detailed markup text, which makes it possible to use it in a large number of tasks related to the information extraction in zero-shot regime, i.e. without any specific resources written in the processed language. The paper presents a method based on series analysis. The image is segmented into words. The text is characterized by a sequence of features, which are invariant to the author's writing style: normalized lengths of the segmented words. These features can be used for both handwritten and machine-readable texts. The computational experiment is conducted on IAM dataset of English handwritten texts and the dataset of real images of handwritten school essays.

Keywords: handwritten text analysis, near-duplicate detection, word segmentation, time series analysis
DOI: 10.28995/2075-7182-2021-20-47-57

Поиск почти дубликатов рукописных текстов без распознавания текста

Бахтеев О. Ю.
Антиплагиат, ФИЦ ИУ РАН
Москва, Россия
bakhteev@ap-team.ru

Кузнецова Р. В.
МФТИ
Москва, Россия
rita.kuznetsova@phystech.edu

Хазов А. В.
Антиплагиат
Москва, Россия
khazov@ap-team.ru

Огальцов А. В.
Антиплагиат
Москва, Россия
ogaltsov@ap-team.ru

Сафин К. Ф.
МФТИ
Москва, Россия
kamil.safin@phystech.edu

Горленко Т. А.
Антиплагиат
Москва, Россия
gorlenko@ap-team.ru

Суворова М. А.
Антиплагиат
Москва, Россия
suvorova@ap-team.ru

Ивахненко А. А.
Антиплагиат
Москва, Россия
ivahnenko@ap-team.ru

Ботов П. В.
Антиплагиат
Москва, Россия
botov@ap-team.ru

Чехович Ю. В.
Антиплагиат, ФИЦ ИУ РАН
Москва, Россия
chehovich@ap-team.ru

Моттль В. В.
ФИЦ ИУ РАН
Москва, Россия

Аннотация

Рассматривается задача поиска почти-дубликатов в коллекции сканированных изображений школьных сочинениях. Сочинение представляется набором изображений рукописного текста, написанного автором. Актуальность задачи обусловлена наличием больших библиотек школьных сочинений, которые могут использоваться школьниками в качестве источника заимствования при написании собственного сочинения. На текущий момент не существует автоматических методов анализа сочинений на наличие заимствований. Несмотря на успехи в области распознавания рукописного текста,применение данных методов для рассмотренной задачи затруднительно. Для решения задачи предлагается рассматривать текст, находящийся в изображении, как последовательность. Предлагается метод, заключающийся в сегментации слов в изображении. Текст характеризуется последовательностью признаков, полученных на основе сегментации. В качестве такого признака выступает нормализованная длина извлеченных из изображения слов. Полученные статистики являются инвариантными по отношению к почерку автора, а также могут использоваться как для рукописных, так и для машиночитаемых текстов. Предложенный метод поиска почти-дубликатов не требует наличия аннотированных корпусов изображений, и потому может быть применим для низкоресурсных языков. Для подтверждения работоспособности метода проводятся эксперименты на англоязычном корпусе IAM, а также выборке реальных изображений рукописных текстов школьных сочинений.

Ключевые слова: рукописные изображения, поиск почти-дубликатов, сегментация слов, анализ временных рядов

1 Introduction

The paper is devoted to the analysis of academic essays and textual reuse detection in them. We consider the problem on the example of school essays written during high school final exams in Russia. A standardized system of assessment for the essay makes it possible to reuse some text or parts of the texts from open collections of school essays available on the Internet. We refer to the problem as near-duplicate detection, but not plagiarism detection problem because the proposed method is robust to slight changes in compared documents. Also, near-duplicate detection is a more precise formulation since plagiarism is a fact that is approved by experts after a detailed analysis of near-duplicate passages.

The main feature that makes the problem hard to analyze is Russian cursive which is really variable in terms of styles of writing letters and connecting them with each other. This is a known feature, but even now datasets for Russian handwritten text recognition are proprietary and not available for the public. Therefore we can't use state-of-the-art handwritten recognition algorithms due to the lack of datasets to train on. Since word recognition is a crucial component for subsequent text reuse detection it is not possible to obtain decent detection quality.

Our contributions are:

- we propose a novel method that avoids the stage of word recognition and directly applicable to the image of the analyzed document;
- we present the dataset of real handwritten school essays and baseline for the text reuse detection task without word recognition;
- we compare the performance of our method with state-of-the-art recognition-based algorithm using IAM dataset for handwritten text recognition in English.

2 Related Work

The problem of finding sources of textual reuse in academic essays is a challenge and can be considered as critical for the educational system [13, 8, 19]. Despite the probably massive nature of the problem, at the moment there is no automatic method of text reuse detection in school essays. The closest work in this area [15] involves an automated system for collecting and analysing essays written in English. The methods described in it are not directly applicable to our problem since the student works considered in [15] are written using printed letters, which are simpler to analyse. The method to compare two document images is presented in [5]. It is based on a similarity measure on the top of word bounding boxes vector representations that are obtained by convolution neural network. The authors pointed out the problem of low data resources, but they deal with it by generating synthetic data and perform transfer learning on IAM dataset. In contrast, we have a slightly different task of comparison handwritten documents with a collection on properly printed documents. Also, we propose not to generate additional data, but perform in zero-shot manner without any learning.

The major works in the area of handwritten text analysis are based on the text recognition methods [3, 15, 18, 12]. Currently, the methods based on deep learning achieve rather good performance on the handwritten recognition task [3, 18], which potentially makes it possible to use it with a combination of modern plagiarism detection systems [2]. The main disadvantage of such methods is the requirement for the presence of markup: to optimize the parameters of recognition models, a significant corpus of annotated texts is required. Therefore this method is not applicable if the documents are written in the language with a lack of such markup. An example of such language is Russian: despite the amount of works devoted to handwritten text and distinct characters recognition [6, 11], to the best of our knowledge currently there is no available publicly annotated corpus for the Russian language.

This paper presents a simple yet efficient method for near-duplicate school essay detection. The method is based on the word segmentation with further analysis of word lengths extracted from the texts. The word segmentation is a well-studied problem, which can be conducted much simpler than word recognition and basically does not require any markup, therefore the proposed method can be applied as a zero-shot method for languages without any annotated corpus for recognition model training. We analyse the lengths of words extracted from the texts and empirically show that they can be considered as features for the information retrieval algorithms invariant to the author's writing style. For similar text detection we employ different methods of time series and sequence alignment [4, 14]. The computational experiment is conducted on two datasets of scanned images: IAM dataset of handwritten texts [9] and the real dataset of the images of handwritten school essays.

3 Problem statement

Consider the problem of near-duplicate detection as an information retrieval problem. Given a dataset of school essays, which are represented by scanned images of handwritten text:

$$D_{\text{susp}} = \{d^i_{\text{susp}}\}.$$

There is also given a collection of documents $D = \{d^j\}$, which can be represented both as scanned images or text in the machine-readable format. We suppose that for each essay $d^i_{\text{susp}} \in D_{\text{susp}}$ there is only one document in collection, which was employed as a source of text reuse:

$$g : D_{\text{susp}} \to D.$$

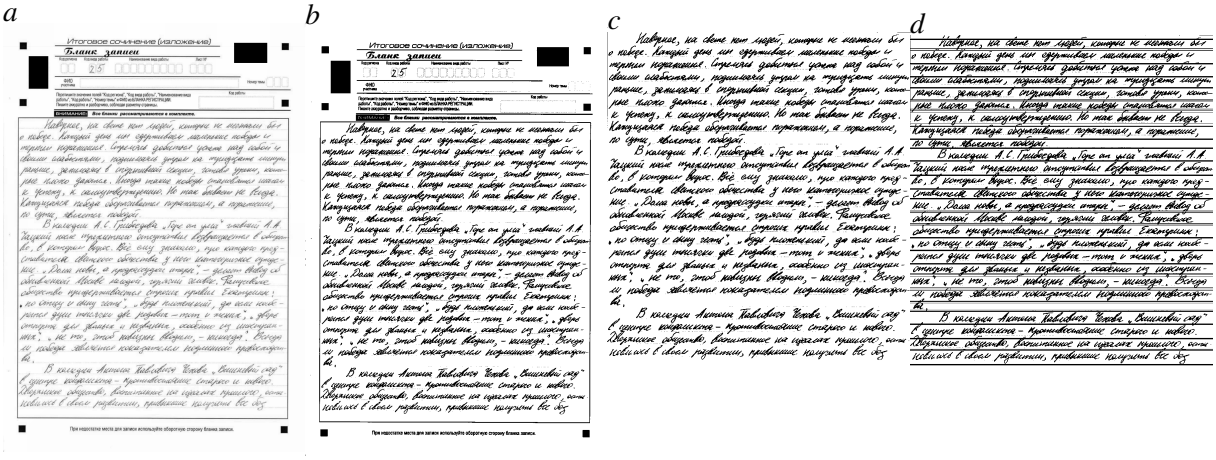


Figure 1: School essay page example: *a)* original image; *b)* binarized image; *c)* text area segmentation; *d)* line segmentation.

The major quality criterion for this task is Recall@K maximization, where Recall@K is a ratio of relevant documents in the most similar K documents retrieved by our method:

$$\text{Recall@}K = \frac{1}{|D_{susp}|} \sum_{d^i_{susp}} |f(d^i_{susp})@K \cap \{g(d^i_{susp})\}|, \tag{1}$$

where f is a document retrieval model, $f(d^i_{susp})@K$ is a set of top-K documents the most similar to the document d^i_{susp}.

After the model found a probable text reuse source for the suspicious document, the source should be verified by the expert. In practice the expert can analyse only a small number of retrieved documents, therefore the formal optimization task is to find a mapping, that maximizes Recall@1 for our dataset:

$$\hat{f} = \arg\max_{f \in \mathcal{F}} (\text{Recall@1}(f, g, D, D_{\text{susp}})),$$

where $\mathcal{F}$ is a family of considered retrieval models.

3.1 Near-duplicate detection using word segmentation

The proposed method is based on the considering the text as a series of features [10]. We propose to segment the document into words without its further recognition. The extracted words are further transformed into features. In this paper we use only the word length as such feature, however other features, such as word height or number of ligatures can potentially improve current near-duplicate detection quality. Opposite to challenging text recognition problem these features are rather simple to extract from the handwritten document.

The school essay d^i_{susp} is represented as a sequence of scanned images of handwritten text. The essay form is standardized and has clear ruled lines, therefore the problem of line segmentation can be solved using a rule-based line segmentation algorithm. The image preprocessing consists of image binarization, text area extraction and line segmentation. The example of preprocessing steps application is shown in Figure 1.

The further image analysis step is word segmentation. For this problem, we use the method based on connectivity component analysis [7]. To take into account word cursive during word length analysis we use a deslatning algorithm similar to [17]. This step is significant especially if the essay author uses significant letter tilt and also helps to segment words more accurately. After that, we extract connectivity components and determine the thresholds for spaces between words and between characters. Since these

thresholds depend on the author writing style we determine them dynamically using a Gaussian mixture with 2 components:

$$s \sim \alpha\mathcal{N}(m_1, s_1) + (1 - \alpha)\mathcal{N}(m_2, s_2),$$

where s is a distance between connectivity components, $\alpha \in (0, 1)$. We suppose that the component with a smaller mean corresponds to the space distance between characters in words. We unite the connectivity components with distance between that is more likely to correspond to this component. The example of word segmentation is shown in Figure 2. The list of numbers presented in Figure 2.e is a list of lengths in pixels of the bounding boxes of the extracted words. We normalize them by dividing by the average box length extracted from the text. The resulting sequence of normalized word length is shown in Figure 2.f.

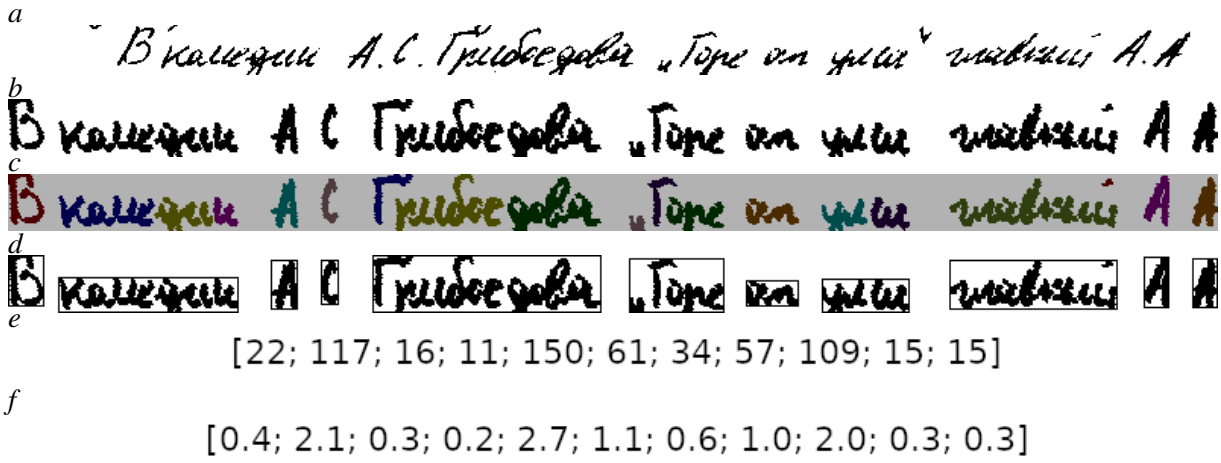


Figure 2: Word segmentation example: *a)* original image line; *b)* deslanting; *c)* connectivity component extraction; *d)* word segmentation result; *e)* word length extraction; *f)* word length normalizing.

After the word segmentation for each image we obtain a sequence of lengths extracted from the image. We normalize this sequence by average extracted word length and consider it as a feature that characterizes the essay. For the essay comparison we employ functions based on the dynamic time warping method [4]:

$$\text{DTW}(x^1, x^2) = \text{dtw}_{t_1,t_2},$$

$$\text{dtw}_{i,j} = ||x_i^1 - x_j^2||_2 + \min(\text{dtw}_{i,j-1}, \text{dtw}_{i-1,j-1}, \text{dtw}_{i-1,j}),$$

where x^1, x^2 are the sequences of lengths t_1 and t_2 correspondingly.

The computational complexity of DTW which is $O(t_1 \cdot t_2)$. In this paper we employed DTW function and its modification FastDTW [14], which has linear computational complexity. Although these methods are well-known for our knowledge there is no research of usage such representation for near-duplicates detection of handwritten texts. We are inspired by the work [10], which shows that considering text as time series and subsequent outlier detection is a fruitful approach to the problem of intrinsic plagiarism detection.

4 Experiment

In order to demonstrate the performance of the proposed method we conducted computational experiments with two image datasets: IAM dataset [9] and a dataset of real school essays [1]. The brief statistics about the used dataset is represented in Table 1. For better experiment reproducibility we used two datasets as a document collection D for the Russian language: a collection of essays mined from the Internet and Taiga corpus [16]. To the best of our knowledge currently there is no available publicly available

[1]The dataset is available at http://bit.ly/ap_handwritten

Table 1: Statistic about the used dataset

Suspicious documents, D_{susp}			
Dataset	Language	Document number	Average word number
IAM [9]	English	336	76
School essays	Russian	89	263
Document collection, D			
Dataset	Language	Document number	Average word number
IAM [9]	English	992	75
School essays	Russian	17361	503
Taiga[16]	Russian	15197	287

dataset annotated for handwritten text recognition, therefore we used IAM dataset for comparison with the text recognition-based model.

As a quality criteria for both experiments we used recall function (1): Recall@1, Recall@10 and Recall@100.

Experiment on IAM dataset. The dataset consists of handwritten images of text segmented into lines. Each line has an annotation file with information about each word in line. The dataset is split into *Train, Test* and *Validation* parts. We used the images from the *Test* split as a set of suspicious documents $D_{\text{susp}}, |D_{\text{susp}}| = 336$ and all the text documents from the dataset as a collection of documents $D, |D| = 992$. We used *Train* part of dataset to tune hyperparameters of the proposed algorithm. Some of the images of the dataset contains identical texts written by different authors. We did not use this information and considered all the images as independent objects.

This dataset was used to compare the proposed method with text recognition-based models. For the comparison we used a model from [3], a neural network-based model achieving state-of-the-art results on multiple handwritten text recognition datasets. We trained the model with different percentages of the images from *Train* subset: $\{10\%, 20\%, 50\%, 100\%\}$. The performance of these models is presented in Table 2. For each percentage we ran the training procedure 5 times for 1000 epochs, the results were averaged.

We evaluated the word segmentation algorithm used in the proposed method using the methodology described in [1]. For the used word segmentation algorithm we got *Precision*=0.8, *Recall*=0.7, F_1=0.75, which is quite comparable to other word extraction algorithms.

As a distance function between the documents we used a cosine distance between the collection document and text extracted from the image:

$$\cos(\mathbf{v}_1, \mathbf{v}_2) = \frac{\langle \mathbf{v}_1, \mathbf{v}_2 \rangle}{||\mathbf{v}_1||_2 \cdot ||\mathbf{v}_2||_2},$$

where $\mathbf{v}_1, \mathbf{v}_2$ are the bag-of-words vectors of the texts of the compared documents.

For the methods based on series analysis we filtered one-character words from the document collection D. We found that this heuristic gives a significant performance improve since the large amount of "a" words in texts lowers the chances to correctly align short texts. The results for the experiment are shown in Table 2. The *Ground Truth Word lengths* method corresponds to the application of series analysis to the word lengths from the dataset annotation. The results for this method show a performance that can be potentially achieved if the word segmentation method works perfectly without any error. The results show that the performance of the text recognition-based model dramatically decreases with size of the training dataset. As we can see, the proposed method performance is comparable with state-of-the-art recognition method that is trained on half of the dataset, however the proposed algorithm achieves comparable performance in zero-shot manner. It can be used for languages with a lack of ground-truth data for handwritten word recognition which is actually very frequent case.

Table 2: Word error rates (WER) for the recognition-based models

Method	**WER**
[3], 10% of the dataset used for training	0.921 ± 0.001
[3], 20% of the dataset used for training	0.836 ± 0.010
[3], 50% of the dataset used for training	0.546 ± 0.027
[3], 100% of the dataset used for training	0.187 ± 0.000

Table 3: Experiment results for the IAM dataset.

Method	**Recall@1**	**Recall@10**	**Recall@100**
[3], 10% of the dataset used for training	0.00 ± 0.00	0.04 ± 0.01	0.23 ± 0.03
[3], 20% of the dataset used for training	0.02 ± 0.01	0.15 ± 0.04	0.47 ± 0.04
[3], 50% of the dataset used for training	0.74 ± 0.05	0.89 ± 0.03	0.98 ± 0.01
[3], 100% of the dataset used for training	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Proposed, DTW	0.66	0.78	0.89
Proposed, FastDTW	0.55	0.69	0.82
Ground Truth Word lengths, DTW	0.97	0.99	0.99
Ground Truth Word lengths, FastDTW	0.92	0.95	0.97

Experiment on the dataset of real handwritten school essays. The dataset of suspicious essays D_{susp} consists of 89 images of school essays. For each image we have the corresponding text without per-word annotation. We split the dataset into two parts: 18 images for *Train* part and 71 images for *Test* part. *Train* part was used to tune hyperparameters of the proposed algorithm.

As a collection of texts D we used two different datasets. The first dataset is a dataset of school essays mined from the Internet. The dataset consists of 17361 documents. In order to increase the reproducibility of the experiment we also used the second dataset, which was constructed as a subset of Taiga corpus [16]. We used a subset of *proza.ru* texts included in this corpus. We used only texts from the year 2009, which length is similar to typical essay length: from 150 to 400 words. The final collection size was 15197 texts. Both the datasets does not contain the real sources of the suspicious documents D_{susp}. For each document d^i_{susp} we also added into collection D the real source of the document, thus during the experiment there is only 1 real source document in the collection D.

The results for the experiment are shown in the Table 4, Table 5. As we can see the proposed method gives rather good results for both collections. For the Taiga corpus we also estimated the time for one school essay processing. All the experiments were run on the computer with 16 GB RAM and Intel Core i5 CPU. For both the experiments we used only one core. As we can see, FastDTW performs significantly faster, however, the quality of the proposed method with DTW is better. One of the further directions in the development of the proposed method is the combination of these functions in order to obtain a trade-off between the quality and speed of the method.

We also analyzed the dependence of the proposed method on the essay length and its similarity to the original texts. Firstly, we conducted an experiment truncating all the analyzed sequences, extracted from the collection D and suspicious documents D_{susp}. We considered different truncation percentage: from 10% to 90%. For the experiment DTW function was used. The document collection D is Taiga corpus. The results are shown in Figure 3. As we can see, the proposed method works poorly on the small texts, which also can explain the difference in performance on the IAM dataset and the dataset of the school essays: the average essay length is much longer than the average document from the IAM dataset, 65 words after removing short words in IAM dataset versus 257 words in essays.

Secondly, we analyzed the performance of the proposed method for the case, when the original text and suspicious document are partially different. For this experiment instead of including into the document

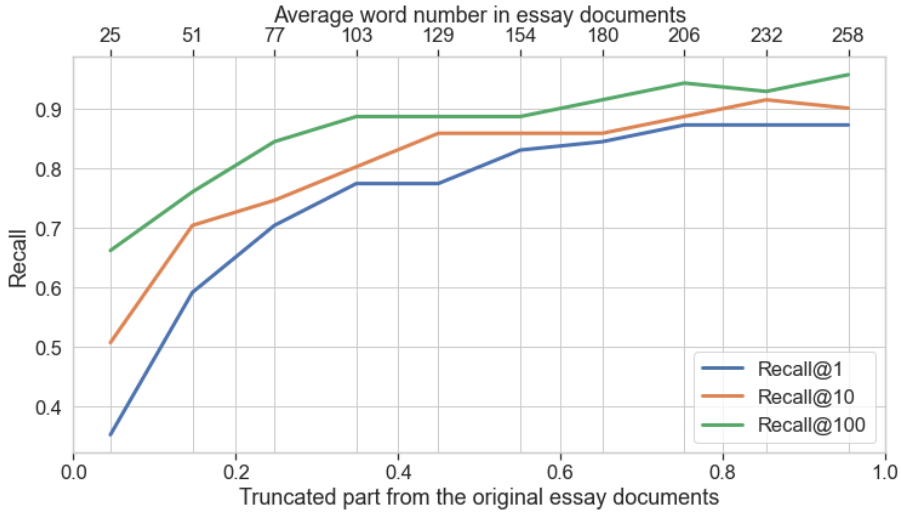


Figure 3: The dependency of the performance of the proposed method on the analyzed sequences lengths.

Table 4: Experiment results for the the dataset of real handwritten school essays using documents mined from the Internet as a document collection D.

Method	**Recall@1**	**Recall@10**	**Recall@100**
Proposed, DTW	0.93	0.99	1.0
Proposed, FastDTW	0.80	0.91	0.99

collection D the original essay text we randomly mixed it with another document from the collection. We considered different percentage p of the original essay text for this procedure. More formally, we conducted the following steps:

1. select the original essay text, extract word length sequence from this text;
2. randomly select subsequence of the sequence with $p\%$ of the original sequence;
3. randomly select document d from the collection D, extract word length sequence from this text;
4. randomly select subsequence of the collection document text series with $(100 - p)\%$ of the original sequence;
5. randomly insert the subsequence of the essay text into the subsequence of the collection document;
6. add the resulting subsequence into the series of the collection D;
7. remove the sequence of the document d from the collection D.

This algorithm simulates the situation, when the text was copied from the origin partially, with $p\%$ of text reuse. For this experiment we mix the original text with one of the documents d from the collection D, therefore the number of ground-truth source documents increases. We believe that this differs from a real-world setting, when the student often copies the text only from one origin. Therefore we remove the series of the document d from the collection D in order to have only one ground-truth origin for each essay. We considered different mixture percentage: from 70% to 90%. The experiment was run 5 times, the results were averaged. As for the previous experiment, we used DTW function and Taiga corpus for the document collection D. The results are shown in Figure 4

For further analysis we collected 25 essay images that use one text as a source of reuse. We built an alignment matrix [4] between them to demonstrate the proposed method operability. The matrix for these texts is shown in Figure 5a. In comparison, we also built alignment matrices between these essay images and random school essay texts. The result is shown in Figure 5b. The alignment matrix for the essay image and true text reuse source is strongly diagonal while the matrix between random texts does not demonstrate this matrix property.

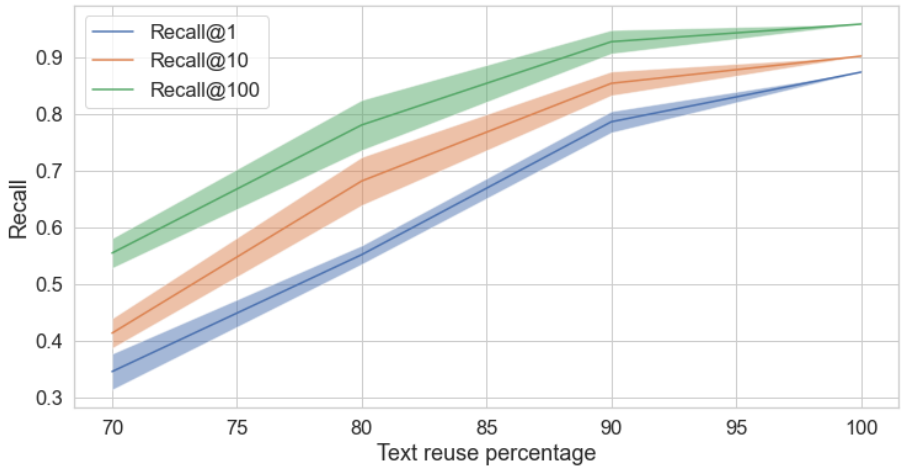


Figure 4: The dependency of the performance of the proposed method on the similarity between suspicious document and the original collection document. The results are averaged between different experiment runs.

Table 5: Experiment results for the the dataset of real handwritten school essays using subset of Taiga corpus as a document collection D.

Method	**Recall@1**	**Recall@10**	**Recall@100**	**Time per one essay, sec**
Proposed, DTW	0.87	0.90	0.96	73.4 ± 13.1
Proposed, FastDTW	0.66	0.70	0.79	3.9 ± 0.2

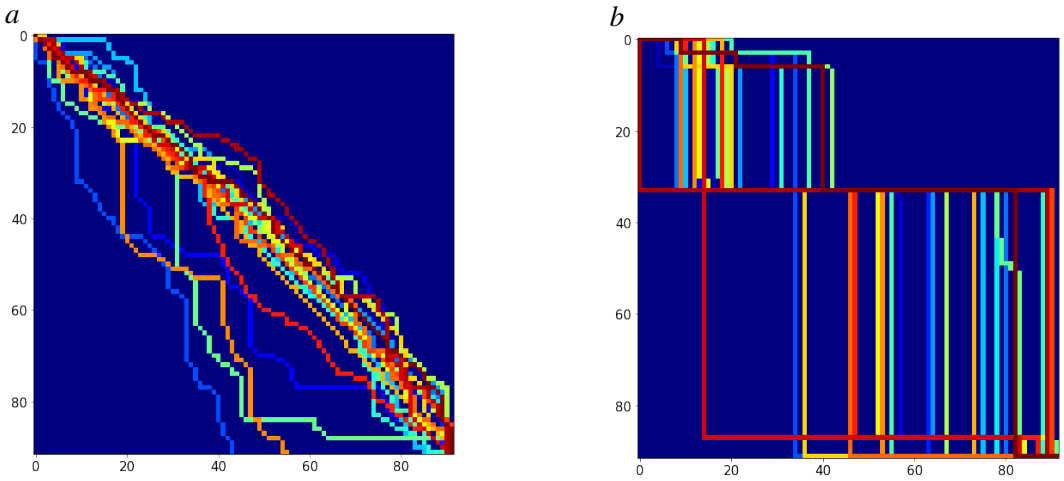


Figure 5: Alignment for image documents: *a)* with real text reuse source; *b)* with random documents.

To conclude the proposed method showed rather good quality for near-duplicate handwritten document retrieval. The method has a performance comparable to the performance of the text recognition-based methods and can be especially useful for low-resource languages that have no markup for recognition model training.

5 Conclusion

The paper is devoted to the problem of near-duplicate detection in handwritten school essay collections. The proposed method is based on word segmentation and further analysis of the extracted word lengths.

As a distance function between the essays, we analysed functions based on the dynamic time warping function. The computational experiment showed that the proposed method can efficiently work on large collections of school essays and comparable to the state-of-the-art handwritten text recognition methods. The future work includes analysis of different similarity functions and usage of different features that can be extracted from the text without its recognition.

Acknowledgements

This research is funded by RFBR, grant 19-29-14100.

References

[1] Axler Gregory, Wolf Lior. Toward a dataset-agnostic word segmentation method // 2018 25th IEEE International Conference on Image Processing (ICIP) / IEEE. — 2018. — P. 2635–2639.

[2] Discovering text reuse in large collections of documents: A study of theses in history sciences / Anton Khritankov, Pavel Botov, Nikolay Surovenko et al. // 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT) / IEEE. — 2015. — P. 26–32.

[3] End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network / Denis Coquenet, Clément Chatelain, Thierry Paquet, Ioan Grozny // arXiv preprint arXiv:2012.03868. — 2020.

[4] Giorgino Toni et al. Computing and visualizing dynamic time warping alignments in R: the dtw package // Journal of statistical Software. — 2009. — Vol. 31, no. 7. — P. 1–24.

[5] Krishnan Praveen, Jawahar C. V. Matching handwritten document images // European Conference on Computer Vision / Springer. — 2016. — P. 766–782.

[6] Liepieshov Kostiantyn, Dobosevych Oles. On recognition of Cyrillic Text // Workshop on Document Intelligence at NeurIPS 2019. — 2019.

[7] Louloudis Georgios, Gatos Basilios et al. Text line and word segmentation of handwritten documents // Pattern recognition. — 2009. — Vol. 42, no. 12. — P. 3169–3183.

[8] Ma Hongyan Jane, Wan Guofang, Lu Eric Yong. Digital cheating and plagiarism in schools // Theory Into Practice. — 2008. — Vol. 47, no. 3. — P. 197–203.

[9] Marti Urs-Viktor, Bunke Horst. The IAM-database: an English sentence database for offline handwriting recognition // International Journal on Document Analysis and Recognition. — 2002. — Vol. 5, no. 1. — P. 39–46.

[10] Methods for Intrinsic Plagiarism Detection and Author Diarization. / Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, Vadim Strijov // CLEF (Working Notes). — 2016. — P. 912–919.

[11] Mustakimova Elmira. Offline Recognition of Russian Handwriting : Master's thesis / Elmira Mustakimova. — 2016. — Master thesis.

[12] Pandey Om, Gupta Ishan, Mishra Bhabani S. P. A Robust Approach to Plagiarism Detection in Handwritten Documents // International Symposium on Visual Computing / Springer. — 2020. — P. 682–693.

[13] Prevalence of Plagiarism among Medical Students / Vedran Frković, Josip Ažman, Tamara Turk et al. // Book of Abstracts. ZIMS 4. — P. 38–38.

[14] Salvador Stan, Chan Philip. Toward accurate dynamic time warping in linear time and space // Intelligent Data Analysis. — 2007. — Vol. 11, no. 5. — P. 561–580.

[15] Scaling Handwritten Student Assessments With a Document Image Workflow System / Vijay Rowtula, Varun Bhargavan, Mohan Kumar, C. V. Jawahar // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. — 2018. — P. 2307–2314.

[16] Shavrina Tatiana, Shapovalova Olga. To the methodology of corpus construction for machine learning:“Taiga” syntax tree corpus and parser // Proceedings of “CORPORA-2017” International Conference. — 2017. — P. 78–84.

[17] Vinciarelli Alessandro, Luettin Juergen. A new normalization technique for cursive handwritten words // Pattern recognition letters. — 2001. — Vol. 22, no. 9. — P. 1043–1050.

[18] Voigtlaender Paul, Doetsch Patrick, Ney Hermann. Handwriting recognition with large multidimensional long short-term memory recurrent neural networks // 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) / IEEE. — 2016. — P. 228–233.

[19] Wrigley Stuart. Avoiding ‘de-plagiarism’: Exploring the affordances of handwriting in the essay-writing process // Active Learning in Higher Education. — 2019. — Vol. 20, no. 2. — P. 167–179.

Idiomaticity of a Text as a Matter of the Individual Style: A Quantitative Approach

A.N. Baranov
Russian Language Institute of RAS /
Moscow, Volkhonka 18/2
Baranov_anatoly@hotmail.com

D.O. Dobrovol'skij
Russian Language Institute of RAS /
Moscow, Volkhonka 18/2
Institute of Linguistics / Moscow,
B. Kislovskiy 1
Stockholm University / 10691
Stockholm Sweden
dobrovolskij@gmail.com

Abstract

The paper suggests one of the ways to formally define the degree of idiomaticity of a given text. Text idiomaticity is understood as the density of the use of idioms per text unit. The assessment of the degree of idiomaticity is carried out in the proposed approach as the ratio of the total number of idioms to the volume of the text in which they met. The conducted corpus experiment allows us to conclude that the degree of idiomaticity of the most important representatives of the prose of the second half of the 19th century varies significantly. Thus, the degree of idiomaticity of the text turns out to be an essential factor of the individual style.

Keywords: the degree of idiomaticity; idioms; individual style; quantitative methods of analysis
DOI: 10.28995/2075-7182-2021-20-58-67

Об одном подходе к количественной оценке идиоматичности текста как характеристике авторского стиля

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

Д.О. Добровольский
Институт русского языка РАН /
Москва, Волхонка 18/2
Институт языкознания РАН / Москва,
Большой Кисловский 1
Стокгольмский университет / 10691
Стокгольм
dobrovolskij@gmail.com

Аннотация

В докладе предлагается один из способов формального определения идиоматичности – той части этого феномена, которая связана с функционированием идиом. Оценка степени идиоматичности осуществляется в предлагаемом подходе как отношение общего количества идиом к объему текста, в котором они встретились. Проведенный корпусный эксперимент позволяет сделать вывод, что степень идиоматичности важнейших представителей художественной прозы второй половины XIX века варьирует в существенных пределах. Тем самым, идиоматичность текста оказывается существенным стилеобразующим фактором.

Ключевые слова: степень идиоматичности; идиоматика; индивидуальный стиль; количественные методы анализа

1 Феномен идиоматичности и возможности его формализации

Феномен идиоматичности, понимаемый как нестандартный способ сочетания смыслов, распространяется на весь язык. Тем самым в языке широко представлены идиоматичные формы (например, некомпозициональные сочетания морфем, метафоры, метонимические сдвиги нерегулярного характера и пр.) [Makkai 1978; Баранов, Добровольский 2013: 44–62]. Однако формализация феномена идиоматичности, представление его в объективной (количественной) форме, допускающей независимую проверку, в настоящее время невозможно. Да и в будущем – вряд ли, поскольку в этом случае потребуется разработка качественных критериев («весов»), которые позволят сравнивать и оценивать идиоматичность, возникающую за счет разных видов отклонения от регулярности (ср., например, речевые акты в несобственных употреблениях и грамматические формы русского императива, не передающие побуждения – *Знай я, чем это кончится, никогда бы не согласился в этом участвовать*).

Еще более широкое понимание идиоматичности представлено в [Апресян 1995], где идиоматичность связывается с выполнением сложившихся в данном языке норм: «идиоматично, т. е. в соответствии со сложившимися в данном языке и подчас трудно мотивируемыми нормами синтаксической, семантической и лексической сочетаемости» [Апресян 1995: 11]. Среди названных Ю.Д. Апресяном факторов к идиоматичности в нашем понимании относится только выполнение «трудно мотивируемых норм», в частности, следование правилам лексической сочетаемости, не выводящимся из толкования (или не описываемым в терминах семантических классов), то есть не мотивируемым семантически. Количественная оценка такого понимания идиоматичности также вряд ли возможна.

Значительная часть идиоматичности реализуется за счет употребления идиом (об идиомах см. [Баранов, Добровольский 2013: 69–72; 2020б]). Идиоматичность идиом более гомогенна и поддается объективизации.

Таким образом, термин «идиоматичность» здесь используется в более узком смысле как характеристика использования именно идиом. Отметим, что категория идиоматичности даже при таком сужении дает важную информацию об индивидуальном стиле автора, позволяя сравнивать практики употребления идиом у разных носителей языка. Нас будет интересовать идиоматичность (в указанном смысле) текстов Достоевского и ряда его современников – Толстого, Гончарова, Салтыкова-Щедрина, Тургенева. Особенности стиля указанных авторов находятся в центре исследовательского проекта по выявлению формальных характеристик авторского стиля (см. по этому поводу [Баранов, Добровольский 2018; 2019]).

Следует, однако, иметь в виду, что параметр идиоматичности сам по себе недостаточен для отождествления стиля автора, хотя в совокупности с другими речевыми особенностями может дать важную информацию об авторстве. Возможности использования параметра идиоматичности в стилеметрии в данной работе не обсуждаются.

Естественно вычислять индекс идиоматичности, характеризующий плотность употребления идиом в тексте, как соотношение количества идиом в данном фрагменте текста к общему количеству содержащихся в нем знаков. Вычисление индекса идиоматичности предполагает проведение эксперимента с текстами Достоевского и его современников. До проведения эксперимента нам казалось, что наиболее идиоматичны в своих произведениях Достоевский и Салтыков-Щедрин. В наименьшей степени идиоматика присутствует в прозе Л.Н. Толстого. Интересно было проверить, насколько наша интуиция соответствует объективным данным.

Эксперимент включал в себя несколько этапов. На первом этапе были отобраны фрагменты произведений указанных авторов. Каждый фрагмент включал порядка 60 тыс. знаков. Если у кого-то автора в исследование попадал существенно больший фрагмент, то это компенсировалось в последующем за счет меньшего объема фрагментов других произведений.

В основном мы ориентировались на крупные романы. Кроме того, по личным предпочтениям выбирались отдельные произведения малой прозы, которые иногда брались целиком. Причем во всех случаях мы ориентировались на такие фрагменты, где присутствуют диалоги персонажей. Достоевский представлен в большем объеме, чем другие авторы. Это объясняется тем, что основная цель проекта заключается в изучении идиостиля именно этого автора. В связи с этим было важно учесть его главные романы – «Преступление и наказание», «Бесы», «Идиот», «Братья Карамазовы» и «Подросток», – а также малую прозу. Как примеры малой прозы мы выбрали «Село

Степанчиково и его обитатели» и «Дядюшкин сон» из-за того, что они диалогичны: разговоры персонажей занимают большое место в текстах этих произведений.

Л.Н. Толстой представлен в проводившемся эксперименте романами «Война и мир», «Анна Каренина» и «Воскресенье». Малая проза – «Смерть Ивана Ильича» (полностью) и «Дьявол» (полностью).

Из произведений И.С. Тургенева взяты романы «Отцы и дети», «Накануне», «Новь», из малой прозы – повесть «Вешние воды», рассказ «Степной король Лир» (полностью).

И.А. Гончаров представлен романами «Обрыв», «Обломов», «Обыкновенная история», путевыми очерками «Фрегат Паллада» и повестью «Счастливая ошибка».

Из корпуса текстов М.Е. Салтыкова-Щедрина были выбраны фрагменты романов «Господа Головлевы», «Пошехонская старина», «История одного города» и фрагменты рассказов, объединенных в циклы «Сказки» и «Помпадуры и помпадурши».

В каждом выбранном фрагменте отмечались вхождения всех идиом, вне зависимости от их модификаций и характера их использования. Не учитывались коллокации (*предложить вопрос, положить резолюцию*), пословицы (*близок локоть, да не укусишь; даром-то и прыщ на носу не вскочит; гром не грянет, мужик не перекрестится*), грамматические фразеологизмы типа *стало быть, по крайней мере*. Подробнее о типах фразеологизмов и правилах их выделения см. [Баранов, Добровольский 2013: 67–98][1].

2 Частотное распределение употребления идиом по авторам

Рассмотрим последовательно частоту употребления идиом выбранных авторов, ориентируясь на фрагменты, выделенные для эксперимента. Начнем с произведений Тургенева и Гончарова, поскольку, как будет показано ниже, они демонстрируют близкие характеристики частотности в отношении идиоматики. Далее проанализируем полученные результаты по текстам Л.Н. Толстого, а завершим изложение двумя наиболее «идиоматичными» авторами – Салтыковым-Щедриным и Достоевским.

Как уже отмечалось выше, при выборе фрагментов для проведения эксперимента отдавалось предпочтение диалогам персонажей. Так, во фрагменте из романа «Отцы и дети» было выявлено 53 контекста употребления идиом. Например, в разговоре с Аркадием Николай Петрович использует идиому *на славу*:

(1) – Теперь уж недалеко, – заметил Николай Петрович, – вот стоит только на эту горку подняться, и дом будет виден. Мы заживем с тобой *на славу*, Аркаша; ты мне помогать будешь по хозяйству, если только это тебе не наскучит. [И.С. Тургенев. Отцы и дети]

Это, однако, не означает, что идиомы в словах автора не учитывались. В том же фрагменте идиомы встречаются и в речи автора:

(2) Он без нужды растягивал свою речь, избегал слова «папаша» и даже раз заменил его словом «отец», произнесенным, правда, *сквозь зубы*; с излишнею развязностью налил себе в стакан гораздо больше вина, чем самому хотелось, и выпил все вино. Прокофьич *не спускал* с него *глаз* и только губами пожевывал. [И.С. Тургенев. Отцы и дети]

Общее количество идиом и их распределение по фрагментам отдельных произведений приводится ниже в таблице 1.

[1] Словник идиом XX и начала XXI века представлен в «Тезаурусе русских идиом: семантические группы и контексты» [Тезаурус 2018]. В эксперименте учитывались также идиомы, представленные в русской классической литературе XIX века, подтвержденные материалом Национального корпуса русского языка.

Произведение	**Абсолютная частота**	**Относительная частота**[2]
Накануне	41 (56851)[3]	0,72
Новь	53 (54440)	0,97
Вешние воды	19 (47794)	0,4
Отцы и дети	53 (64357)	0,82
Степной король Лир	148 (136223)	1,09
Общее количество	**314 (359665)**	**0,87**

Таблица 1: Идиоматика И.С. Тургенева по данным эксперимента

Наибольшее значение индекса идиоматичности приходится на рассказ «Степной король Лир», что неудивительно: серьезный конфликт между персонажами стимулирует активное употребление идиом – даже не обязательно конфликтной семантики. Ср., например:

(3) – И как он мне сказал, ваш-то Володька, – с новой силой подхватил Харлов, – как сказал он мне, что мне в моей горенке больше не жить, а я в самой той горенке каждое бревнышко собственными руками клал – как сказал он мне это – и *бог знает*, что со мной приключилось! В головушке помутилось, *по сердцу как ножом*... Ну, либо его зарезать, либо из дому вон!.. Вот я и побежал к вам, благодетельница моя, Наталья Николаевна... И куды ж мне было *голову приклонить*? [И.С. Тургенев. Степной король Лир]

Внутренний конфликт присутствует и в повести «Вешние воды», однако он не проявляется в явном виде в прямой речи персонажей. Меньшая эмоциональность дискурса снижает частоту употребления идиоматики.

Частотные характеристики идиом в произведениях Гончарова близки соответствующим характеристикам в текстах Тургенева; см. таблицу 2.

Произведение	**Абсолютная частота**	**Относительная частота**
Счастливая ошибка	64 (61495)	1,04
Фрегат Паллада	30 (60020)	0,5
Обыкновенная история	106 (59434)	1,78
Обрыв	53 (59991)	0,88
Обломов	76 (58735)	1,29
Общее количество	**329 (299675)**	**1,10**

Таблица 2: Идиоматика И.А. Гончарова по данным эксперимента

Из приведенной таблицы видно, что чаще всего идиомы употребляются в проанализированном фрагменте из романа «Обыкновенная история». Можно предположить, что это связано с особенностями речи конкретного персонажа, участвующего в диалогах выбранного фрагмента:

(4) – Ну, сядь, сядь! – отвечала она, наскоро утирая слезы, – мне еще много осталось поговорить... Что бишь я хотела сказать? *из ума вон*... Вишь, нынче какая память у меня... да! блюди посты, мой друг: это великое дело! В среду и пятницу – *бог простит*; а в великий пост – *боже оборони*! Вот Михайло Михайлыч и умным человеком считается, а что в нем? Что мясоед, что страстная неделя – все одно жрет. Даже *волос дыбом становится*! Он вон и бедным помогает, да будто его милостыня принята господом? Слышь, подал раз старику красненькую, тот взял ее, а сам отвернулся да плюнул. Все кланяются ему и *в глаза-то бог знает что* наговорят, а *за глаза* крестятся, как поминают его, словно шайтана какого. <…> – Береги пуще всего здоровье, – продолжала она. – Как заболеешь – чего *боже оборони*! – опасно, напиши... я соберу все силы и приеду. Кому там ходить за тобой? Норовят еще обобрать больного. Не ходи ночью по улицам; от людей зверского

[2] На 1000 знаков текста.

[3] В скобках указывается объем фрагмента в знаках.

вида удаляйся. Береги деньги... ох, береги *на черный день*! Трать с толком. От них, проклятых, всякое добро и всякое зло. Не мотай, не заводи лишних прихотей. Ты будешь аккуратно получать от меня две тысячи пятьсот рублей в год. Две тысячи пятьсот рублей *не шутка*. Не заводи роскоши никакой, ничего такого, но и не отказывай себе в чем можно; захочется полакомиться – не скупись. – Не предавайся вину – ох, оно первый враг человека! – Да еще (тут она понизила голос) берегись женщин! Знаю я их! Есть такие бесстыдницы, что сами *на шею будут вешаться*, как увидят этакого-то... [И.А. Гончаров. Обыкновенная история]

В примере (4) представлен дискурс наставления: Анна Павловна Адуева, речь которой образна и передает высокую степень эмпатии, провожает в дорогу своего сына Александра.

В противоположность роману «Обыкновенная история», путевые заметки «Фрегат Паллада» по жанру не предполагают воспроизведение живой речи, что отражается на использовании идиом.

Как показал эксперимент, наименее идиоматичны тексты Л.Н. Толстого, причем распределение частот по разным проанализированным произведениям оказывается близким; ср. таблицу 3.

Произведение	**Абсолютная частота**	**Относительная частота**
Воскресенье	14 (41966)	0,33
Дьявол	40 (80325)	0,5
Смерть Ивана Ильича	44 (108464)	0,41
Война и мир	36 (60718)	0,59
Анна Каренина	21 (60271)	0,35
Общее количество	**155 (351744)**	**0,44**

Таблица 3: Идиоматика Л.Н. Толстого по данным эксперимента

В целом полученные данные соответствуют интуитивным представлениям об использовании идиоматики Толстым. Это не означает, что произведения Толстого неидиоматичны, но идиоматичность может обеспечиваться за счет других средств передачи нерегулярной семантики. Анализ выбранных фрагментов показывает, что совместная дискурсивная встречаемость идиом (то есть в репликах персонажей, объединенных одним коммуникативным ходом) – весьма редкое явление. Приведем, однако, пример такого рода:

(5) – Ну, ну, хорошо! – сказал старый граф, – все горячится. Все Бонапарте всем *голову вскружил*; все думают, как это он из поручиков попал в императоры. Что ж, *дай бог*, – прибавил он, не замечая насмешливой улыбки гостьи. [Л.Н. Толстой. Война и мир]

В пример (5) идиомы *вскружить голову* и *дай бог* относятся к нейтральному стилю, и это очень характерно для практик использования идиом в текстах Толстого.

В отличие от Толстого (и даже Гончарова и Тургенева), Салтыков-Щедрин широко использует идиоматику, что и проявляется в индексе идиоматичности, см. таблицу 4.

Произведение	**Абсолютная частота**	**Относительная частота**
Сказки	141 (58888)	2,39
Помпадуры и помпадурши	57 (65632)	0,87
История одного города	63 (61503)	1,02
Пошехонская старина	104 (62297)	1,67
Господа Головлевы	112 (66030)	1,7
Общее количество	**477 (314350)**	**1,52**

Таблица 4: Идиоматика М.Е. Салтыкова-Щедрина по данным эксперимента

Из таблицы 4 следует, что индекс идиоматичности у Салтыкова-Щедрина оказывается самым высоким по сравнению с уже рассмотренными авторами. Наибольшей величины он достигает в

текстах «Сказок» – 2,39. Полученное значение индекса идиоматичности отвечает интуиции, поскольку речь рассказчика и персонажей произведений Салтыкова-Щедрина содержит много идиом различной стилистической направленности:

(6) *Понурил* Ловец *голову*, потому что знал, что Ловчихино слово твердое. Снял он с себя пальто – и вдруг словно преобразился совсем! Так как совесть осталась, вместе с пальто, на стенке, то сделалось ему опять и легко, и свободно, и стало опять казаться, что *на свете* нет ничего чужого, а всё его. И почувствовал он вновь в себе способность глотать и загребать. – Ну, теперь вы у меня не отвертитесь, дружки! – сказал Ловец, *потирая руки*, и стал поспешно надевать на себя пальто, чтоб *на всех парусах* лететь на базар. [М.Е. Салтыков-Щедрин. Пропала совесть]

В приведенном примере обнаруживаются и литературные идиомы (*понурить голову, на всех парусах*), и нейтральные (*на свете*), и разговорные (*потирать руки*). Такое распределение до известной степени объясняется тем, что в примере (6) присутствует как авторская речь, так и речь персонажа.

В иных случаях последовательность идиом оказывается стилистически более однородной, ср. (7):

(7) – Ну, делай как знаешь! В Головлеве так в Головлеве ему жить! – наконец, сказала она, – окружил ты меня кругом! опутал! начал с того: как вам, маменька, будет угодно! а под конец заставил-таки меня *под свою дудку плясать*! Ну, только слушай ты меня! Ненавистник он мне, всю жизнь он меня *казнил да позорил*, а наконец и над родительским благословением моим надругался, а все-таки, если ты его *за порог выгонишь* или *в люди* заставишь *идти* – нет тебе моего благословения! Нет, нет и нет! Ступайте теперь оба к нему! чай, он и *буркалы-то свои проглядел*, вас высматриваючи! [М.Е. Салтыков-Щедрин. Господа Головлевы]

Идиомы *плясать под чью-либо дудку, казнить да позорить*[4]*, выгнать за порог, в люди*, скорее, относятся к разговорным, а *проглядеть буркалы* – к просторечным. Сходные по концентрированности идиом в речи и по стилистическим характеристикам примеры встречаются и в других произведениях Салтыкова-Щедрина:

(8) – Ишь печальник нашелся! – продолжает поучать Анна Павловна, – уж не *на все ли четыре стороны* тебя отпустить? *Сделай милость*, воруй, голубчик, поджигай, грабь! Вот ужо в городе тебе покажут... *Скажите на милость*! целое утро словно в котле кипела, только что отдохнуть собралась – *не тут-то было*! солдата *нелегкая принесла*, с ним валандаться изволь! *Прочь с моих глаз*... поганец! Уведите его да накормите, а не то еще издохнет, *чего доброго*! А часам к девяти приготовить подводу – и *с богом*! [М.Е. Салтыков-Щедрин. Пошехонская старина]

Такие случаи характерны для прямой речи персонажей, наделенных яркой речевой индивидуальностью.

Достоевский по дискурсивным практикам использования идиом близок Салтыкову-Щедрину; ср. таблицу 5.

[4] Форма *казнить [да/и] позорить* была употребительной как идиома в XIX в. По данным НКРЯ встречается также в произведениях Т.Г. Шевченко и М.Н. Загоскина.

Произведение	**Абсолютная частота**	**Относительная частота**
Подросток	78 (61657)	1,27
Идиот	96 (64746)	1,48
Дядюшкин сон	103 (61598)	1,67
Село Степанчиково и его обитатели	137 (64869)	2,11
Бесы	79 (64262)	1,23
Преступление и наказание	91 (61597)	1,48
Братья Карамазовы	67 (62609)	1,07
Общее количество	**651 (441338)**	**1,48**

Таблица 5: Идиоматика Ф.М. Достоевского по данным эксперимента

Как видно из таблицы 5, индекс идиоматичности текстов Достоевского достаточно велик. Вместе с Салтыковым-Щедриным они образуют «лидерскую группу». И по многим другим лингвистическим характеристикам индивидуальные стили этих авторов оказываются близкими, ср. в частности употребление дискурсивных слов [Баранов, Добровольский 2020а].

Наиболее идиоматичной оказывается повесть «Село Степанчиково и его обитатели», индекс идиоматичности которой превышает 2. Это проявляется в частности в том, что в коммуникативном ходе одного персонажа может использоваться значительное количество идиом, ср. (9):

(9) – Да, сударь, я вам такое могу рассказать, что вы только *рот разинете* да так и останетесь *до второго пришествия с разинутым ртом*. Ведь я прежде и сам его уважал. Вы что думаете? Каюсь, открыто каюсь: был дураком! Ведь он и меня обморочил. Всезнай! *Всю подноготную знает*, все науки произошел! Капель он мне давал: ведь я, батюшка, человек больной, сырой человек. Вы, может, не верите, а я больной. Ну, так я с его капель-то чуть *вверх тормашки* не полетел. Вы только молчите да слушайте; сами поедете, всем полюбуетесь. Ведь он там полковника-то *до кровавых слез доведет*; ведь *кровавую слезу прольет* от него полковник-то, да уж поздно будет. <…> Я, дескать, ученый. Да что ж, что ученый! Так из-за того, что ученый, уж так непременно и надо заесть неученого?.. И уж как начнет ученым своим *языком колотить*, так уж та-та-та! та-та-та! <…> Зазнался, *надулся как мышь на крупу*! Ведь уж туда теперь лезет, *куда и голова его не пролезет*[5]. Да чего! Ведь он там дворовых людей по-французски учить выдумал! <…> Небось, и вы по-французски: «та-та-та! та-та-та! *вышла кошка за кота*!» – прибавил Бахчеев, смотря на меня с презрительным негодованием. [Ф.М. Достоевский. Село Степанчиково и его обитатели]

Конечно, не следует думать, что все персонажи повести столь идиоматичны – в приведенном примере это, разумеется, часть речевого портрета героя.

Персонажи произведений Достоевского не просто используют идиомы, но и обмениваются ими в репликах, ср. характерный пример:

(10) – С наступающим днем! Да ты смотри, сколько дня-то ушло, человек несообразный! – *Ври, Емеля, – твоя неделя*! – По-нашему, *хоть на час, да вскачь*! [Ф.М. Достоевский. Село Степанчиково и его обитатели]

Помещик Васильев репликой *ври, Емеля, – твоя неделя* выражает свое недоверие к сообщению о том, что он проспал почти весь день. В ответной реплике *по-нашему, хоть на час, да вскачь* обращается внимание на причину столь долгого сна – сильное опьянение.

Проведенный комплекс замеров указывает на то, что наименее идиоматичен текст романа «Братья Карамазовы». Не исключено, что это связано с выбором одного фрагмента из довольно большого литературного произведения. Однако и в «Братьях Карамазовых» встречаются реплики, с концентрированным употреблением идиом:

[5] Встречается в XIX в. как идиома, например, у А.П. Чехова.

(11) – Совершенно *как дома*? То есть *в натуральном-то виде*? О, этого много, слишком много, но – с умилением принимаю! Знаете, благословенный отец, вы меня на *натуральный-то вид* не вызывайте, не рискуйте... до *натурального вида* я и сам не дойду. Это я, чтобы вас охранить, предупреждаю. [Ф.М. Достоевский. Братья Карамазовы]

Федор Павлович в ответ на предложение старца Зосимы чувствовать себя как дома провокационно интерпретирует эту идиому как санкцию на неприличное поведение, эвфемистически называя его *натуральным видом*. Эта идиома довольно часто встречается в текстах XIX века, обозначая в основном голого человека. Федор Павлович употребляет ее явно расширительно.

3 Варьирование индекса идиоматичности по исследуемым авторам

Выявленные значения индекса идиоматичности у исследованных авторов представлены в таблице 6 и графике 1.

Авторы	**Значение индекса идиоматичности**
Л.Н. Толстой	0,44
И.С. Тургенев	0,87
И.А. Гончаров	1,10
Ф.М. Достоевский	1,48
М.Е. Салтыков-Щедрин	1,52

Таблица 6: Распределение значений индекса идиоматичности

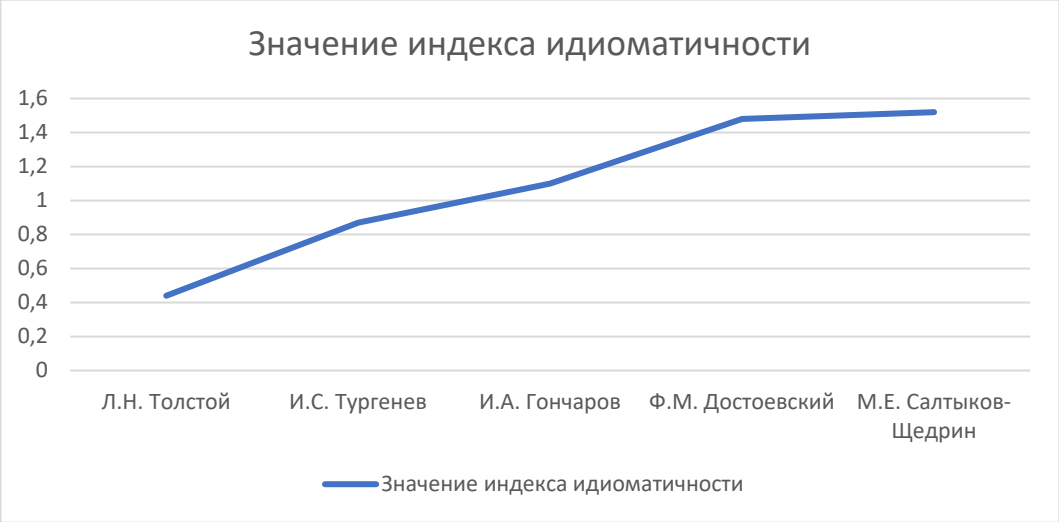


График 1: Графическое представление индекса идиоматичности

Из приведенной таблицы и графика хорошо видно, что Достоевский и Салтыков-Щедрин очень близки по степени идиоматичности (в указанном понимании) и существенно превосходят по этому параметру всех остальных исследованных авторов. В то же время тексты Толстого отличаются наименьшей идиоматичностью. Остальные два автора – Тургенев и Гончаров, – хотя и отличаются друг от друга, но не столь сильно, как они вместе от Толстого, с одной стороны, и от Достоевского и Салтыкова-Щедрина – с другой.

4 Заключение

Технология подсчета индекса идиоматичности, выбранная в данном эксперименте, далеко не идеальна. К ней легко можно предъявить претензии. Действительно, по приведенным данным видно,

что индекс идиоматичности варьирует не только от автора к автору, но и внутри текстов каждого автора от одного произведения к другому, причем степень варьирования оказывается весьма значительной – более, чем в два раза. Единственный писатель, индекс идиоматичности которого не обнаруживает большое отклонение от произведения к произведению – это Л.Н. Толстой. Конечно, это зависит от выбора фрагмента для подсчета идиом, но в то же время это определяется и жанром текста, теми художественными задачами, которые ставит перед собой писатель в данном произведении, и даже составом персонажей.

При всех ограничениях выбранного способа подсчета идиоматичности окончательный результат оказался весьма правдоподобным, что объясняется выбором для анализа нескольких важных для конкретного автора произведений. Действительно, стиль Толстого не оставляет места для широкого использования идиом – даже в тех случаях, когда идет эмоционально насыщенный диалог персонажей, они редко используют разговорную идиоматику. С другой стороны, и Салтыков-Щедрин и Достоевский нагружают реплики героев идиомами, сознательно прибегая к такому художественному средству выражения смысла. Анализ «Записных книжек» Достоевского, не рассматриваемых в данной работе, показывает, что он последовательно фиксировал услышанные идиомы живой речи и вкладывал их в уста персонажей. Тургенев и Гончаров явно более идиоматичны, чем Толстой, но идиоматику в их произведениях нельзя рассматривать как характерный художественный прием.

Благодарности

Работа выполнена в рамках проекта по гранту РФФИ № 18-012-90025. Авторы благодарят фонд за поддержку.

References

[1] Apresjan Ju.D. Selected Writings [Izbrannye trudy] — Vol. 1: Lexical Sematics [Leksicheskaya semantika] — Moscow: Yazyki russkoy kul'tury, 1995.
[2] Baranov A.N., Dobrovol'skij D.O. Fundumentals of Phraseology [] — Moscow: Flinta, Nauka, 2013.
[3] Baranov A.N., Dobrovol'skij D.O. (2018), *Kstati* and *nekstati*: on Dostoevsky's discourse practices [*Kstati* i *nekstati*: k rechevym praktikam Dostoyevskogo] — Russkiy yazyk v nauchnom osvescchenii, 2018. — No. 1 (35). — P. 33–45.
[4] 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 "Dialog 2019" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2019"], Moscow, pp. 41–52.
[5] Baranov A.N., Dobrovol'skij D.O. (2020a), Style dynamics of the Russian written speech of the 19th century: a corpus study [Dinamika stilya russkoy pis'mennoy rechi XIX veka: korpusny eksperiment], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2020" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2020"], Moscow, pp. 48–61.
[6] Baranov A.N., Dobrovol'skij D.O. (2020b), Idiom [Idioma], Russky yazyk. — Moscow: Ast-Press, 2020. — pp. 197–198.
[7] Makkai Adam. Idiomaticity as a language universal // Greenberg J.H. (ed.) Universals of human language. Stanford, 1978.

Литература

[1] Апресян Ю.Д. Избранные труды. Т. 1: Лексическая семантика. Изд. 2-е, испр. и доп. М.: Языки русской культуры, 1995.
[2] Баранов А.Н., Добровольский Д.О. Основы фразеологии (краткий курс). М.: Флинта, Наука, 2013.
[3] Баранов А.Н., Добровольский Д.О. (2018), *Кстати* и *некстати*: к речевым практикам Достоевского — Русский язык в научном освещении, 2018. — № 1 (35). — С. 33–45.
[4] Баранов А.Н., Добровольский Д.О. (2019), 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., 2019. — С. 41–52.

[5] Баранов А.Н., Добровольский Д.О. (2020а), Динамика стиля русской письменной речи XIX века: корпусный эксперимент // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог». 2020. Выпуск 19 (26). М., 2020. — С. 48–61.
[6] Баранов А.Н., Добровольский Д.О. (2020б), Идиома // Русский язык: Энциклопедия / ИРЯ РАН; Под общ. ред. А.М. Молдована. — 3-е изд., перераб. и доп. — М.: Аст-Пресс Школа, 2020. — С. 197–198.
[7] Тезаурус 2018 — Тезаурус русских идиом: семантические группы и контексты. — 2-е изд., испр. и доп. — М.: Лексрус, 2018.
[8] Makkai Adam. Idiomaticity as a language universal // Greenberg J.H. (ed.) Universals of human language. Stanford, 1978.

The order of objects in Russian: a corpus study*

Bazhukov M.O.
NRU HSE
Moscow
mobazhukov@hse.ru

Chubarova L.I.
NRU HSE
Moscow
lchubarova@hse.ru

Slioussar N.A.
NRU HSE, SPbU
Moscow
nslioussar@hse.ru

Toldova S. Yu.
NRU HSE
Moscow
stoldova@hse.ru

Abstract

The paper presents the results of a corpus study of the order of direct and indirect objects in ditransitive constructions in Russian (like *Petya dal Mashe yabloko* 'Petya gave Masha an apple' or *Petya dal yabloko Mashe* 'Petya gave an apple to Masha'). This topic has been widely discussed in the literature, but previous hypotheses have been based on individual examples and have never been tested on corpus data. Based on earlier research, we have selected parameters that affect the order of the objects, such as the length, depth, animacy and role of individual verbs and statistically tested their real effect on two subsamples: with a dative indirect object and with a prepositional one.

Keywords: word order; ditransitive constructions; animacy; argument prominence hierarchy; information structure; Russian

DOI: 10.28995/2075-7182-2021-20-68-78

Порядок дополнений в русском языке: корпусное исследование

Бажуков М.О.
НИУ ВШЭ
Москва
mobazhukov@hse.ru

Чубарова Л.И.
НИУ ВШЭ
Москва
lchubarova@hse.ru

Слюсарь Н.А.
НИУ ВШЭ, СПбГУ
Москва
nslioussar@hse.ru

Толдова С.Ю.
НИУ ВШЭ
Москва
stoldova@hse.ru

Аннотация

В статье представлены результаты корпусного исследования порядка прямого и косвенного дополнения в русских дитранзитивных конструкциях (типа *Петя дал Маше яблоко* и *Петя дал яблоко Маше*). Эта тема ранее широко обсуждалась в литературе, однако предыдущие гипотезы были основаны на отдельных примерах и никогда не тестировались на корпусных данных. Основываясь на предыдущих работах, мы отобрали параметры, влияющие на порядок дополнений, такие как длина, глубина, одушевленность дополнений и влияние отдельных глаголов и проверили их влияние на двух подвыборках: с косвенным дативным и предложным дополнениями.

Ключевые слова: порядок слов; дитранзитивные конструкции; одушевлённость; иерархия доступности аргументов; информационная структура; русский язык

1 Introduction

One of the topics that is widely discussed in word order studies is the ordering of direct and indirect objects (DOs and IOs). In the languages in which different word orders are possible, like in German or in Russian, both functional and formal studies seek to determine which factors govern the distribution of these orders and which order can be considered basic, neutral or canonical (these three terms are often used interchangeably). For Russian, these questions have been addressed by many generative syntacticians

* This work is an output of a research project "Interface phenomena in grammar of languages of Russia: a formal approach" implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

([4, 5, 6, 11, 19, 30, 37, 38, 39, 40], among others). The authors do not agree which order is neutral, IO before DO or DO before IO, and include different factors in the analysis.

However, their conclusions have never been tested in a corpus study — all these papers are based on the analysis of individual examples. Early functional corpus studies of the Russian word order looked at object ordering [45], but did not take almost any potentially relevant factors into account. In the present paper, we aim to close this gap in a corpus study analyzing the role of various factors that were discussed in the literature for Russian and for other languages.

2 Previous Studies

Russian word order has been analyzed in many functional and formal studies. Early functional studies like Sirotinina [45] and Kovtunova [24] demonstrated that new information tends to follow given and elements that are part of the rheme tend to follow those that are not. However, these studies did not consider ditransitive verbs in any detail. Sirotinina only noted that the 'IO DO' order is more frequent in Russian. This observation is repeated in the Russian Grammar [34].

Junghanns and Zybatow [19] was the first generative study focusing on ditransitive verbs in Russian. The authors showed that the 'DO IO' order, as in (1b), is compatible only with a given DO and/or a narrow focus on the IO. They concluded that 'IO DO' is the basic order in Russian, while 'DO IO' is derived by direct object scrambling.

(1) *Odna ženščina podarila mal'čiku jabloko.* S V IO DO
one$_{NOM}$ woman$_{NOM}$ gave boy$_{DAT}$ apple$_{ACC}$
'A woman gave a/the boy an/the apple.'
Odna ženščina podarila jabloko mal'čiku. S V DO IO
one$_{NOM}$ woman$_{NOM}$ gave apple$_{ACC}$ boy$_{DAT}$
'A woman gave the apple to a boy.'

Dyakonova ([11, 12]) provided further arguments in favor of this conclusion. She showed that idioms with ditransitive verbs often include the verb and the DO (e.g. *stroit' komu-to glazki* 'to flirt with smb', *peremyvat' komu-to kostočki* 'to gossip about smb'), but almost never the verb and the IO. When a part of the VP is topicalized, it can include the verb and the DO, as in (2a), but much less readily the verb and the IO, as in (2b).

(2) [*Čitat' skazki*] *roditeli detjam očen' ljubjat.*
read$_{INF}$ tales$_{ACC}$ parents$_{NOM}$ kids$_{DAT}$ very love
??/**Čitat' detjam roditeli skazki očen' ljubjat.*
read$_{INF}$ kids$_{DAT}$ parents$_{NOM}$ tales$_{ACC}$ very love

However, Bailyn ([4, 5]) argued for the opposite view relying on the asymmetries in reciprocal and variable binding and examples involving instrumental secondary predicates. Let us consider (3a–d) with reciprocals. (3a) is fine, while (3c) is worse. (3b), which, according to Bailyn, is derived from (3a), is also fine, while (3d), which could be derived from (3c) in the same way, is ungrammatical.

(3) *Mama predstavila Petrovyx drug drugu.*
mother$_{NOM}$ introduced Petrovs$_{ACC}$ each other$_{DAT}$
'The mother introduced the Petrovs to each other.'
Mama predstavila drug drugu Petrovyx.
mother$_{NOM}$ introduced each other$_{DAT}$ Petrovs$_{ACC}$
?*Mama predstavila Petrovym drug drugu.*
mother$_{NOM}$ introduced Petrovs$_{DAT}$ each other$_{DAT}$
**Mama predstavila drug drugu Petrovym.*
mother$_{NOM}$ introduced each other$_{DAT}$ Petrovs$_{DAT}$

Titov [40] supports Bailyn's ([4, 5]) position arguing that in the analysis of Russian word order, not only widely conceived information structure (the distinctions like given/new, topic/focus, referential/non-referential), but also animacy should be taken into account. She introduces the Argument Prominence

Hierarchy: ±presupposed, ±referential, ±animate and well-formedness constraints requiring the arguments that are higher in the hierarchy to precede those that are lower. Thus, (1a) rather than (1b) is used in an all-new context because the IO is animate and the DO is inanimate in this sentence. Titov [40] claims that when the objects do not differ with respect to animacy or information structure, the 'DO IO' order is used.

Here, it is interesting to compare the concept of the basic / neutral / canonical word order in the formal and functional approaches. In the latter, it is the most frequent order. In the former, it is the order that is used when all relevant factors are balanced. If IOs happen to be animate more often that DOs, this might obscure the picture on the surface making some non-canonical word order the most frequent[1] . We consider both approaches worthwhile and incorporate both perspectives in our corpus study.

Finally, Boneh and Nash (2017) found different patterns for different individual verbs. They argued that IOs can occupy different positions in the syntactic structure, depending on their semantic role and the type of the predicate. Similar ideas were expressed, for example, in [16] for German.

While most previous studies on Russian focused on information structure, our paper analyzes animacy and several other factors that have been applied to ditransitive constructions crosslinguistically, but did not receive enough attention in Russian. The role of animacy in ditransitive constructions has been discussed in many typological studies (e.g. [14, 21, 18]) and in corpus and experimental work on individual languages: a corpus study on German [20] or an acceptability judgment study on Croatian [42] can be taken as examples. Faltz [14] explained it by "the greater cognitive salience of the typically animate IO argument over the typically inanimate DO argument" (p. 84). More recent accounts share this insight.

Other factors identified as relevant include the presence or absence of the preposition in the IO and the heaviness of the objects. As is well known, in English the order of the objects is fixed, and the IO precedes the DO if and only if it is not introduced by a preposition. This rule does not hold in case of the so-called heavy NP shift: DOs that are especially long and syntactically complex may follow prepositional IOs. The role of heaviness for constituent order has been discussed in numerous studies, including [17, 13, 44, 43].

Another well-known factor is prominalization. Pronouns have various properties ranging from phonology to semantics and information structure that affect their syntactic behavior. In double object constructions, pronouns have a very strong tendency to precede full noun phrases, often cliticizing on the verb, if word order alternations are possible (in the languages like German or Russian). We will look for similar effects in Russian in our corpus study and will test the role of other factors mentioned above.

3 Our corpus study

3.1 Data

The dataset used in this study was obtained from the SynTagRus corpus [1], [36]. It is a dependency treebank developed by the Computational Linguistics Laboratory, A.A. Kharkevich Institute of Information Transmission Problems. Currently the treebank contains over 1.100.000 tokens (over 77.000 sentences) from the texts of various genres.

Initially, we tried to work with a larger Taiga corpus [35], which is automatically annotated. But the error rate was extremely high: about 35% of the extracted sentences had annotation errors, either morphological like verb lemmatization, or syntactic. For instance, about half of the errors were the cases when a DO modifier was annotated as an IO. Conversely, SynTagRus has a comprehensive manually corrected morphological and syntactic annotation in the spirit of the dependency grammar. This allowed extracting the relevant examples with virtually no errors. Besides, the detailed syntactic annotation in SynTagRus allowed us to differentiate verb arguments from adjuncts.

We extracted corpus sentences containing a verb that governs an accusative object and an indirect object, i.e. word forms that are marked with the following relations to the verb: '1-компл' [first comple-

[1]How such non-canonical orders are derived is a central question for formal theories, but it is outside of the scope of the present paper. Let us only note that in case of double object constructions, both in Russian and cross-linguistically, there is a big debate whether non-canonical orders result from syntactic movement (e.g. [3, 4, 5, 19, 9, 25, 37]) or are base-generated (e.g. [8, 10, 29, 31, 40])

ment] and '2-компл' [second complement][2] . Finally, we limited our dataset to 6398 contexts with dative IOs and IOs with a preposition. These two groups were discussed in the literature and would be interesting to compare. We excluded a large number of sentences with IOs in the instrumental case, leaving them for further research, and several other examples with genitive or nominative IOs (primarily with the verbs *lišit'* 'to deprive' and *nazyvat'* 'to name'). The distribution of word orders in the resulting dataset is given in Table 1.

Word order	Number and percentage of examples
DO V IO	1087 (17%)
IO V DO	616 (9,6%)
DO IO V	208 (3,2%)
IO DO V	125 (2%)
V DO IO	2591 (40,5%)
V IO DO	1771 (27,7%)

Table 1: The distribution of word orders in the final dataset before filtering.

As Table 1 makes clear, one of the objects or both of them often precede the verb in Russian. The distribution of VO vs. OV orders is a separate topic in Russian syntax (e.g. [7, 9, 22, 23, 24, 27, 28, 45, 46]), so in the present study, we decided to focus on the cases in which both objects follow the verb. We also filtered out examples with pronominal objects because their syntactic position is primarily determined by their "lightness". Figure 1 shows the distribution of word orders in these sentences. If only one of the objects is pronominal, it precedes the other in the absolute majority of cases. Needless to say, this tendency is highly statistically significant (χ^2 = 823,13, p<0,001). As for the sentences with two pronominal objects, there are too few of them for any further analysis.

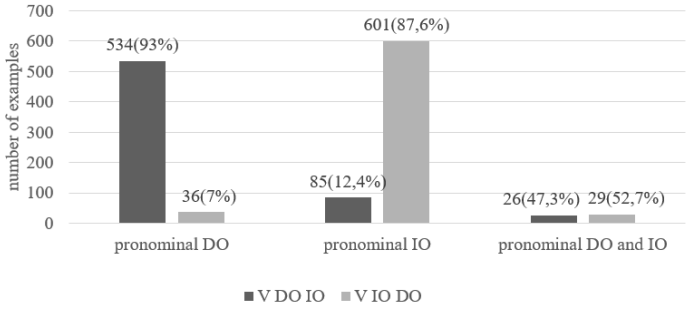


Figure 1: Word order in the sentences with pronominal objects.

After filtering out preverbal and pronominal objects and non-dative IOs without a preposition we had the final dataset containing 3017 contexts that we used for further analysis. In this dataset, grammatical features of the verb and the objects were annotated (including animacy that is especially relevant for the current study), as well as some additional features like the syntactic depth of the objects (the maximal distance from each node in an object subtree to its head, where the distance of a head-dependent pair is 1) and their length in words.

3.2 Analysis and discussion

The distribution of word orders in the final dataset is given in Table 2. The difference between the two subsets, with dative IOs and prepositional IOs, is statistically significant (χ^2=147,71, p<0,001). In the former, there is a modest, but statistically significant predominance of the 'IO DO' word order (χ^2=4,65, p=0,036). In the latter, the 'DO IO' word order is much more frequent (χ^2=180,73, p<0,001).

[2]A similar algorithm was used for our initial searches in Taiga, but the relevant relations were 'obj' and 'iobj' or 'obl', respectively. The code is available at `https://github.com/bamaxi/dir-indir`.

Word order	V DO IO	V IO DO
Dative subset	313 (44,2%)	394 (55,8%)
Prepositional	subset 1598 (69,5%)	701 (30,5%)
Total	1911 (63,5%)	1095 (36,5%)

Table 2: Numbers and percentages of sentences with different word orders in the final dataset.

The factors of interest. Now let us estimate the role of different factors that could influence this distribution: the length and syntactic depth of the objects and their animacy. The main problem for the analysis is that these factors are not balanced in the two subsets and correlate with each other. Table 3 illustrates this showing the average length and depth for animate and inanimate IOs and DOs in the two subsets. In particular, inanimate objects (direct and indirect, in both subsets) have higher average length and depth than animate ones.

		Average depth		Average length	
		Animate	Inanimate	Animate	Inanimate
Prepositional subset	IO	2,1	2,4	3,6	4,2
	DO	1,0	1,3	2,5	3,0
Dative subset	IO	0,9	1,3	2,3	2,9
	DO	1,3	2,1	3,3	4,2

Table 3: Average length and depths of animate and inanimate objects.

Statistical analysis reveals a very strong correlation between length and depth (ρ=0,92, p<0,001 for DOs; ρ=0,92, p<0,001 for IOs). For this reason, we will use only one of these factors in some further analyses. To estimate their correlation with animacy, we chose an arbitrary grouping for the variable 'length' (one word, 2–4 words, more than four words) and demonstrated that the differences between the resulting length and animacy groups are weak, but significant (χ^2=161,67, p<0,001, Cramer's V = 0,16[3]). This is shown in more detail on the mosaic plot in Figure 2 (red indicates that the observed values for the group are significantly larger than expected, blue indicates that they are significantly lower than expected).

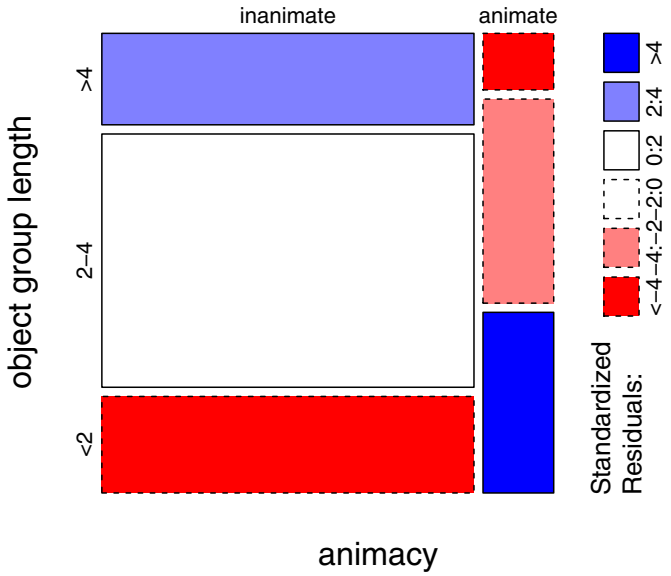
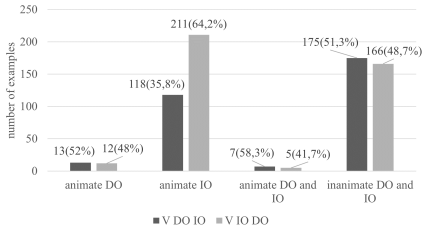


Figure 2: Mosaic plot showing the correlation between animacy and length.

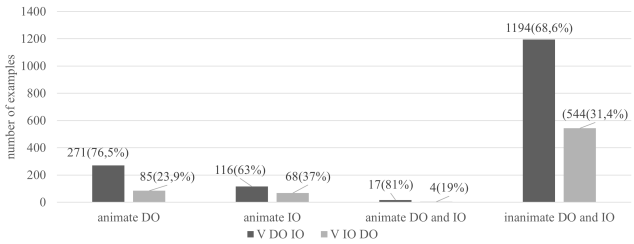
[3]Cramer's V is used to estimate the effect size (V>0,5 means a large effect, 0,3–0,5 is medium, 0,1–0,3 is small [26]).

The reasons why length and depth are closely connected are obvious, while their connection to animacy may be indirect. We mentioned earlier that animate arguments tend to be more salient. We also know independently that more salient arguments tend to have shorter descriptions (e.g. [2, 41, 33]) Below, we will first look at animacy, then at length and depth, and finally will consider all factors together in a regression model.

Animacy. Let us first consider the differences between the two subsets. In the prepositional subset, the majority of objects are inanimate: 2094 (91%) IOs and 1922 (84%) DOs. In the dative subset, an even larger share of DOs are inanimate (670, or 95%), but only slightly more than half of the IOs are (366, or 52%). The distribution of word orders depending on the animacy of the objects is shown in Figures 3a and 3b.



(a) Word order and animacy in the dative subset.



(b) Word order and animacy in the prepositional subset.

Figure 3: Numeric data on word orders as dependent on animacy for the two subsets

As Figures 3a and 3b make clear, the number of sentences in which both objects are animate is too small in both subsets to make any conclusions. When both objects are inanimate, there is a clear preference for the 'DO IO' order in the prepositional subset. When either IO or DO is animate, this is still the preferred order, although its share changes (according to Cramer's V measure, the effect is too weak: V=0,07).

In the dative subset, sentences with two inanimate objects have an equal distribution of the two word orders. It changes significantly when only the IO is animate (χ^2=17,85, p<0,001, V=0,16), which is illustrated by the mosaic plot in Figure 4 (examples in which only DO is animate are rare). Thus, Titov's [40] prediction that the 'DO IO' order will prevail once animacy is balanced was not supported, but we confirmed her intuition that the overall distribution of word orders is influenced by animacy (by the higher frequency of animate dative IOs and the tendency of animate objects to precede inanimate ones).

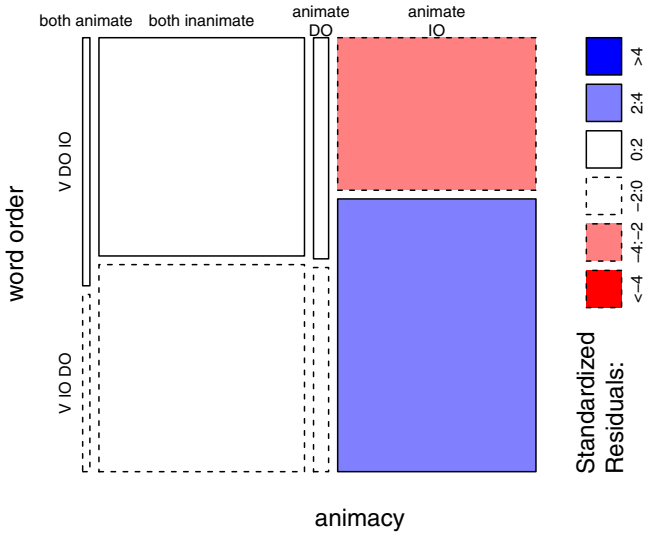
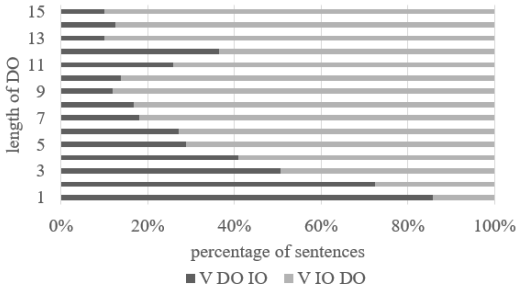
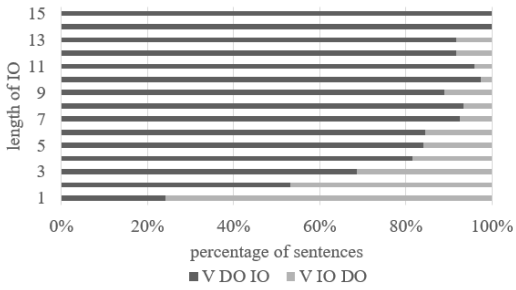


Figure 4: Mosaic plot showing the correlation between animacy and word order in the dative subset.

Length and syntactic depth. Firstly, let us note some differences between the two subsets. As Table

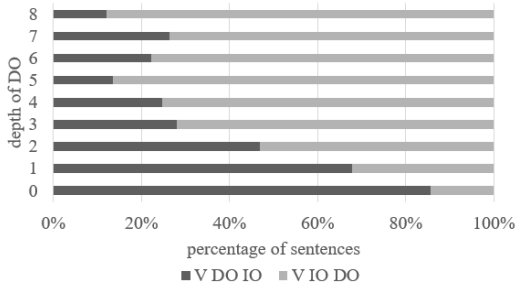


(a) Word order and the length of DOs.

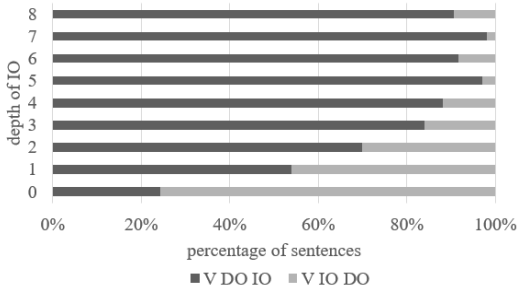


(b) Word order and the length of IOs.

Figure 5: Plots of word order as dependents on objects' lengths



(a) Word order and the syntactic depth of DOs.



(b) Word order and the syntactic depth of IOs.

Figure 6: Plots of word order as dependent on objects' syntactic depths

3 above shows, IOs (both animate and inanimate) have higher average length and depth than DOs in the prepositional subset, while the opposite is true for the dative subset. This might have to do with a different distribution of semantic roles in these two subsets and the relative salience of these roles. As for the influence of these two factors on the choice of the object order, we will first illustrate it and then will estimate it statistically in a regression model. As Figures 5a and 5b show, shorter objects tend to precede longer ones.

The syntactic depth of the objects has a similar effect on the word order. This is illustrated in Figures 6a and 6b. The more syntactically complex the object is, the higher its probability to follow the other object.

Regression model. To analyze different parameters together we performed a regression analysis fitting a binomial mixed effect model (logistic regression) on the whole dataset. We used the lme4 package [15] in the R software [32]. Verb lemmas were treated as the random effect. The fixed effect variables were the IO type (dative, coded as 1, or prepositional, coded as 0), the length of the IO and DO and their animacy (animate, coded as 1, or inanimate, coded as 0), as well as two interaction terms: 'IO length * animacy' and 'DO length * animacy'. Table 4 presents the outputs of the model.

	β	SE	z value	p-value
IO type: dative	0,73	0,19	3,83	<0,001
IO length	-2,12	0,18	-12,05	<0,001
DO length	1,80	0,11	15,80	<0,001
IO animacy: animate	0,39	0,20	2,00	0,049
DO animacy: animate	-0,25	0,20	-1,27	0,203
IO length * animacy	0,32	0,34	0,95	0,341
DO length * animacy	0,65	0,39	1,68	0,092

Table 4: The outputs of the regression model.

The model shows that the two subsets are significantly different and that DO and IO lengths are highly

significant factors. The effect of the IO animacy is much weaker, while the DO animacy factor did not reach significance (this is consistent with the results we got analyzing animacy in isolation). Discussing the role of animacy in ditransitive constructions, Titov [40] predicted it to play a moderate role. What has not been predicted either in this or in any other study is the crucial role of length and depth, although these results are not unexpected from a cross-linguistic perspective. Examples with prepositional IOs received very little attention in general, so the difference between them and the sentences with dative IOs has not been discussed for Russian as well.

Individual verbs. Finally, we estimated the role of individual verbs because Boneh and Nash [6] showed that different verbs have different preferences. We selected verb lemmas with more than 20 examples in our dataset and conducted a logistic regression analysis like above. Verb lemmas were treated as fixed effects. The results are presented in Figure 7: it shows β coefficients and standard errors for every verb (the ones that are significantly different from average are shown in red).

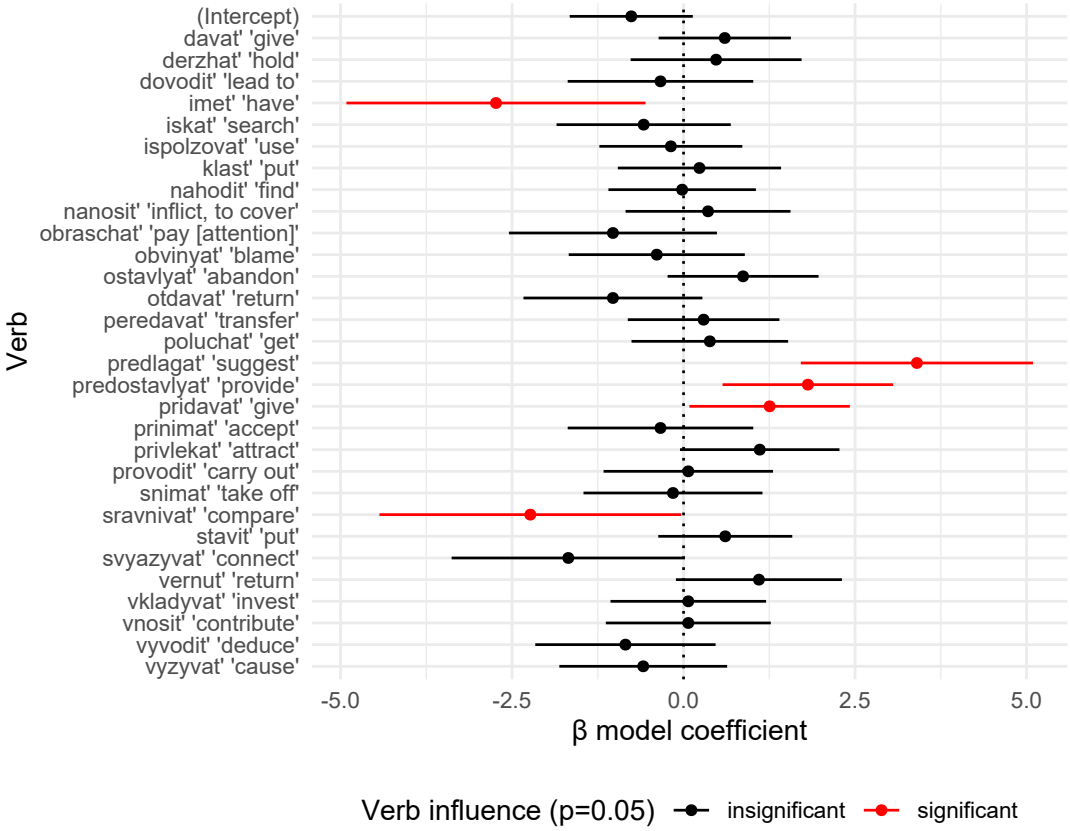


Figure 7: Word order: individual verb preferences.

Two verbs, *predlagat'* 'to offer' and *predostavlyat'* 'to provide', have a significant preference of the 'IO DO' word order. They are almost exclusively used with an animate dative IO and an inanimate DO. The opposite is true for three verbs: *imet'* 'to have', *poluchat'* 'to receive' and *sravnivat'* 'to compare'. They are used with prepositional IOs. The majority of contexts with *imet'* 'to have' involves an idiomatic expression *imet' delo s kem-to* 'to deal with somebody', so the DO is not only very short, but also forms a whole with the verb semantically.

4 Conclusions

The order of arguments with ditransitive verbs in Russian has been extensively discussed in the literature, especially in the generative tradition. However, most studies relied on the analysis of individual examples and did not take into account many factors that were identified as relevant cross-linguistically. To close this gap, we conducted a corpus study and analyzed such factors as animacy, length and syntactic depth of

the objects, as well as the role of individual verbs. We showed that these factors correlate with each other (especially length and depth) and that length and depth, which have not been discussed in the previous studies on Russian, heavily influence the choice of the word order, while animacy plays a moderate role.

We showed that for prepositional IOs, which have not received enough attention in the previous studies, 'DO IO' is definitely the neutral word order. In the sentences with dative IOs, the 'IO DO' order is more frequent, so it qualifies as basic in the functional approach. In the formal tradition, the basic word order is not necessarily the most frequent — it is the order that is used when all relevant factors are balanced. Analyzing sentences with two inanimate objects does not lead to a definitive answer: the shares of the 'DO IO' and 'IO DO' orders are virtually the same in them. Thus, even if the basic word order can be established using some syntactic tests — e.g. asymmetries in reciprocal and variable binding and scope taking, as Bailyn [4, 5], suggests — one would still have to explain this result: why don't we see the prevalence of this order? This could be explained by different preferences of individual verbs, as Boneh and Nash [6] suggested, but we did not see any clear evidence for that in our corpus data. We plan to explore these finings in our further research.

Word order distribution in the sentences with dative and prepositional IOs that we found can be associated with other properties of these sentences. In particular, animate IOs are much more frequent than animate DOs in the former, while the opposite is true for the latter (although the difference is much smaller). In the former, IOs tend to be shorter and less complex than DOs, while the reverse picture is found in the latter. These observations may be connected to the properties of semantic roles typically assigned to dative and prepositional IOs. We are going to study this connection in more detail in our subsequent work.

References

[1] Annotated corpus of Russian texts: concept, annotation tools, informations types [Annotirovanniy korpus russkih tekstov: kontseptsiya, instrumenty razmetki, tipy informatsii] / I. Boguslavskiy, N. Grigoriev, S. Grigorieva et al. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2000". — Moscow : Russian State University for the Humanities, 2000. — P. 106—118.

[2] Ariel M. Accessing noun-phrase antecedents. — London : Routledge, 1990.

[3] Baylin J. A configurational approach to Russian 'free' word order : Ph. D. thesis / J. Baylin ; Cornell University. — Ithaca, NY, 1995.

[4] Baylin J. What's inside VP? New (and old) evidence from Russian // Formal Approaches to Slavic Linguistics 18. — Ann Arbor, MI : Michigan Slavic Publications, 2010. — P. 21—37.

[5] Baylin J. The syntax of Russian. — Cambridge : Cambridge University Press, 2012.

[6] Boneh N., Nash L. The syntax and semantics of dative DPs in Russian ditransitives // Natural Language & Linguistic Theory. — 2017. — Vol. 35, no. 4. — P. 899–953. — Access mode: https://doi.org/10.1007/s11049-017-9360-5.

[7] Bonnot Ch. I., Fougeron I. Is the initial sentence accent in Russian always a sign of expressiveness or familiarity[L'accent de phrase initial en russe est-il toujours un signe d'expressivité ou de familiarité] // Bulletin de la Société de Linguistique de Paris. — 1982. — Vol. 87. — P. 309–330.

[8] Bruening B. QR obeys Superiority: Frozen scope and ACD // Linguistic Inquiry. — 2001. — Vol. 32. — P. 233–273.

[9] Configuring topic and focus in Russian / Ed. by T. H. King. — Stanford, CA : CSLI Publications, 1995.

[10] Cuervo C. Datives at large : Ph. D. thesis / C. Cuervo ; MIT. — Cambridge, MA, 2003.

[11] Dyakonova M. Russian double object constructions // ACLC Working Papers. — 2007. — Vol. 2. — P. 3–30.

[12] Dyakonova M. A phase-based approach to Russian free word order : Ph. D. thesis / M. Dyakonova ; University of Amsterdam. — Amsterdam, 2009.

[13] Faghiri P., Samvelian P. Constituent Ordering in Persian and the Weight Factor // Empirical Issues in Syntax and Semantics 10. — 2014. — P. 215–232.

[14] Faltz L. M. On indirect objects in universal syntax // Chicago Linguistic Society 14. — 1978. — P. 76–87.

[15] Fitting linear mixed-effects models using *lme4* / D. Bates, M. Mächler, B. Bolker, Walker S. // Journal of Statistical Software. — 2015. — Vol. 67. — P. 1–48.

[16] Haider H. *Mittelfeld* Phenomena: Scrambling in Germanic // The Wiley Blackwell Companion to Syntax, Second Edition. — Hoboken, NJ : Wiley Blackwell, 2017. — P. 1–73.

[17] Heaviness vs. newness: The effects of complexity and information structure on constituent ordering / J.E. Arnold, T. Wasow, A. Losongco, R. Ginstrom // Language. — 2000. — Vol. 76. — P. 28–55.

[18] Heine K., König Ch. On the linear order of ditransitive objects // Language Sciences. — 2010. — Vol. 32. — P. 87–131.

[19] Junghanns U., Zybatow G. Syntax and information structure of Russian clauses // Formal Approaches to Slavic Linguistics 4. — Ann Arbor, MI : Michigan Slavic Publications, 1995. — P. 289––319.

[20] Kempen G., Harbusch K. A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment // Multidisciplinary Approaches to Language Production. — Berlin : De Gruyter Mouton, 2004. — P. 87–116.

[21] Kittilä S. Object-, animacy- and role-based strategies: a typology of object marking // Studies in Language. — 2006. — Vol. 30, no. 1. — P. 1–32.

[22] Kodzasov S.V. On the accent structure of the constituents [Ob akcentnoj strukture sostavlyayushchih] // Experimental phonetic analysis of speech, Vol 2 [Eksperimental'no-foneticheskiy analiz rechi, T. 2]. — Saint-Petersburg : LGU, 1989. — P. 122–127.

[23] Kodzasov S.V. Combinatorial model of phrasal prosody [Kombinatornaya model' frazovoy prosodii] // Prosodic structure of Russian speech [Prosodicheskiy stroy russkoy rechi]. — Moscow : Russian Language Institute of the Russian Academy of Sciences, 1996. — P. 85–123.

[24] Kovtunova I. Modern Russian. Word order and information structure [Sovremenniy russkiy jazyk. Poryadok slov i aktualnoe chlenenie predlozhenija]. — Moscow : Prosveschenie, 1976.

[25] Larson R. K. On the double object construction // Linguistic Inquiry. — 1988. — Vol. 19. — P. 335–391.

[26] Mangiafico S.S. Summary and analysis of extension program evaluation in R, version 1.18.1. — Acess mode: `http://rcompanion.org/handbook/`. — 2016.

[27] Mykhaylyk R. Optional object scrambling in child and adult Ukrainian : Ph. D. thesis / R. Mykhaylyk ; Stony Brook University. — New York, 2001.

[28] Mykhaylyk R. Middle Object Scrambling // Journal of Slavic Linguistics. — 2011. — Vol. 19, no. 2. — P. 231–272.

[29] Neeleman A., van de Koot H. Dutch scrambling and the nature of discourse templates // Journal of Comparative Germanic Linguistics. — 2008. — Vol. 11. — P. 137–189.

[30] Pereltsvaig A. Syntax of denominal and ditransitive verbs reconsidered. — Sheffield : University of Sheffield, 2003.

[31] Pylkkänen L. Introducing arguments. — Cambridge, MA. : MIT Press, 2008.

[32] R Core Team. R: A language and environment for statistical computing. — Acess mode: `http://www.R-project.org/`. — 2013.

[33] Reference in discourse / Ed. by A. A. Kibrik. — Oxford : Oxford University Press, 2011.

[34] Russian grammar. Vol. 2: Syntax [Russkaya grammatika, T. 2: Sintaksis] / Ed. by N. Shvedova. — Moscow : Nauka, 1980.

[35] Shavrina T., Shapovalova O. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. — Saint-Petersburg : CORPORA 2017, 2017. — P. 78–84.

[36] SinTagRus today[SinTagRus segodnya] / E.S. Inshakova, L.L. Iomdin, L.G. Mityushin et al. // Proceedings of the V.V. Vinogradov Institute of the Russian Language[Trudy Instituta russkogo yazyka im. V.V. Vinogradova]. — 2019. — P. 14–40.

[37] Slioussar N. Grammar and information structure: A study with reference to Russian : Ph. D. thesis / N. Slioussar ; Utrecht University. — Utrecht, 2007.

[38] Soschen A. Derivation by phase: Russian applicatives // 2005 Canadian Linguistic Association Annual Conference. — Michigan Slavic Publications, 2005.

[39] Titov E. Information structure of argument order alternations : Ph. D. thesis / E. Titov ; University College London. — London, 2012.

[40] Titov E. The canonical order of Russian objects // Linguistic Inquiry. — 2017. — Vol. 48. — P. 427–457.

[41] Topic Continuity in discourse: a quantitative cross-language study / Ed. by T. Givón. — Amsterdam : John Benjamins Publishing, 1983.

[42] Velnić M. The influence of animacy, givenness, and focus on object order in Croatian ditransitives // Studia Linguistica. — 2019. — Vol. 73. — P. 175–201.

[43] Wasow T. Remarks on grammatical weight // Language Variation and Change. — 1997. — Vol. 9. — P. 81–105.

[44] Wasow T., Arnold J. Post-verbal constituent ordering in English // Topics in English Linguistics. — 2003. — Vol. 43. — P. 119–154.

[45] Word order in Russian [Poryadok slov v russkom jazyke] / Ed. by O. Sirotinina. — Saratov : Saratov State University, 1965.

[46] Yanko T. E. Communicative structure of declarative sentences with verb preposition [Kommunikativnaya struktura povestvovatel'nykh predlozheniy s prepozitsiyey glagola] // Language and culture. Facts and values. — Moscow : Languages of Slavic culture, 2001. — P. 371–383.

Corpus regional lexicography: principles, methods, and preliminary results

Belikov V. I.
MIPT, ABBYY Lab
vibelikov@gmail.com

Dubyaga A. O.
RSUH
dubiaga.al@gmail.com

Rvanova L. Y.
MIPT, ABBYY Lab
rvanova.lyu@phystech.edu

Selegey V. P.
ABBYY
vladimir_s@abbyy.ru

Abstract

The article summarizes the results of the long-term project “Languages of Russian Cities” (LoRC) of the regional vocabulary collecting and researching, which, unfortunately, was not depicted in any academic publications for a number of reasons. About 4 thousand pieces of regional materials were collected, systematized, and became the basis of the typology of regional differences consideration and the concept of a regional norm discussion. Reliability issues and methods of computer-based regional corpus research, including automatic text classification and author profiling, are paid attention to. Along with this article, the "reincarnation" of the LoRC project is also returning to the fund of open lexicographic resources basing on the joint portal for distinctive sociolinguistic research, which includes the General Web-corpus of Russian Language and the interactive dictionary “Languages of Cities and People” (LoC&P)

Keywords: WAC; regional diversity; regionalism; regionally biased vocabulary; regional norm; automatic regional classification; author profiling

DOI: 10.28995/2075-7182-2021-20-79-93

Корпусная региональная лексикография: принципы, методы и предварительные результаты

Беликов В. И.
МФТИ, ABBYY Lab
vibelikov@gmail.com

Дубяга А. О.
РГГУ
dubiaga.al@gmail.com

Рванова Л. Ю.
МФТИ, ABBYY Lab
rvanova.lyu@phystech.edu

Селегей В. П.
ABBYY
vladimir_s@abbyy.ru

Аннотация

В статье подводятся итоги многолетнего проекта «Языки Русских Городов» (ЯРГ) по сбору и исследованию региональной лексики, который, к сожалению, не был «финализирован» по ряду причин в виде академических публикаций. Был собран и систематизирован значительный (ок. 4 тыс. единиц) региональный материал, на базе которого рассматривается типология региональных различий, вводится/обсуждается понятие региональной нормы. Особое внимание уделяется вопросам надежности и методикам компьютерных региональных корпусных исследований, включая автоматическую классификацию текстов и профилирование авторов. Вместе с этой публикацией возвращается в фонд открытых лексикографических ресурсов и «реинкарнация» проекта ЯРГ – теперь на базе объединенного портала для дифференциальных социолингвистических исследований, включающего интернет-корпус ГИКРЯ и интерактивный словарь ЯГеЛь (Языки Городов и Людей).

Ключевые слова: WAC; региональная вариативность; регионализм; регионально смещенная лексика; региональная норма; автоматическая региональная классификация; авторское профилирование

1 Проект Языки Русских Городов: мотивы и итоги

Можно утверждать, что до старта проекта «Языки русских городов» реального представления о масштабах региональной вариативности в норме языка (см. далее о региональных нормах) не было. Обычная публика охотно принимала анекдоты в популярных изданиях про куру, гречу и поребрик, но за пределами московско-питерской темы существовали только отдельные плоды лексического краеведения, в которых не всегда проводилась грань между сельскими диалектами, топонимикой, автохтонной лексикой и собственно региональной вариативностью «городской» языковой нормы.

Можно сказать, что сама постановка исследовательского вопроса со смещением интереса с устного языка на письменный не была очевидной. В проекте впервые ставилась задача определить масштаб и типологию регионального варьирования РЯ на основании исследования текстов региональных СМИ и социальных сетей и опросов пользователей форумов Lingvo.

Реализация проекта «внутри» Lingvo community оказалось очень правильной идеей: в начале нулевых языковые форумы Lingvo были наиболее популярной дискуссионной площадкой по вопросам не только перевода, но и русского языка, с большим региональным разбросом участников, многие из которых были профессиональными переводчиками и/или филологами.

Технология была простой: участники обсуждений предлагали свои варианты, которые проверялись профессиональными редакторами с помощью базы региональных СМИ Интегрум и работавших на тот момент Яндекс.Блогов.

В результате нескольких лет функционирования форума был собран уникальный материал – несколько тысяч региональных слов (в разной стадии проверки) с высоким индексом цитирования в региональных СМИ и соцсетях, что позволило сделать вывод о глобальном характере проявления языковой вариативности в норме языка. Некоторым «апофеозом» этой деятельности стало издание словаря «Языки Русских Городов», который вошел в юбилейную версию системы Lingvo с номером «ХЗ». В этот словарь вошло около тысячи слов из числа обсуждавшихся на форуме. Форум послужил источником материала для значительного количества научных статей, там же разрабатывалась методика лексикографической работы с материалами интернета в целом [1; 5; 6; 7; 8]. Из работы на форуме вырос фундаментальный словарь неофициальной топонимии России и ближнего зарубежья [2].

К сожалению, проект ЯРГ прекратил активное существование из-за смены лексикографической политики команды Lingvo, которая отказалась от собственных лексикографических проектов и перешла на лицензирование контента. Но причина задержки с публикацией промежуточных итогов более глобальна — она отражает общее падение интереса к кропотливой лексикографической работе в угоду автоматизации, стремление к максимальному покрытию и скорости (дешевизне) получения результата в ущерб качеству. К сожалению, сегодня исчезает понятие «авторитетного» словаря даже для академических толковых словарей (хотя и раньше их авторитетность в случае региональной лексики не означала корректности, полноты и последовательности в описании).

Авторы надеются, что эту тенденцию не поздно еще изменить, причем не возвратом к старому, а за счет применения методов анализа больших корпусных данных, которые являются не только средством получения новых объектов описания, но и верификации этих описаний.

2 Типология региональной лексики

Анализ собранного материала позволяет сделать некоторые выводы. Регионализмы можно классифицировать по разным основаниям.

По происхождению:

- Из местных сельских диалектов (что не всегда легко подтверждается в силу малодоступности диалектных словарей; по СРНГ [16] часто не удается выявить диалектный ареал). Есть и экзотика. *Баской/баский* ‘красивый, хороший’ очень слабо представлено в городском узусе на Европейском севере и Сев. Урале, в Сибири — только в сельских диалектах. Но в 1990-х — начале 2000-х в молодежном жаргоне от Норильска до сев.-вост. Казахстана было (сейчас, вероятно, ушедшее) прил. *бащный* — ‘отличный’ — вероятно, от сравнительной степени баще ‘лучше’.

- Из распространенного в ареале нерусского языка, заимствование (*махалля* < узб. *mahalla*) или калька (*самориск* < латышск. *pašrisks*). А также расширение значения: *урюк* ‘дерево и его свежие плоды’ (среднеазиатский *урюк* производит впечатление метонимии, но это результат контактов с местными тюркскими языками).
- Заимствовано из местного неродственного языка, но в городскую речь проникало и из диалектов, и из самого языка — *калега* ‘брюква’ в Удмуртии.
- Свободно порождается системой языка (*башня/свечка* — одноподъездный «высокий» дом); *политсила* ‘организация, активно участвующая в политической жизни страны; партия, политический блок’ (Украина).
- Заимствовано из проф. узуса *точечный дом* (а этап *точечный дом* → *точка* — уже переработка системой языка).
- Неясно откуда (*мультифора* ‘файл для бумаг’, центральная Сибирь) — внутрення форма очевидна: излат. *multus* ‘многочисленный’ и *foro* ‘дырявить’, но пути появления этой номинации не ясны.

По фиксации в толковых словарях:

- Отсутствуют.
- Присутствуют параллельно с основным с неточным толкованием. Ср. в БТС [10] общее *сурок* (‘небольшое животное сем. беличьих, зимой впадающее в спячку’) и его синонимичные для незоолога именования *байбак* ‘степной грызун из рода сурков, осень и зиму проводящий в спячке’ и ‘грызун рода сурков, обитающий в Забайкалье, на Алтае, в Монголии и Северном Китае’.
- Присутствуют с косвенным указанием на региональность, ср. в [10]: «1. На Дальнем Востоке и в Сибири: небольшая гора с округлой вершиной, курган, холм. 2. На Камчатке и Курильских островах: вулкан».
- Присутствует с ошибочной стилистической пометой (обычно это слова со слабо выраженной региональностью). Ср. в [10]: «**ХОЛОДЕЦ**, -дца*; **м**. Нар.-разг.* 1. =Студень» (*студень* помет не имеет; в современном петербургском узусе, исключая самое старшее поколение, *холодец* частотнее *студня*).
- Присутствует, но региональность не маркируется. Ср. в [10] *банлон* (без помет), *водогрей* (без помет), *хабарик* (*жарг.*). При этом первое с такой фонетикой давно устарело, ср. реакцию петербургских блоггеров при обсуждении вариативности: *Слово «бонлон» я слышала разве что от папы* (1973 г. р.). *Нам банлон не нужен. Мы хотим бадлон!* (1978 г. р.).
- Некогда фиксировалось словарями в разных значениях, в том числе и региональном, а в современных словарях значение заужено: *толока* ‘безвозмездная общественная работа’ (Украина, Белоруссия, Эстония, Латвия) — значение, среди прочих, фиксировалось в словаре Ушакова [18]. В БТС [10] только явно устаревшее ‘поле под паром, используемое для выпаса скота; выпас скота на таком поле с целью удобрения почвы’, которое плохо стыкуется и с региональным, и с распространяемым Яндекс.Толокой: «Заработок в интернете. Простые задания за вознаграждения» (toloka.yandex.ru).

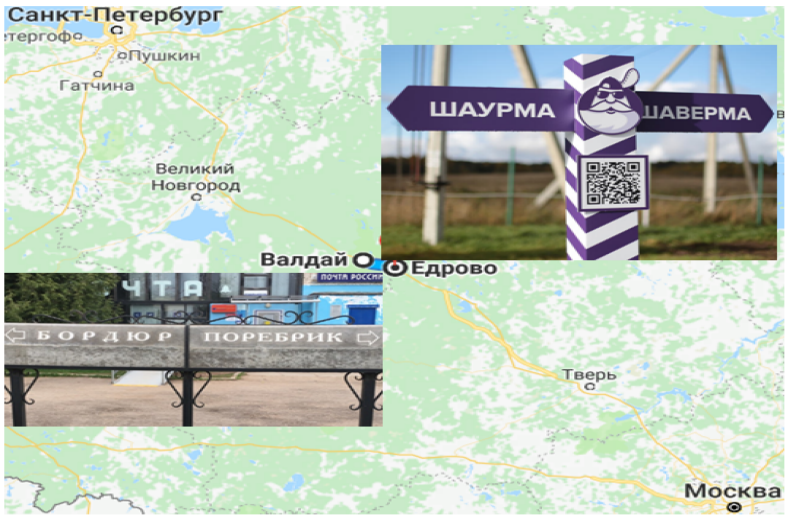
По связи с общенормативным:

- Синоним: *качок* ‘насос’, *лейка* ‘воронка’ (они же и омонимы).
- Как бы синоним: петерб. *латка* = «почти ‘утятница’», но *латка* может быть круглой.
- Как бы «метонимия»: *гардина* ‘карниз для занавесок’, *гамаши* ‘рейтузы’, ср.-аз. *урюк* ‘дерево и его свежие плоды’ (такой *урюк* производит впечатление метонимии, но это результат контактов с местными тюркскими языками)
- Омонимия: дальневосточн. *медведка* (в стандарте — живущее в почве насекомое *Gryllotalpa gryllotalpa*) и *чилим* (в стандарте — *Trapa natans*, водяной орех) — крупные креветки.

По осознаваемости носителями:

- Специфика *массово* не осознается.
- Специфика общеизвестна и в ареале, и за его пределами.

В силу слабости межрегиональных контактов традиционно таких слов было мало, вне ареала слово становилось узнаваемым через беллетристику (*изба/хата*) или из фактов личной биографии (*цветет урюк* у Солженицына; то же в «Золотом теленке» — *Цветет урюк под грохот дней / Дрожит зарей кишлак* … вне региона, вероятно, массово воспринималось как «плохие стихи»).
Интернет породил «мемические» знания, часто неточные: указатель «Шаурма» в Едрово — в сторону Тверь→Москва, но в Твери *шава́рма*):



- Специфика общеизвестна в ареале, служит своеобразным шибболетом. Популяризации такого знания в пределах ареала способствовали распространенные в 2000-х годах бродячие списки *Ты из* [название города], *если* [следует длинный список местных реалий и элементов языковой специфики], а также публикации типа «Только у нас известны слова …»; фактически ареал может быть заметно шире «заявленного» (*грядушка* [спинка] *кровати* — якобы чисто воронежское, но шире используется в Волгоградской обл.) или ареал может быть не единственным (*отсветка* ‘ксерокопия’ — якобы только Амурская обл., но также и Литва). **Вырожденный случай** — поморские *баско* и *порато* в Архангельске (повседневно слова не используются, имеют лишь символическое значение):

Архангельск: парикмахерская и магазин головных уборов

По границам ареала:

- Точечное: *явочка* ‘талон к врачу’ (Ярославль, но не Ярославская область)
- Ареал большой; граница может быть четкой и размытой (образуется периферия ареала): (*по*)*ставить укол/прививку*: западная граница четкая, восточной как бы и нет, но на Дальнем Востоке и в Казахстане *делать укол* сильно частотнее, чем *ставить*.

- Ареал сравнительно небольшой (одна-две-три области или сопредельные части областей), граница четкая: *на зеленую / на зеленой* — о выезде «на природу» с рекреационными целями (Ярославская, Костромская, северо-восток Ивановской обл.). Госграница — не помеха *фыгать* ‘курить анашу’ (только Оренбургская и Актюбинская обл.), граница ареала четкая.
- Территориально далекие независимые ареалы (во всяком случае связь не прослеживается): *отсветка* ‘ксерокопия’: Литва и Амурская обл., *лента* ‘сдвоенный академический час в вузе, «пара»’: Днепропетровск (ныне *Днепр*) и частично соседи, а также Красноярск;
- Точные границы ареала трудно определить. *баллон* ‘трехлитровая банка’ — «южное»: от Украины до Средней Азии, но детали ареала (ареалов) неясны. Примерно такова же ареальность *газводы*, но здесь явно в соответствии с надписями на автоматах *Газвода* vs. *Газированная вода*, — где производились и куда поставлялись те и другие сейчас вряд ли можно установить.
- По географии нас интересует постсоветское пространство. Но специфику языка дальнего зарубежья тоже необходимо учитывать, если она явлена в беллетристике или публицистике, широко читаемой в России (в частности, в российских изданиях). В беллетристике Журнального зала неоднократно встречаются загадочные для многих слова *хенди/хэнди* (Германия), *пелефон* (Израиль) ‘мобильник’, которые не всегда ясны из контекста[1]. То же касается и необычного управления, ср. *на интернете* вм. *в интернете* (характерно для США и, кажется, Франции).

По типу «пользователя»:

- Общее разной стилистики. Есть заведомо **нормативное**, использующееся в местном законодательстве: *углярка* ‘угольный сарай’ (Кемерово), судебной практике: *отбивать* (кассовый) *чек* (сев. Урал)[2]; *хабарик* ‘окурок’ (СПб, разговорное), *чибон* ‘окурок’ (Пермь, сленг = «общий жаргон», стилистически гораздо ниже петербургского хабарика).
- Детское. Тут сложно говорить о региональности, поскольку детская лексика плохо фиксируется словарями, то есть «общая кодифицированная норма» в этой сфере неполна (нет, например, явно общих *бибики* и *бибикать*, при том что слово не сугубо детское: есть цитата из СМИ, как Путин, ехавший на Афон, *бибикал ослу*, взрослые собственный автомобиль нередко зовут *бибикой*).
- Жаргон. В первую очередь речь о молодежном жаргоне, хотя понятие «молодежный» довольно туманно. Если потребность в конкретной номинации с возрастом не затухает, то «бывшая молодежь» продолжает его использовать, переходя в средний возраст. Жаргонное, естественно, может быть необщеизвестным в ареале: *зимбура* (Мурманск) «знают автомобилисты и бомжи».
- Профессиональное: *туалет прямого падения* — используется в системе архангельского ЖКХ, поэтому знают его все местные жители; *опанелка* (дверная обналичка, пиломатериал) — непрофессионалы знают, только если сталкивались (как и «московскую» *обналичку*).
- Сугубо административное, вряд ли встречающееся в повседневном узусе: *освободить от транспорта* ‘закрыть проезд на определенных улицах’ (Мурманск). Типы поселений при

[1] И так мы ходили вчера и позавчера, пели “Христос воскресе из мертвых” под писк хэнди, окутанные густым запахом духов, который не может выветрить из нашего крестного хода никакой ветер (Михаил Шишкин. Взятие Измаила // Знамя, № 12, 1999)

[2] По данным базы СМИ «Интегрум» в документах Федерального арбитражного суда Московского округа **Oper** от (*кассовый*) *чек* представлен глаголами *пробивать* (94%) и *выбивать* (6%), ясно, что норма здесь *пробивать*, но допустимо и *выбивать*; tertium non datur. Похожая ситуация и в документах ФАС Северо-западного округа, соответственно, 90% и 10%. Совершенно иное положение в ФАС Уральского округа, где в этом контексте преобладает *отбивать* — 54% (*пробивать* — 35%, *выбивать* — 11%). Преобладание возникает в основном за счет документов северной части Уральского округа — в нижестоящем 17 арбитражном апелляционном суде (Удмуртия, Пермский край, Свердловская область) *отбивать* дает 67% таких контекстов, а на юге, в 18 арбитр. апелл. суде (Башкирия, Челябинская, Оренбургская, Курганская обл.), только 40%. Разумеется, эта практика не кодифицирована ни в толковых словарях, ни в законодательных документах. Основывается она на региональном узусе, ядро ареала которого на севере, а юг Урала — периферия. ***Отбивание** кассовых чеков* противоречит московской и петербургской норме, но **для Урала это норма**, de facto закрепляемая судебными документами.

железнодорожных разъездах официально именуются по-разному. Судя по территориальному распределению, определение типа поселения зависело от региональной администрации (наименование типа поселения приводятся по ОКТМО). есть, *разъезд* и *железнодорожный разъезд* — это разные типы поселений. В одних областях представлены только *разъезды* (в Вологодской обл. их 12, в Пензенской — 8, в Ростовской — 10, в Ульяновской — 19), в других — только *железнодорожные разъезды* (в Волгоградской — 10, в Кировской — 10, в Саратовской —18). В Мордовии есть только более редкий вариант *поселок разъезд*, их 16. В Свердловской области только два поселения-разъезда, но именуются они сложно: *поселок при железнодорожном разъезде 99 км* и *поселок при железнодорожном разъезде 136 км*.

Распространенность по времени:

- Пришло-ушло (обычно жаргонное, но к уходу может подталкивать не только смена моды на экспрессивные слова, но и утрата актуальности: *сотыга* ‘100 рублей’, сибирское).
- Историзм: *мосовская машина* (от советских автомобильных номеров с буквами *МОС* (с 1959 г.) для «больших начальников») — есть литературные примеры, где контекст ничего не разъясняет.
- Утрачивающаяся номинация. Известное с 1970-х гг. петерб. *лабаз* ‘магазин, обычно винный’, судя по возрасту использующих слово, устаревает; при этом в более младших возрастах значение расширяется: ‘любой магазин’.

По степени опознаваемости неносителем:

- от «**неправильно**» (*сайка чёрного* ‘буханка’, Волгоград) и «понятно, но **неграмотно**» (*ставить укол*, *отбивать чек*) через «ясно, что такое» (*красноголовик* ‘подосиновик’) и «в контексте **несложно догадаться**» (*пастик* ‘стержень шариковой ручки’, *химица*, *математица* и т. п. ‘учительница химии’ и т. п.) до **ошибочного понимания** (*медведки к пиву*, *бегал на морозе в одном гольфе* (=водолазке) и **непонимания** (*поместить рисунок в мультифору*, *мосовская машина*).
- Стандартные для Удмуртии примеры типа *Мой отец — удмурт, а мама — русская, из потомственных дворян, и я очень жалею, что не знаю удмуртского языка ‹…› Приедешь, бывало, на ферму, колхозники толкуют о своих проблемах, а ты хоть **не** толкай соседа: «О чем речь?»* (Председатель Союза журналистов Удмуртии Людмила Прокошева, в интервью) в печатном виде, вероятно, будут поняты правильно, но получат реакцию вроде «при редактировании текста забыли убрать ***не***».

3 Общеязыковая и региональная норма

Теоретические взгляды на то, что следует считать литературной нормой, изменчивы, меняется языковая ситуация, положение русского языка в мире, эволюционирует и сам язык. Поэтому представляется естественным регулярно возвращаться к тому, что́ именно в языке подлежит нормированию, какие объективные и субъективные препятствия встают на этом пути, насколько строгими могут и должны быть нормы в отношении разных аспектов такого сложного и во многом все еще непознанного феномена, каким является язык.

Мы исходим из концепции В. А. Ицковича, понимавшего под нормой «комплекс закрепленных речевой практикой языковых средств и закономерностей их реализации, объективно существующие в данное время в данном языковом коллективе» [12: 8]. По Ицковичу, норма представлена двумя ипостасями: «Имплицитно норма выступает в виде образца или, точнее, текстов, считаемых образцовыми ‹…› Эксплицитно, в явном виде, сформулированной, норма предстает перед носителями языка в кодификации, отражающей представление авторов грамматических пособий и словарей о языковой норме. Кодификация — это фиксация объективно существующей языковой нормы, сформулированная в виде правил (предписаний) в авторитетном лингвистическом издании (типа грамматики, учебника, словаря) и адресованная всем членам языкового коллектива» [12: 11–12].

Следует отметить, что норму кодифицируют не только «авторитетные лингвистические издания». Есть **терминологические словари**, которые содержат вполне адекватное описание профессиональной нормы для некоторых научных дисциплин и хозяйственных отраслей; но не для всех: вполне очевидно, что не всякий лингвист сочтет любой словарь лингвистических терминов (общий или специализированный, например, словарь социолингвистических терминов) полностью адекватным описанием лингвистической терминологии.

Есть официально утвержденные **стандарты**. Стандартизация продукции, а значит, и унификация терминологии необходима практически во всех производственных отраслях. Специализированный орган, занимающийся стандартизацией в рамках всего государства, существовал в СССР с 1925 г., именование его менялось. В РФ с 2004 года стандартизация находится в ведении Федерального агентства по техническому регулированию и метрологии (Росстандарт). Стандартизируется п р о д у к ц и я, но описание этой продукции кодифицирует с л о в о у п о т р е б л е н и е. Стандарты бывают отраслевыми, государственными и межгосударственными (в СНГ), а также международными (СССР, а затем Россия — член ISO/ИСО). Далеко не вся кодифицированная таким способом лексика имеет узкоспециальный характер. Есть «простые» слова, по разным (обычно загадочным) причинам игнорируемые толковыми словарями, но давно четко зафиксированные в ГОСТах. Ограничимся одним примером: ни в одном толковом словаре нет свиной *рульки*, широко продающейся в сыром и копченом виде. Давным-давно гостирован разруб свиной туши, где фигурирует и *рулька*, на производство и упаковку готовых мясных изделий также существуют стандарты, эти обязательные к исполнению общегосударственные документы можно считать кодифицирующими именование понятия «рулька».

Кроме ГОСТов, Росстандарт ответствен за разработку разнообразных **общероссийских классификаторов**. Есть «Общероссийский классификатор валют», где кодифицированы слабо отраженные толковыми словарями наименования денежных единиц постсоветского пространства[3]. Существует «Общероссийский классификатор территорий муниципальных образований» [15], где перечислены все поселения с указанием для каждого его типа, изредка неопределенно: *населенный пункт*, но обычно конкретизировано: *город*, *поселок*, *село*, *деревня*, *хутор*, *слобода*, *станция*, *погост*[4] и мн. др. Только что перечисленное — «обычные» русские слова, но в толковых словарях они могут объясняться неточно или не полностью. *Слобода* в БТС толкуется лишь как историзм («В России 11–17 вв. ‹…›»), у Шведовой в [20] есть также значение ‘посёлок около города, пригорода’, но помечено оно *устар*., *станция* — только пункт остановки транспорта, *погост* в обоих словарях только ‘кладбище’. В этом классификаторе закреплены и этнически маркированные типы поселений (188 *улусов* в Бурятии, 52 *аала* в Хакасии, 20 *арбанов* в Туве), словарями они толкуются либо неточно, либо игнорируются.

Есть **общие справочные издания**, в частности, энциклопедии. Адекватность сообщаемой там информации зависит от профессионализма авторов и единообразия интерпретации конкретной единицы на русскоязычном пространстве. Их основным назначением является сообщение энциклопедической информации, но вряд ли принятое в БСЭ (или в современной РСЭ) словоупотребление следует считать ненормативным узусом. А в подобных изданиях хватает общеизвестных лексических единиц, не попавших пока в толковые словари.

Есть **законодательство** и другого рода административные документы, при этом «на местах» и законотворцы, и те, кто ведет официальный документооборот, следуют местному узусу, создавая региональную норму (разнообразные примеры см. в [4]). Некоторые типы региональной документации требуют утверждения на общегосударственном уровне и, как следствие, местная норма de facto утверждается как общероссийская.

Слово *своротъ* ‘поворот (дороги, пути)’ в МАСе[13] получило помету *прост*., то есть в официальном речевом обиходе фигурировать не должно. Тем не менее оно регулярно используется в Иркутской области и Красноярском крае. Подготовленные там документы послужили основой

[3] Есть и занятное. Ко времени введения в действие современного классификатора (принят Постановлением Госстандарта России от 25 декабря 2000, действует 1.07.2001) в Таджикистане была введена денежная единица *сомони*, а в предыдущем классификаторе валют (действовал с 1.07.1995 по 1.07.2001) отразился переход в марте 1997 от именования *таджикский рубль* к таджикизированному написанию *таджикский рубл*. Это особенно примечательно, поскольку на советских рублях были и надписи на таджикском языке: *як сӯм* ‘один рубль’, *панҷ сӯм* ‘пять рублей’ и т. п.; независимый Таджикистан перешел от теоретического *сӯм* к реально функционировавшему и в советские времена *рубл*.

[4] Например, *погост Старая Никола* в составе Вахромеевского муниципального образования Камешковского района Владимирской обл. (код ОКТМО 17625408181).

«Паспорта инвестиционного проекта „Комплексное развитие Нижнего Приангарья“», утвержденного распоряжением Правительства РФ № 1708-р от 30.11.06, где среди «мероприятий, реализуемых в рамках проекта», упоминается *реконструкция автомобильной дороги Канск — Абан — Богучаны на участке Черемухово до* ***поворота*** *на Покатеево (км 124 — км 133) в Абанском районе Красноярского края*.

В нормированном языке юга Западной Сибири в том же значении используется слово *свороток*, последалевской общей толковой лексикографией вообще не фиксируемое. В Республике Алтай, Алтайском крае, Кемеровской области оно широко используется в кадастровой документации, в планах развития дорожного хозяйства и в других случаях, где необходимо упомянуть поворот на второстепенную дорогу. Археолог из Барнаула так указывает местоположение описываемого кургана: *приблизительно в 1,3 км к ЮВ от устья р. Куюм, напротив лесопилки, у своротка с Чемальского тракта к последней* (Степанова Н. Ф. Погребения в каменных ящиках и их датировка // Погребальный обряд древних племен Алтая. Барнаул, 1996, стр. 54).

Главным различием нормы и кодификации Ицкович считает вполне естественное запаздывание последней. Представляется, что куда важнее субъективность кодификаторов. Как и «рядовые» носители литературного языка, при определении нормативности «кодификаторы ориентируются в первую очередь на собственный узус, во вторую — на узус с в о е г о круга, но лишь настолько, насколько этот узус пассивно знаком самим лексикографам» [3: 361; там же см. разнородные примеры, подтверждающие этот тезис]. Незнакомые реалии толкуются по не всегда аккуратным источникам. Рыба, именуемая на международном языке биологов *Stenodus leucichthys*, на северных реках называется *нельмой*, а в бассейне Каспийского моря (где она практически исчезла) — *белорыбицей*. В словаре Шведовой [20] *нельма* толкуется как «крупная северная рыба сем. лососевых», а *белорыбица* — как «северная промысловая рыба сем. сиговых с серебристой блестящей чешуей»[5]; в петербургских словарях — не совсем так, но столь же ошибочно.

Трактовка фауны и флоры в толковых словарях ориентирована на научную картину мира, как видим, делается это не всегда аккуратно. А «биологически аккуратные» толкования часто плохо соотносятся с картиной мира образованного русскоязычного «обывателя», язык которого и должен быть отражен в словаре. Один из авторов настоящего текста многократно в разных аудиториях (лингвисты, школьные учителя, студенты-филологи) и разных регионах (от Воронежа до Благовещенска и от Петербурга до Волгограда) предъявлял изображения трех растений с вопросом, которое из них *камыш*. Обычно большинство указывало на *Typha latifolia*, который толковые словари вслед за ботаниками именуют «рогозом», с ним конкурировал *Phragmites australis* (словарный «тростник»), а «камыш» (*Scirpus lacustris*) иногда вообще никто не считал *камышом*. Показательно, что некогда популярный строительный материал *камышит* изготовлялся из того, что в словарях является тростником.

Определенная доля общерусской лексики в повседневном узусе столиц (и лексикографов) не встречается. Как кажется, ни в Ленинграде/Петербурге, ни в Москве — в отличие от большинства городов СССР — не функционировали *уличкомы*[6]; этого слова нет и в словарях. Между тем утвержденное в 1996 г. «Положение об уличных комитетах (уличкомах) г. Воронежа» попало в качестве типового образца в хрестоматию по муниципальному праву [14: 339–342]. В этом случае можно говорить о своеобразной **антирегиональности**: «везде» есть, но где-то не встречается.

* * *

Существование региональных различий в норме вполне очевидно. Очевидны и причины игнорирования этого факта официальной русистикой. В постсоветской истории это всего лишь традиция, а прежде была и идеология.

В первые послереволюционные годы во многом оказалось неизбежным расшатывание нормы: с одной стороны, в общегосударственный коммуникативный процесс вовлекались широкие массы населения, недостаточно владевшие нормативным языком, с другой стороны, менялся

[5] Для жителей Дагестана и Азербайджана это «восточная рыба», в Казахстане и Туркмении — «западная», а «северной» она оказывается лишь при взгляде из Ирана.

[6] В окраинной Москве *уличкомы* вполне могли существовать, в частности, на территориях, вошедших в черту города 17 августа 1960 (города Очаково, Кунцево, Тушино, Бабушкин, Перово, Люблино, села Медведково, Тропарево и мн. др.).

спектр функций литературного языка и охватываемая им проблематика. Происходило это одновременно по всей стране, так что о единообразии результатов не могло быть и речи[7]. С середины 1930-х гг. можно говорить об относительной стабилизации нормы, но в области лексики «новая» норма довольно заметно отличалась от «старой»[8].

Реальное осмысление сложившейся социолингвистической ситуации началось лишь с 1950-х гг. В немногих работах, посвященных региональным лексическим расхождениям в языке города, они интерпретировались как влияние местных диалектов и просторечия. Редким исключением оказалась статья Р. Р. Гельгардта «О литературном языке в географической проекции», справедливо утверждавшего, что «местные различия ‹…› литературного языка могут и не иметь источников в народной диалектной среде. Тогда они являются только вариантами литературной нормы» [11: 98].

Однако такие взгляды были признаны идеологически вредными: «старые» диалекты отмирают, любые их следы в городской речи — пережитки, «новых» различий, возникших в рамках литературного языка, в принципе не может быть, поскольку для них нет социальной базы. Возобладал декларативный тезис о полном единообразии русского литературного языка на всей территории его распространения, «общеобязательности его норм как образцовых для всех, кто им владеет и пользуется, независимо от социальной, профессиональной и территориальной принадлежности» [19: 3]. Писалось это за семь лет до предполагавшегося стирания классовых границ, а бесклассовому обществу положен монолитный язык.

Упоминание социальной принадлежности параллельно с профессиональной и территориальной указывает на довольно примитивное понимание *социального* всего лишь как *классового*, что характерно для далекой от социологии части научного сообщества, которая использовала собственные элементарные познания в этой сфере в административно-идеологической борьбе[9]. Профессионалам же известно, что любой город и другой населенный пункт представляет собой самостоятельный социальный организм, устроенный иногда сложно, иногда очень сложно. Каждый из них занимает собственное место в иерархически организованной системе «однотипных» социальных организмов — поселений. Однотипность тут условная, всякое поселение имеет свое лицо, определяемое многими показателями: историей образования, численностью населения, родом занятий жителей, местом в административной иерархии, физико- и экономикогеографическим положением, развитостью культурной среды, сетью учебных заведений и другими параметрами, вплоть до локальных мифологем. И было бы удивительно, если бы все это не находило отражения в лексиконе, в частности, в его нормативной части.

4 Технологии поиска и верификации регионализмов

Для задачи выявления региональной вариативности наиболее принципиальным вопросом является определение надежных регионально маркированных источников данных, причем в том количестве, которое позволяет выносить статистически значимые суждения. Стандартный «пайплайн», приводящий к появлению новых словарных входов регионального словаря, состоит из следующих этапов:

1. Поиск кандидатов на статус региональных нормированных вариантов значений, входящих в некоторый универсальный национальный Инвентарь Значений, сущности, которая, увы, в реальности никак не представлена. Далее мы будем называть **регионализмами** именно такие единицы, отличая их от топонимов и другой **регионально смещенной лексики**.
2. Определение достоверной региональной картины употребления регионализмов.
3. Лексикографическое описание.

Остановимся более подробно на каждом из этих этапов.

[7] По замечанию Р. О. Шор, в дореволюционном языке имелись значительные лексические лакуны, например, среди «терминов кухни и домашнего хозяйства». «Очевидно, что при отсутствии соответствующих слов в „литературном языке“ „образованные классы“ общества принуждены заимствовать их из народных говоров данной местности» [21: 137].

[8] Не случайно в словаре под ред. Д. Н. Ушакова появились пометы *новое* («слово или значение возникло в русском языке в эпохи мировой войны и революции») и *дореволюционное* («слово обозначает предмет или понятие, вытесненные послереволюционным бытом») [18: XXVII—XXVIII].

[9] Вообще-то такой взгляд заслуживает старой советской этикетки «вульгарный социологизм». В действительности социально в человеке все, что не обусловлено исключительно биологией.

4.1 Поиск кандидатов

Есть целый ряд проблем, типичных для современных социолингвистических исследований, препятствующих эффективному поиску таких объектов автоматически:

- малое число надежных полномасштабных источников региональных данных;
- проприетарность таких источников или существенные ограничения в их академическом использовании;
- ложная/неточная региональная атрибуция текстов и/или авторов, связанная, в частности, с принципиальным различием геометок, ассоциированных с местом текущего пребывания автора (геолокация) и местом его рождения. Заметим, что на идиолект могут действовать оба фактора, при этом первый является случайно смещающим реальную региональную картину.

На первом этапе проекта ЯРГ основным источником потенциальных регионализмов были предложения, сделанные участниками форумов Lingvo, в основном, профессиональными переводчиками, представляющими самые разные регионы России и Ближнего Зарубежья и привыкшими не только внимательно относиться к нюансам в употреблении слов, но и обсуждать эти нюансы с “peer-to-peer” коллегами. Это обусловило высокий КПД обсуждений, позволивший быстро набрать наиболее очевидные частотные регионализмы. Число таких регионализмов, несколько тысяч, оказалось сюрпризом как для донаторов (для которых сама идея наличия таких слов в их собственных идиолектах вовсе не была очевидной), так и для идеологов проекта.

Этот результат можно считать наиболее значительным, поскольку он показывает **реальность существования региональной вариативности** не только в узусе, но и **в норме** в статистически значимых объемах: вывод, с которым обязаны теперь считаться любые лексикографы, занимающиеся толковыми словарями русского языка.

С другой стороны, переход от наиболее частотных и очевидных регионализмов к менее частотным требует уже иных методов: увеличение числа участников обсуждений неизбежно приводит к падению среднего качества предложений, росту «фейковых» обсуждений и т.п. Это подвело естественную черту под первым этапом проекта ЯРГ.

Второй этап проекта начался с появлением корпуса ГИКРЯ, который позволяет не просто проверять региональное смещение запроса, но и проводить сплошную обработку регионально маркированных текстов в поисках кандидатов на регионализмы. Сразу скажем, что этот этап проекта еще не полностью реализован, и речь пойдет о тех подходах, которые активно исследуются, и о некоторых предварительных выводах из этих исследований.

Здесь следует немного отвлечься от основной темы статьи и коснуться некоторых побочных, но важных тем, связанных с оценкой качества регионально маркированных подкорпусов ГИКРЯ, особенностям распределения регионально окрашенной лексики и возможности ее использования для задач извлечения кандидатов и автоматической региональной классификации текстов и регионального профилирования авторов (как для расширения корпуса, так и для верификации априорной разметки).

Основные выводы следующие:

1. Регионализмы имеют очень низкую плотность распределения в текстах. Кроме того, в отличие от других социолингвистических категорий (например, гендера и возраста), региональные признаки крайне неравномерно представлены в доступных для сплошного компьютерного анализа данных, что препятствует получению статистически значимых результатов для многих «маленьких», хотя и, возможно, интересных в языковом отношении регионов.
2. В результате исследований по автоматической региональной классификации текстов на основании использования региональных словарей был получен вполне естественный негативный результат: высокая точность идентификации при крайне низкой полноте [17]. Конкретные цифры есть в статье, но сейчас неважны, поскольку были получены на довольно грязных данных.
3. Автоматическая региональная классификация текстов стандартными методами классического и глубокого обучения дает результаты, которые значимо выше случайных (в особенности в искусственных условиях подбора максимально ортогональных и при этом хорошо представленных укрупненных регионов, например, «Краснодарский край и Кавказ», «Урал

и Сибирь», «Восточная Украина и Киев»), но не позволяют рассматривать такие методы как надежное средство верификации априорной или получения новой региональной разметки.

4. При исследовании значимости признаков выявлено, что основной вклад в качество региональной идентификации произвольного текста вносят **не регионализмы, но другие виды регионально смещенной лексики**, прежде всего, топонимы и прочие регионально значимые именованные сущности и нерегиональная по сути лексика, связанная с важными локальными событиями. Это, разумеется, вполне предсказуемый результат, который, тем не менее, стоило проверить.

Таким образом, для решения двойной задачи расширения регионально размеченных данных и автоматического поиска регионализмов необходимо:

- Перейти от задачи классификации текстов к задаче авторского профилирования (тем самым решая проблему низкой плотности и неравномерности распределения регионально смещенной лексики любого типа).
- Задачу поиска кандидатов в регионализмы решать статистическими методами, элиминируя из ранжированных списков прочие типы регионально смещенной лексики.
- Универсальным средством повышения качества является очистка данных. Эта задача в значительной степени решена в новой версии ГИКРЯ (см. [9].)

4.2 Определение «карты» употребления регионализмов

Итогом первого этапа является список кандидатов в регионализмы, полученный как в результате предложений участников проекта (в версии ЯРГ), так и автоматическим анализом регионально смещенной лексики по данным ГИКРЯ.

Этап проверки реальной картины регионального распределения не получается пока делать полностью автоматически. Эту проблему нам еще предстоит решить, прежде всего увеличением как общего объема корпуса, так и применением автоматических методов профилирования для увеличения регионально маркированной части, которая и сейчас весьма значительна, но крайне неравномерно представляет регионы, которые, естественно, существенно различаются числом авторов в соцсетях.

Для верификации регионального распределения регионализмов в проекте ЯРГ использовалась, помимо соцсетей, и база данных региональных СМИ «Интегрум». К сожалению, этот замечательный ресурс не имеет API-доступа к текстам для пользователей, что не позволяет использовать его не только для верификации, но и для поиска кандидатов. Возможно, этот вопрос удастся когда-нибудь решить.

Результатом верификации являются данные, подобные приведенному ниже распределению вариантов *магаз/магазин/лабаз*:

В Ленинграде 1980-х лабаз — обычно винный магазин, у современных авторов младших возрастов — любой магазин. Вот цитаты из старшего поколения:

В. Гаврильчик (1929–2017): *Одиннадцать пропикало, / Народ бежит в лабаз (1978); Нас мотало в метро. И в лабазах давили. / В зной и стужу стояли мы у пивного ларь*ка (1979).

*За водкой и более деликатными алкоголями бились (в буквальном смысле слова) в специальных отделах, полуподвалах, лабазах, и все равно только БЛАТ давал вожделенную влагу в нужном для праздника количест*ве (Сергей Юрский. Вспышки, 2001, В Москве с 1978 года, с 43 лет.)

Региональный и возрастной анализ в ЖЖ ГИКРЯ (без учета семантики):

Нужен эталон для сравнения. 1. нейтральное *в магазин* и 2. «молодёжное» *в магаз*.

Избранные регионы (число словоупотреблений)

Регион	*в магаз*	*в магазин*	*в лабаз*
Мск	2621	38001	58
СПб	749	10893	52
Моск. обл.	100	1835	4
Ленобласть	37	611	6
Мск/СПб	3,50	3,49	1,12
Всего	10130	178215	350

Доля избранных регионов, %.

Регион	*в магаз*	*в магазин*	*в лабаз*
Мск	25,9	21,3	16,6
СПб	7,4	6,1	14,9
Моск. обл.	1,0	1,0	1,1
Ленобласть	0,4	0,3	1,7

Слово явно петербургское.

Разбивка по возрасту (число словоупотреблений)

Год рождения	*в магаз*	*в магазин*	*в лабаз*	*в магазин / в лабаз*
1950–1969	72	3508	18	194,9
1970–1979	609	13582	58	234,2
1980–1999	2448	30279	30	1009,3
1950–1999	3129	47369	106	446,9

Доля отдельных когорт, %

Год рождения	*в магаз*	*в магазин*	*в лабаз*
1950–1969	2,3	7,4	17,0
1970–1979	19,5	28,7	54,7
1980–1999	78,2	63,9	28,3

Выражение *в магазин* дает обычное возрастное распределение для нейтральной семантически умеренно маркированной лексики (для немаркированной доля младшей возрастной когорты не сильно превышает 50%, но молодежь походы в магазин упоминает чаще).
Межпоколенная разница в тяге к хождению *в магазин* и *в лабаз* очевидна.

4.3 Лексикографическое описание.

Региональный словарь является толковым. Поэтому вопросы собственно лексикографического описания регионализмов не имеют какой-то очевидной специфики, если не считать наличия в словарной статье собственно региональных признаков.

Лексикографическое описание является трудоемким процессом, особенно в ситуации, когда идиолект лекскографа не включает описываемого значения. Ограниченность «редакторского» ресурса привела к тому, что не все очевидные регионализмы, собранные в ходе проекта ЯРГ, получили полноценное лексикографическое описание. Так, в словарь ЯРГ под Lingvo была включена лишь примерно четверть того, что было предложено к обсуждению, и около половины того, что получило статус проекта словарной статьи.

Образцы словарных статей (в условном формате) можно увидеть ниже:

смитник
мусорный бак, мусорное ведро, помойка
Ex*: Эти, юные тогда одесситы, были затем высоко (аж до Харькова, Киева и Москвы) подняты гребнем 1920-х годов, раздавлены и вышвырнуты на смитник страны в конце 1930-х…* (Порто-Франко, Одесса; 23.03.2007
Syn: альтфатер, жбан, мульда, пухто
Reg: Одесса, возможно, вся Украина

альтфатер
мусорный контейнер с крышкой
Ex*: Женщина (откуда-то с провинции, вроде как с Москвы) сильно хочет выкинуть кулек с мусором в альтфатер…* (блог, Одесса); *Пацаны вышли за хлебом, таз атаковал альтфатер* (форум, Черновцы).

Syn: жбан, мульда, пухто, смитник
Reg: Одесса, Черновцы

пухто
тж. пухта
мусорный контейнер
Ex*: Там, во дворах, двадцать лет стояли баки и пухто, но ЖКС № 2 решил их убрать, хотя деньги за вывоз отходов с жильцов собирает* (Невское время, Санкт-Петербург; 15.12.2000); *Оставленную пустую пухту заполнят за несколько часов, и через день за ней снова приедет КамАЗ* (Сельская новь, Волосово, Ленинградская область; 11.06.2005).
Syn: альтфатер, жбан, мульда, смитник
Reg: Петербург, возможно вся Ленинградская область

мульда
мусорный бак
Ex*: Кстати, скоро в Ростове появится порядка трех тысяч новых контейнеров, но вблизи строек установят не их, а более вместительные мульды* (Наше время, Ростов-на-Дону; 05.08.2003); *Ни кабинок для переодевания, ни душевых кабинок, и даже мульд для мусора* (АиФ Удмуртии; 17.07.2003).
Syn: альтфатер, жбан, пухто, смитник
Reg: Ижевск, Ростов-на-Дону

Объем статьи не дает возможность осветить устройство регионального классификатора: это вопрос, в котором перемешаны социолингвистические, культурно-исторические и административные соображения. Скажем только, что в ГИКРЯ используется трехуровневый классификатор (грубо говоря, страна — регион — город), в сильной степени по необходимости связанный с системой региональных признаков соцсетей. В ЯГеЛе используется несколько отличная система, это несоответствие еще предстоит преодолеть. Общий объем региональных признаков – около 1000. Все в целом делает задачу автоматической классификации весьма нетривиальной.

5 Интерактивный региональный словарь проекта ЯГеЛь

Желание вернуть к жизни форум ЯРГ, перенеся его из ставшего чужим проекта Lingvo в «профильный» ГИКРЯ, возникло давно. Однако этому мешала необходимость решить сначала вопросы строительства новой версии корпуса, что оказалось делом непростым и небыстрым, поскольку проект около 5 лет оставался (и остается) без грантовой поддержки. Как оказалось, эксперты полагают, что «корпус Русского Языка у нас уже есть». Фактически развитие проекта проходило и проходит за счет практики и НИР студентов специальности Компьютерная Лингвистика МФТИ и РГГУ при общей оргподдержке отдела перспективных исследований ABBYY и с 2020 года – Лаборатории ABBYY Lab в Физтех-Школе ПМИ МФТИ.

Проект ЯГеЛь был начат студентами МФТИ в ходе т.н. ИннПрака (рук. Т.О. Шаврина). Была сделана попытка поэкспериментировать с региональной разметкой первой версии ГКИРЯ с прицелом на интересы неискушенного пользователя, «любителя слов».

В дальнейшем идеи этого проекта были несколько переосмыслены и в настоящее время реализуется идея объединенного портала для дифференциальных социолингвистических исследований, включающего:

1. интернет-корпус ГИКРЯ новой расширенной и очищенной от неавторских текстов версии;
2. интерактивный словарь ЯГеЛь (Языки Городов и Людей), в состав которого вошли полностью словарные и методические материалы проекта ЯРГ, импортированные из формата форума Lingvo (уже закрытого) в более современный формат Wiki-dictionary. Таким образом, сотни пользователей этого ресурса вновь получают доступ к нему: https://int.webcorpora.ru/reg2/index.php/Языки_городов_и_людей

В отличие от проекта ЯРГ, который был связан с поиском и описанием только региональной лексики, ЯГеЛЬ включает материалы для словаря паремий и разделы для будущих гендерных и возрастных словарей.

Благодарности

Мы благодарим за соучастие и поддержку всех участников Lingvo-форума «Языки русских городов» и в особенности редактора первой версии регионального словаря Марию Ахметову!

Мы также глубоко признательны участникам первой студенческой версии ЯГеЛя, сделанной в рамках ИннПрака МФТИ под руководством Татьяны Шавриной.

References

[1] Akhmetova M.V. (2014), Lexical regionalisms and localisms in Runet: problems of collecting materials [Leksitscheskie regionalism i lokalismy v russkoyazychnom Internete: problem sbora materiala] // Russian language and new technologies [Russkiy yazyk i novye tehnologii], Moscow, pp. 156-171.

[2] Akhmetova M.V. (2015), From A-Aty till Yarsk: unofficial townnames dictionary [Ot A-Aty do Yarska: slovar neofitsyalnykh nazvaniy naselennyh punktow]. – M:FORUM, 2015. – 496 p.

[3] Belikov V.I. (2009), Stereotypes in literary norms understanding [Stereotipy v ponimanii literaturnoi normy] // Language, communication and culture stereotypes [Stereotipy v yazyke, kommunikatsyi i kulture], Moscow, pp. 357-377.

[4] Belikov V.I. (2009), Lexical usus of official documents and codified dictionary norm [Leksicheskyi uzus ofitsyalnych documentov I kodifitsirovannaja slovarnaja norma] // Social language options – VI [Sotsyalnye variant yazyka - VI], Nizhniy Novgorod, pp. 65-68.

[5] Belikov V.I. (2010), Metodic news in the social lexicography of the XXI century [Metodicheskie novosti v sotsialnoj leksikografii XXI veka] // Slavica Helsingiensia 40. Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian, Helsinki University Press, pp.32 -49.

[6] Belikov V.I. (2012), On the method of identifying isoglossae of urban regionalisms [K metodike vyavlenia isogloss gorodskih regionalismov] // Modern problems of cultural and linguistic regionalism [Sovrememmye problem kulturno-yazykovoi regionalistiki], Perm, PONITSAA, pp. 8-14.

[7] Belikov V.I. (2014), On the methodology of corpora research of vocabulary [K metodike korpusnogo issledovanija leksiki] // Russian and new languages [Russkij yasyk i novyje], Moscow, pp. 99-130.

[8] Belikov V.I. (2016), Беликов В. И. What can a linguist get from digitized texts and its ways [Chto i kak mozhet poluchit lingvist is otsifrovannykh tekstov]// Siberian Philological Journal [Sibiskij filologicheskij zhurnal], No 3, pp. 17 -34.

[9] Belikov V.I., Selegey V.I., Selegey D.V. (2020), Web-corpus as a tool for linguistic research: differentiation, authorization, thematic biases (or corpora we want so much to believe) [Internet-korpus kak instrument lingvisticheskikh issledovanij: differentsialnost, avtorixatsija, tematicheskije smeshenija (ili korpusy, kotorym tak hochetsa verit’] – Computational Linguistics and Intelligent Technologies [Kompjuternaja lingvistika i intellektualnyje tekhnologii].

[10] Comprehensive Explanatory Dictionary of the Russian Language [Bolshoj tolkovyj slovar russkogo yazyka], SPb, 1998.

[11] Gel’gardt R.R. (1959), About the literary language in the geographical area [O literaturnom yazyke v geograficheskoj proektsii] // VJA, No 3.

[12] Itskovich V.A. (1982), Essays on the syntactic norm - M.: Nauka.

[13] Evgenjeva A.P. (1981 – 1984), Russian Language Dictionary, 4 vol. [Slovar russkogo yazyka], vol. 2, Moscow, Rus.yaz.

[14] Belousova E.V. (1999), Municipal Law of the Russian Federation: Reader [Munitsypalnoe pravo Rossijskoj Federatsii: Khrestomatija], M.: Jurist, 544 p.

[15] ARCoMT: All-Russian classifier of municipal territories [OKTMO: Obsherossijskij klassificator territorij municipalnykh obrazovanij], OK 033-2013, M.: Standartinform, 2013.

[16] Dictionary of Russian folk dialects [Slovar russkikh narodnykh govorov], vol. 1 et al, M. – L.: Nauka, 1965 - …

[17] Sorokin A.A. (2015), Automatic regional classification based on the dictionary of regional vocabulary: a trial study [Avtomaticheskaja regionalnaja classificatsija na osnove slovarja regionalnoj leksiki: probnoje issledovanije], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2015” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2015”], Bekasovo.

[18] Ushakov D.N. Explanatory dictionary of the Russian language [Tolkovyj slovar russkogo yazyka], vol. 1, M.:SE, OGIZ, 1935.

[19] Filin F.P. (1973), On the structure of the modern Russian Literary language [O structure sovremennogo russkogo literaturnogo yazyka] // Linguistics issues [Voprosy yazykpznanija], No 3.
[20] Shvedov N.J. Explanatory dictionary of the Russian language with the etymology of the words [Tolkovyj slovar russkogo yazyka s vklucheniem informatsii o proishozhdenii slov] / Institute of the Russian Language of the Russian Academy of Sciences [Institut russkogo yazyka RAN], M.: Azbukovnik, 2007.
[21] Shor R.O. (1926, 2009), Language and Society [Yazyk I obschestvo], vol. 3, M.: Lenand.
[22] Age and Gender Identification in Unbalanced Social Media. Juan Carlos Gomez, Luis-Miguel López-Santamaría, Mario-Alberto Ibarra-Manzano, Dora-Luz Almanza-Ojeda, 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), 2019

Список литературы

[1] Ахметова М. В. Лексические регионализмы и локализмы в русскоязычном Интернете: проблемы сбора материала // Русский язык и новые технологии: М.: Нов. лит. обозрение, 2014. С. 156–171.
[2] Ахметова М. В., От А-Аты до Ярска: словарь неофициальных названий населенных пунктов / Отв. ред. В. И. Беликов. М.: ФОРУМ, 2015. 496 с.
[3] Беликов В. И. Стереотипы в понимании литературной нормы // Стереотипы в языке, коммуникации и культуре. М.: РГГУ, 2009-а. Стр. 357 -377.
[4] Беликов В. И. Лексический узус официальных документов и кодифицированная словарная норма // Социальные варианты языка — VI. Нижний Новгород: НГЛУ, 2009-б. Стр. 65 -68.
[5] Беликов В. И. Методические новости в социальной лексикографии XXI века // Slavica Helsingiensia 40. Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian / Editors: A. Mustajoki, E. Protassova, N. Vakhtin. Helsinki: — Helsinki University Press, 2010. Pp.32 -49.
[6] Беликов В. И. К методике выявления изоглосс городских регионализмов // Современные проблемы культурно-языковой регионалистики. Пермь: ПОНИЦАА 2012. С. 8 -14.
[7] Беликов В. И. К методике корпусного исследования лексики // Русский язык и новые. М.: Нов. лит. обозрение, 2014. С. 99 -130.
[8] Беликов В. И. Что и как может получить лингвист из оцифрованных текстов // Сибирский филологический журнал. 2016, № 3. С. 17 -34.
[9] Беликов В.И., Селегей В. П., Селегей Д. В. Интернет-корпус как инструмент лингвистических исследований: дифференциальность, авторизация, тематические смещения. В сб. «Компьютерная лингвистика и интеллектуальные технологии» 2020.
[10] Большой толковый словарь русского языка / Под ред. С. А. Кузнецова. — СПб., 1998.
[11] Гельгардт Р. Р. О литературном языке в географической проекции // ВЯ, 1959, № 3.
[12] Ицкович В. А. Очерки синтаксической нормы. М.: Наука, 1982.
[13] Словарь русского языка: В 4 т. / Под ред. А. П. Евгеньевой. — 2-е изд., испр. и доп. — М.: Рус. яз., 1981–1984
[14] Муниципальное право Российской Федерации: Хрестоматия / Сост. Е. В. Белоусова. — М.: Юристъ, 1999. 544 с.
[15] ОКТМО: Общероссийский классификатор территорий муниципальных образований. ОК 033-2013. Тт. 1 -8. М.: Стандартинформ, 2013.
[16] Словарь русских народиых говоров. Вып. 1 (и следующие, издание не завершено) М.-Л.: Наука, 1965-…
[17] Сорокин А. А. Автоматическая региональная классификация на основе словаря региональной лексики: пробное исследование, Компьютерная лингвистика иинтеллектуальные технологии. По материалам ежегодной конференции Диалог, Бекасово, 2015
[18] Толковый словарь русского языка: В 4 т. Т. 1. / Под ред. Д. Н. Ушакова. — М.: СЭ; ОГИЗ, 1935.
[19] Филин Ф. П. О структуре современного русского литературного языка // Вопросы языкознания, 1973, № 2.
[20] Толковый словарь русского языка с включением сведений о происхождении слов / Ин-т рус. яз. РАН. Отв. ред. Н. Ю. Шведова. — М.: Азбуковник, 2007.
[21] Шор Р. О. Язык и общество. М., 1926. Издание 3-е. М.: Ленанд; 2009
[22] Age and Gender Identification in Unbalanced Social Media. Juan Carlos Gomez, Luis-Miguel López-Santamaría, Mario-Alberto Ibarra-Manzano, Dora-Luz Almanza-Ojeda, 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), 2019

Influence of speech breathing after physical activity on intonational-pausal segmentation of speech

Belkova Lubov
Lomonosov Moscow State University
belckowa.l@yandex.ru

Abstract

This study raises the problem of the difference between normal and forced (deep) speech breathing. The aim of this work was to study the intonational-pausal segmentation of speech in normal and forced breathing after physical activity. The results of the study show that in the process of reading, the structure of the text determines the organization of breathing, and the breathing rate and respiration depth have an impact on the intonational-pausal segmentation of speech, as well as on the duration and quantity of intonation pauses.

Keywords: phonetics; breathing pause; text localization; intonational-pausal segmentation; forced breathing
DOI: 10.28995/2075-7182-2021-20-94-109

Влияние речевого дыхания при мышечных нагрузках на интонационно-паузальное членение речи

Белькова Любовь
Московский Государственный
Университет им. М. В. Ломоносова
belckowa.l@yandex.ru

Аннотация

В данном исследовании поднимается вопрос о различии обычного и форсированного (глубокого) речевого дыхания. Целью работы явилось исследование интонационно-паузального членения речи при обычном и форсированном дыхании после физической нагрузки. Результаты исследования показывают, что при чтении структура текста определяет организацию дыхания, а частота и глубина дыхания влияют на интонационно-паузальное членение речи, длительность и количество интонационных пауз.

Ключевые слова: фонетика; дыхательная пауза; текстовая локализация; интонационно-паузальная сегментация; форсированное дыхание

1 Introdaction

1.1 Breathing: normal and forced

Breathing is a process consisting of three stages: 1) external respiration; 2) transport of gases by the circulatory system; 3) internal respiration. For the body processes the cells must receive oxygen and emit carbon dioxide – this is how energy is released.

External respiration provides lung ventilation that maintains a constant concentration of oxygen and carbon dioxide in the alveolar air [4].

The respiratory cycle consists of the inspiratory and expiratory phases. During inspiration (which is an active process), the muscles contract, the diaphragm raises the lower ribs, thereby the chest capacity increases. Expiration is a passive process: the muscles relax and the diaphragm descends.

Forced breathing (FB), i.e. deep breathing, differs from normal breathing (NB). FB can be associated with psycho-emotional or physical stress. During muscular exercise, a person needs more oxygen and

thus the tidal volume increases owing to the need to satisfy the metabolic needs of the body. In the forced inspiration, the muscles of the back and neck, rotator cuff, and facial muscles also take part. The chest capacity increases even more compared to a normal inhale. Forced exhalation is also an active process, by contrast with the normal inspiration, since the lung ventilation should be carried out faster; the air is sharply released by the means of auxiliary muscles.

During NB, about 1% of the oxygen consumed by the body is spent on the work of the respiratory muscles; during physical exertion and FB, the energy consumption for lung ventilation increases to 20%. Thereat, in both cases, the breathing mode (frequency and depth) is carried out involuntarily, depending on the physical feasibility of the body [15].

1.2 Breathing and speech

There is a distinction between physiologic and speech breathing [5]. In physiologic respiration, the inhale is equal in duration to the exhalation. The frequency of inspirations for adults is 16 – 18 p.m. In the setting of muscle load, the frequency of inhales can reach 40 – 60 p.m. During speech breathing, the inhale accelerates, and the exhale slows down significantly. The amount of inspiratory air can increase by 3 times compared to the physiological exhalation [2].

Speech breathing is associated not only with the production of sounds, but also with the formation of rhythmic-intonational segmentation of speech. In modern phonetics, it is considered that breathing passively adjusts to intonation pauses (IPs) [6]. A pause is usually understood as a break in articulation and, consequently, a break in the speech signal. Such pauses can be described as temporal (TPs - they are realized by a segment of zero intensity in the signal). Moreover, in the absence of a physical pause (NTP - non-temporal pause) as a pause can be interpreted a sharp change in tone and other prosodic parameters [7]. A pause in speech is often used for inhaling, but this does not mean that it necessarily occurs because of the physiological need. A person in a calm state pauses, including in accordance with the semantic segmentation of speech [17]. Typically, pauses occur between syntactic units. The pause can be filled with an inspiration and then it is called a breathing pause (BP). A non-breathing pause (NBP) is performed without inhaling. The experimental data (Krivnova 2007 [8], Grogean/Coxtins 1979 [3], Zellner 1994 [16]) show that the preferred location of speech inspiration is the end of a sentence (in English and Russian) and the end of a clause[1] (in Russian). In a number of experiments, Krivnova discovered that the hierarchical structure of text units plays an important role in the organization of speech breathing: “Text fragments are organized in descending order of the probability of the next breath in a certain way <...>: paragraph (100%) > sentence within a paragraph (94%) > clause within a sentence (65 %) > component within a clause (34 %)” [12]. At the same time, the length of the BP varies according to its textual localization: “Low PBS (Perceptual Boundary Strength) values usually go hand in hand with a short pause <…>. High PBS values frequently imply a long pause” [13]; see also: “the mean duration of BPs and NBPs is not only a function of speaking rate but also of syntax. Both types of pauses are longer at the End S [end of sentence] location than at any other location and as the linguistic importance of the breaks diminishes, so does the duration of BPs and NBPs” [3].

In case of forced breathing, the localization, number, and type of IPs seem to vary. However, there is no evidence to support this hypothesis in the literature which is known to us.

In this regard, the main research tasks were:

1) to identify the main differences in the organization of intonation segmentation of the text while reading,
2) to analyze the coherence of TPs with the boundaries of different text units,
3) to compare the duration of TPs at the borders of different text units

during NB and FB.

[1] A clause is any group, including a non-predicative one, whose top is a verb, and in the absence of a lexical verb, a copula or a grammatical element that plays the role of a copula [14].

2 Material, participants and research methodology

The material for this study was the reading of the specially constructed text. The experimental text contained 7 graphic paragraphs, 27 sentences (7 - simple, monoclausal; 20 - polypredicative constructions of various types), 68 clausal units that are components of sentences, and 1 utterance (see Table 1 in the **Appendix 1**). The text represents a textual unity with the presence of direct speech, dialogues, and expressive words.

The text after preliminary acquaintance immediately before the start of the experiment was read over from a paper sheet twice in the same sequence for everyone: first - in a normal condition and then after physical exercise. Physical activity was the same for all participants; it was achieved by quick (as possible) climbing the stairs from the 1st to the 10th floor. Each participant's heart rate was measured while reading the text before and after the exercise load. The pulse was measured by palpation of the radial artery. The average frequency of the heart rate at rest is usually 60-80 beats p.m. for healthy people aged 18 to 50. During normal breathing, the pulse of the participants was 61-78 beats p.m. After physical activity, the pulse of all informants varied from 95 to 127 beats p.m., which is typical of the frequent pulse (more than 90 beats p.m.) [1]. The text was read from a paper sheet by 8 participants, women of 18-20 years old, Russians, native speakers of literary pronunciation, students of the Faculty of Philology, but without special announcer training. Further, participants are indicated by letters of the Latin alphabet and numbers 1-2, where 1 is the reading in NB, 2 – in FB (a-1, a-2; b-1, b-2, etc.). Examples of reading the text by the same speaker with two options for pausing (in NB and FB) are represented in the **Appendix 2**.

The reading was recorded on a ZOOM Handy Recorder H4n using a condenser microphone.

The average duration of the sounding text was 88 sec. (73 sec. without pauses) during NB and 94 sec. during FB (72 sec. without pauses).

The type of IPs and the presence of respiratory filling in them in the voiced versions of the text were determined audibly and visually by spectrograms using the sound analyzer *Praat*, version 6.0.33. The duration of pauses and other speech segments was measured in semi-automatic mode using the same analyzer.

Examples of segmentation are shown in Figure 1-a - 2-a in the **Appendix 3**.

3 Results and discussion

3.1 General phonetic pattern of IPs with different text localization during NB and FB

Figure 1 shows the oscillograms and spectrograms of IPs with different text localization during NB.

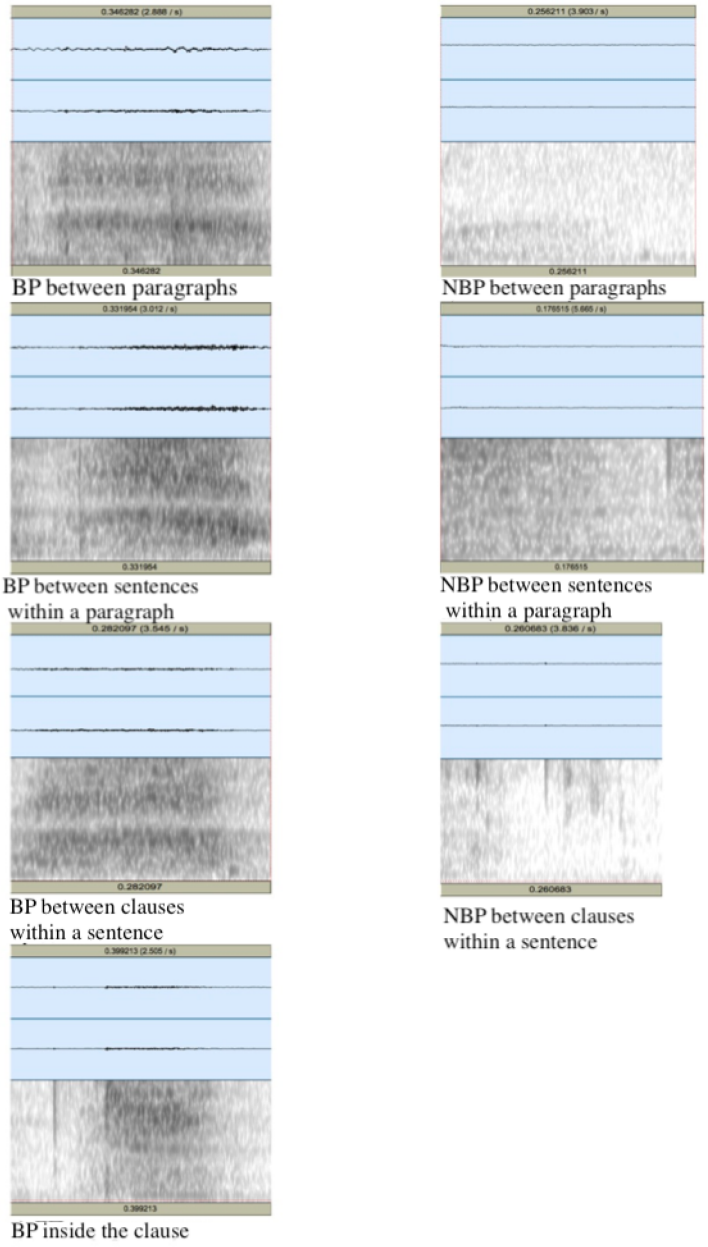


Figure 1: Acoustic pattern of IPs. BPs are on the left. NBPs are on the right. Pauses are represented without capturing of the adjacent sections of the speech signal. The duration in seconds is shown at the top.

As reflected by Figure 1, IPs with different text localization and content differ from each other:

1) Any BPs is characterized by four phases (see Figure 1-a in the **Appendix 3**): a) termination of articulation and, consequently, of the speech signal; b) a short, voiceless part, sometimes filled with a short nasal inhale or exhale; c) noisy oral (rarely nasal) inspiration and smacking, preceding it, swallowing, etc. (which are realized by an explosion, as in obstruent consonants); d) the voiceless part before resuming the speech signal.

2) BPs have a longer duration than NBPs (see 3.3).

3) Pauses are characterized by different durations depending on the hierarchical structure of the text units. However, the pause inside the clause is not subject to this pattern, perhaps due to the fact that the pause was made after the utterance (*Эх*) inside the clause and was perceived as a pause between sentences inside the paragraph. Such cases will be discussed below.

Figure 2 shows the oscillograms and spectrograms of IPs with different text localization during FB (text was read by the same participant).

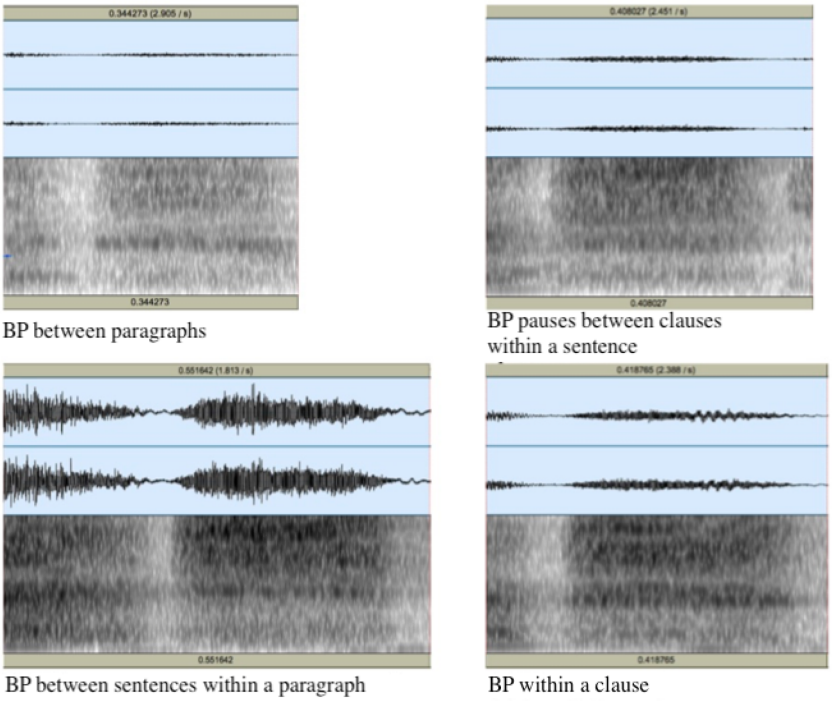


Figure 2: Acoustic pattern of IPs. Pauses are represented without capturing of the adjacent sections of the speech signal. The duration in seconds is shown at the top.

Figure 2 shows only the BPs. In FB, this participant did not make NBPs. Many pauses were not realized where they were implemented before the physical exercise.

Comparison of Figures 1 and 2 allows to identify the following features of forced speech breathing:

1) The BP during FB is characterized by four phases (see Figure 2-a in the **Appendix 3**): a) an exhalation that is inseparable from the speech signal. Exhalation is characterized by the noise similar to the noise of the fricative [x]; b) a short, voiceless part; c) noisy oral or nasal inspiration. It ought to be noted, that during FB before inhaling, there are no smacking, swallowing, etc., because the inhale is always realized immediately after the exhalation; d) the voiceless part before resuming the speech signal.

2) The BP can be prolonged in comparison with the BP during NB, because it includes not only inhaling, but also exhaling; this pattern, however, is not always observed.

3) Inhaling is noisier than the inhaling during NB.

4) The duration of the pause does not always depend on the hierarchical structure of the text.

Thus, we can conclude that the nature of pauses is determined by the features of breathing.

3.2 General features of speech breathing in the readings of the same text by different participants during NB and FB

The general data is given in Table 2 in the **Appendix 4** and in Figures 3 and 4.

Among the most common features of speech breathing in the readings of the same text by different participants during NB and FB we can emphasize the different quantity of IPs and the difference in the speech tempo. The data was determined by the read out text, without taking into account internal TPs. Most participants read the text at an medium tempo. Average tempo speed (Lenneberg, 1967) is 5.6-6.7 syllables p.s. [11]. Thus, the duration of the syllable is 150 - 170 ms.

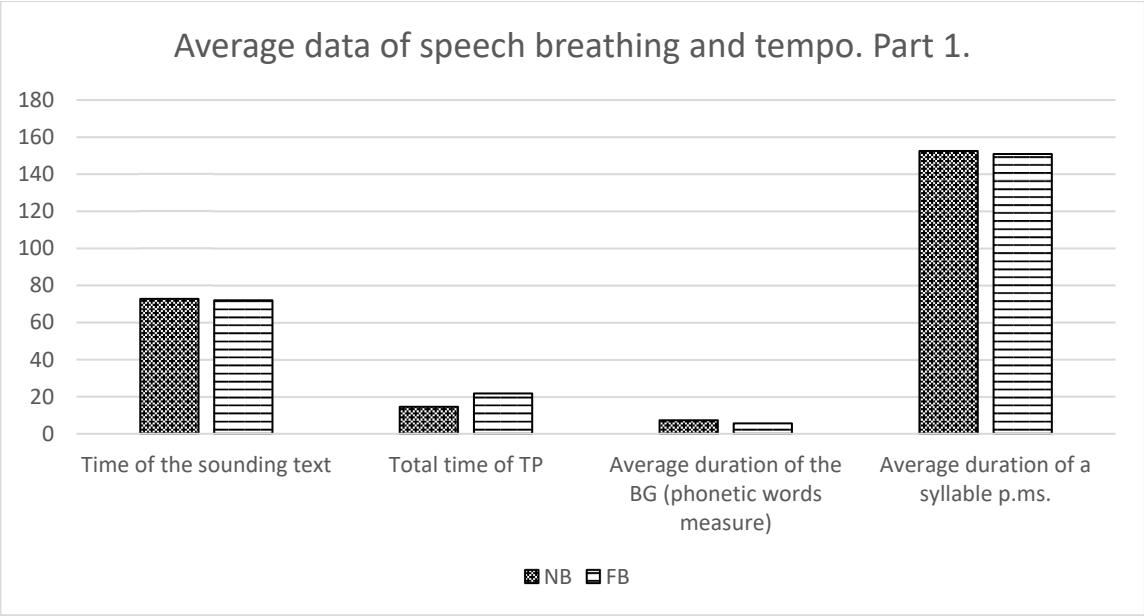


Figure 3: General characteristics of speech breathing and the tempo of reading of the experimental text by different participants during NB and FB

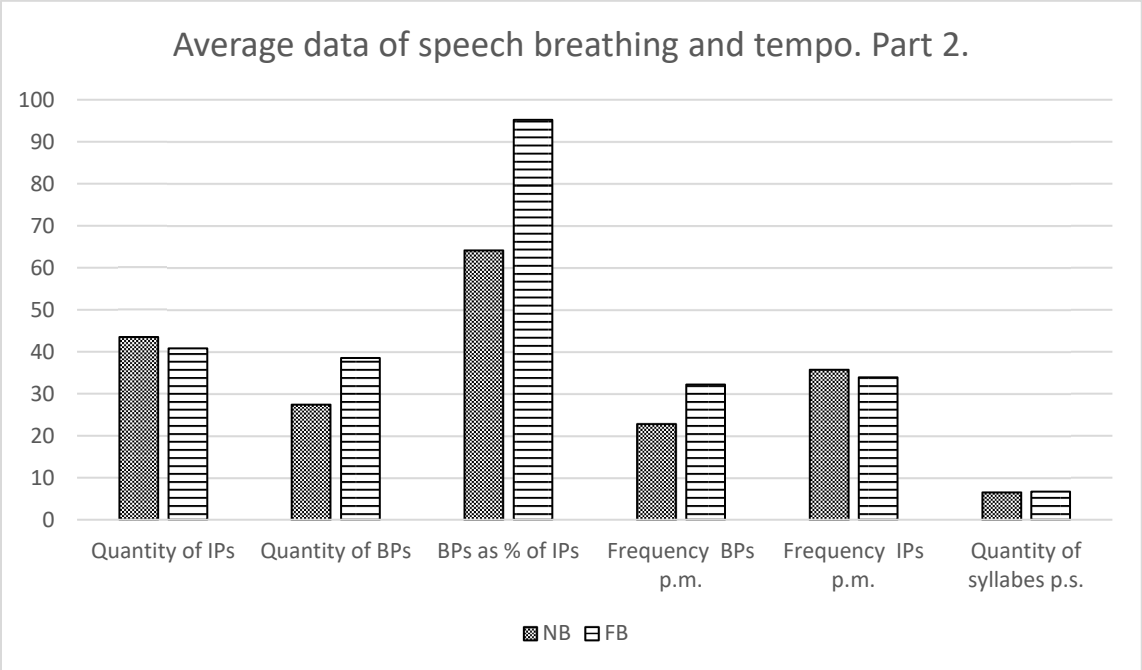


Figure 4: General characteristics of speech breathing and the tempo of reading of the experimental text by different participants during NB and FB

As seen from Table 2 and Figures 3 and 4, the differences in FB are more related to the quantity of BPs than to the tempo and the time of reading the text.

The type of breathing does not affect the total time of reading (the average value is 72.8 seconds during NB and 72.0 seconds during FB) and the tempo of pronouncing (the average duration of a syllable is 153 ms during NB and 151 ms during FB). During FB, the time of TPs increases by 1.5 times. It can be connected with an increase in the depth and, accordingly, the duration of inspiration. The duration of

the breathing groups[2] (BGs) decreases during FB by 1.3 times due to the fact that the participant is forced to pronounce fewer words between the BPs because their quantity and frequency increase (by 1.4 times). Therefore, during FB, BPs take up most of the IPs. The percentage of BPs from the total number of pauses during FB is 1.5 times higher than during NB. The average number of IPs is slightly decreasing in FB (from 44 to 41), as well as the frequency of their implementation per minute (from 36 to 34). This is due to the fact that the participant stops implementing TPs where they are intended, if he does not need to inhale.

Thus, the type of breathing primarily affects the number of BPs during the reading of the text. In FB, BPs take up 95% of the total number of TPs, which also increases compared to NB.

3.3 Correlations between the localizations of BPs and the boundaries of text units in NB and FB

The question of the correlation between the localization of BPs and the boundaries of text units in NB and FB was investigated by Krivnova [8], [9]. She found out that all the participants she interviewed avoided taking inspirations inside the clause, but certainly made inhales after completing the paragraphs, regardless of their length and complexity. A similar pattern was observed at the boundaries of sentences, and at the boundaries of clauses the number of BPs decreased. This suggested that the number and localization of BPs are determined by the intonation pausing strategy, and it is the speaker who sets this strategy: "Important parameters for the implementation of the breath at the interclausal boundaries within the sentence are the longer length of the pronounced clause, its autosemanticity and the expectation (prediction) of the expanded sentence" [8].

In the present study, the task was to determine whether speakers followed the intonation pausing strategy during FB. The data obtained is shown in Table 3 in the **Appendix 5** and in Figures 5 and 6 below.

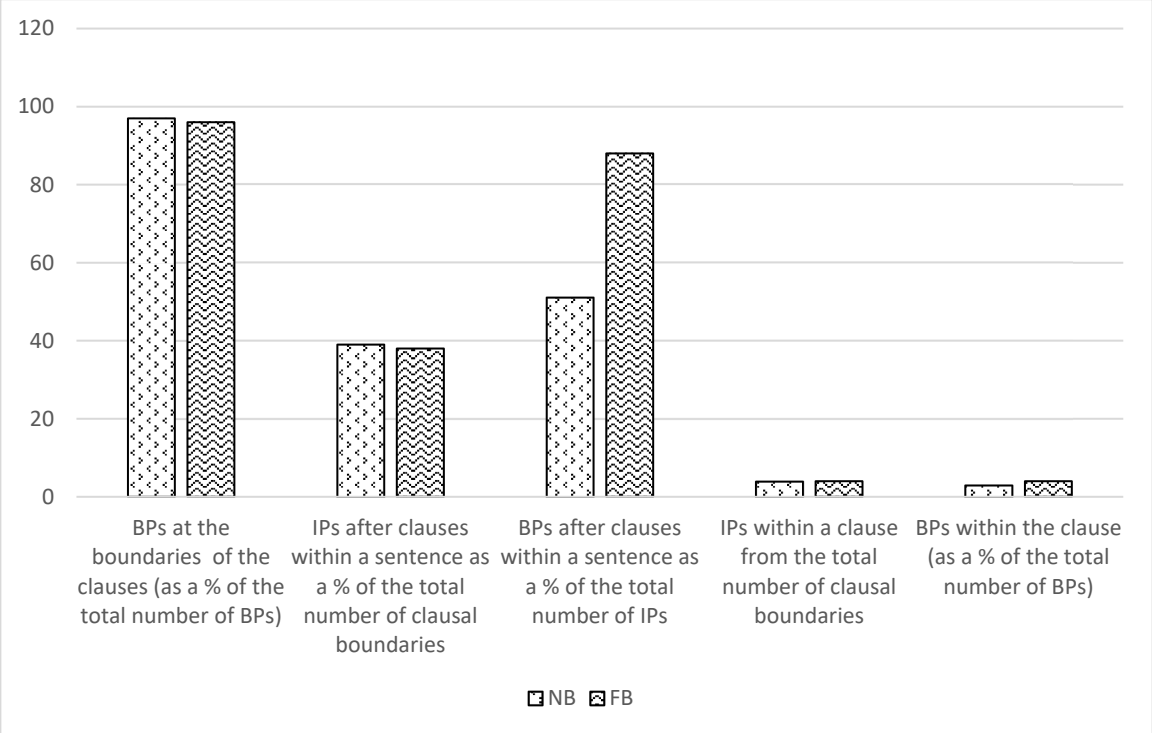


Figure 5: BPs at the boundaries of text fragments with different indexes of syntax boundary strength during NB and FB (part 1)

[2] Breathing group is a chain of words uttered by the speaker during one exhalation [11].

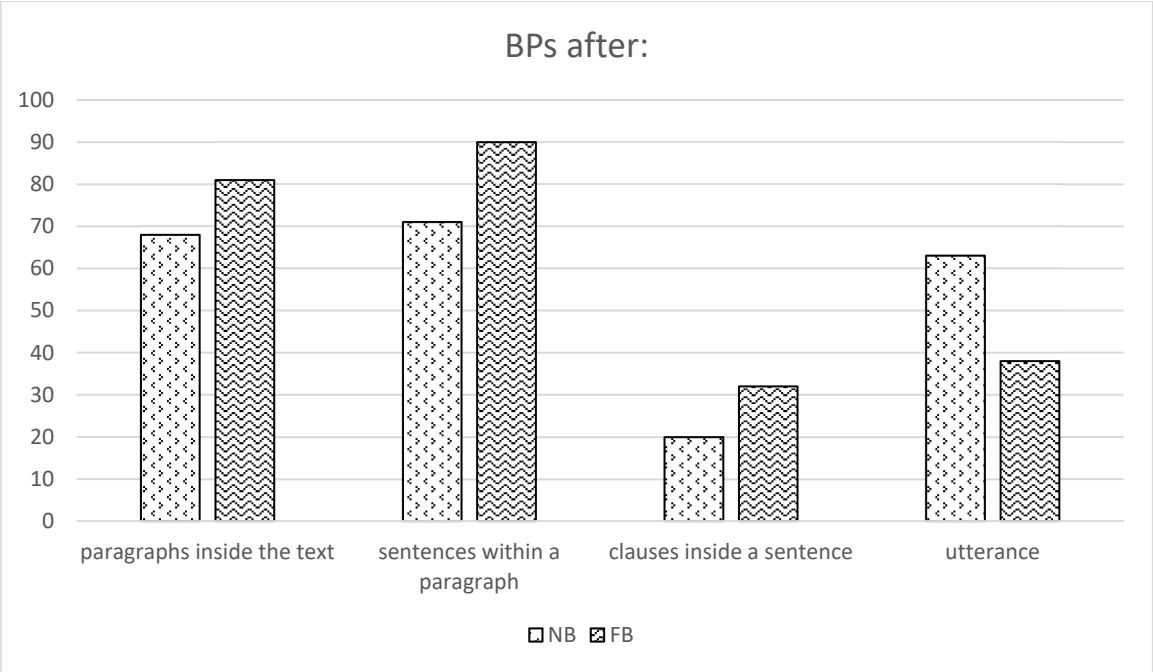


Figure 6: BPs at the boundaries of the text fragments with different indexes of syntax boundary strength during NB and FB (part 2)

As can be seen from the above data, most often the BPs coincide with the boundaries of the clausal boundaries, both in NB (97%) and in FB (96%). After the clauses within the sentence, in NB, the BPs take up about half of the IPs. During FB, this indicator increases by 1.7 times. The participant tries to use any pause for the realization of the inspiration. An increase in the quantity of BPs during FB is observed at all levels of text units. The fewest number of intonation pauses is found inside the clauses, both in NB (20%) and in FB (32%). At the same time, the number of IPs increases during FB by 12%. BPs after paragraphs make up only 68% of the total number of the paragraphs. This discrepancy with the data described by Krivnova [10] can be explained by the fact that the text contains a lot of dialogic utterances, which are graphic paragraphs, but not phonetic. Nevertheless, in FB, the percentage of BPs after paragraphs increases. Breathing pauses after utterance during NB (63%) also converges in percentage terms with the indicators of BPs after a paragraph (68%) and a sentence (71%). However, in FB, BPs after utterance are noted less often (38%) than in NB; however, this cannot be considered a regular process, since this utterance is perceived by participants in different ways: either as a sentence or as part of it.

Based on what was said above, we can conclude that IPs are often used for inhales, this indicator increases during FB. The organization of speech breathing reflects the hierarchical structure of text units, regardless of the depth and frequency of breathing.

3.4 The duration of IPs with different text localization

The assumption that the duration of intonation pauses reflects the hierarchical structure of the text has been expressed by various linguists (Grogean/Coxtins 1979 [3]; Krivnova 2016 [11]). Krivnova conducted a number of experiments confirming this point of view [12].

At this stage of the study, the task was to calculate the average duration of BPs and NBPs in different text localizations during the readings of the same text by different participants in NB and FB and to determine whether the duration of the pause depends on its localization during FB.

The data obtained is shown in Figure 7 below.

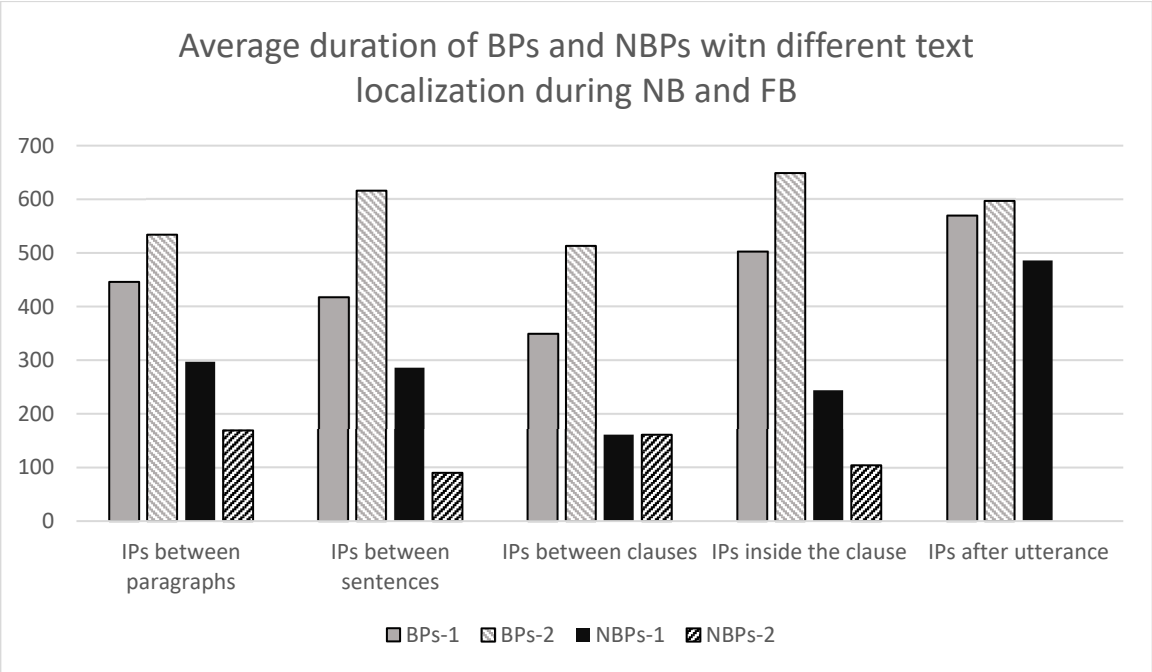


Figure 7: Average duration of BPs and IPs with different text localization during NB and FB. BPs-1 - BPs during NB. BPs-2 - BPs during FB, NBPs-1 - NBPs during NB. NBPs-2 - NBPs during FB. The duration of the pauses was determined in ms

Detailed data on the duration of different types of pauses is provided in Table 4 in the **Appendix 6**.

Analyzing the results of the obtained data, we can conclude that the BP has a longer duration (1.4 times) than the NBPs, regardless of the text localization and type of breathing. In NB: the duration of the IPs reflects the hierarchical structure of the text within the concepts of paragraph > sentence > clause; the IP after the utterance is equal in duration to the IP after the sentence; the IP inside the clause is forced or interpreted as a pause between other text units. The duration of IPs during FB does not reflect the hierarchical structure of the text. The duration of the BPs in FB is longer than in NB, the duration of the NBPs, on the contrary, in FB is less than in normal breathing, except for the NBPs between clauses, where their duration is equal. The average duration of IPs during FB increases by 1.38 times compared to NB (see Table 2).

The study thus shows that the duration of the pause depends on the presence of inspiration in it and the type of breathing. The BP is always longer than the NBP. The duration of the BP during FB increases, and the duration of the NBP decreases.

4 Conclusions

The obtained results confirm that the type of breathing affects the organization of the rhythmic-intonational segmentation of the text. During both normal and forced breathing, the participants tend to pause at the boundaries of text units. However, during forced breathing, not all potential intonation pauses are realized, but only those that are necessary for the realization of inspiration.

The type of breathing most depends on the quantity of breathing pauses that are implemented during the reading of the text. During forced breathing, breathing pauses take up 95% of the total number of temporal intonation pauses - this is 31% more than during normal breathing.

In normal breathing, the duration of the intonation pause depends on its localization and most often shows the hierarchical segmentation of the text: paragraph (446 ms) > sentence (417 ms) > clause (349 ms). In forced breathing, this pattern is not observed. The duration of the pause does not depend on the location of its implementation.

References

[1] Bol'shev A.S., Sidorov D.G., Ovchinnikov S.A. (2017), Heart rate. Physiological and pedagogical aspects [Chastota serdechnyh sokrashchenij. Fiziologo-pedagogicheskie aspekty], NNGASU, Nizhny Novgorod.

[2] Dubrovskiy V.I. (2002), Sports Medicine [Sportivnaya medicina], VLADOS, Moscow.

[3] Grosjean F., Collins M. (1979), Breathing, Pausing and Reading, Phonetica 36, Boston, pp. 98 – 114.

[4] Kayumova A.F. (2016), Physiology of the respiratory system: a textbook [Fiziologiya sistemy dyhaniya: uchebnoe posobie], FSBEI HE BSMU of the Ministry of Health of Russia [FBGOU VO BGMU Minzdrava Rossii], Ufa.

[5] Knyazev S.V., Pozharickaya S.K. (2012), Modern Russian literary language: Phonetics, orthoepy, graphics and orthography [Sovremennyj russkij literaturnyj yazyk: Fonetika, orfoepiya, grafika i orfografiya], Academic Project [Akademicheskij proekt], Gaudeamus, Moscow.

[6] Kodzasov S.V., Krivnova O.F. (2001), General phonetics [Obshchaya fonetika], RSUH [RGGU], Moscow.

[7] Krivnova O.F., Chardin I.S. (1999), Pausing in automatic speech synthesis [Pauzirovanie pri avtomaticheskom sinteze rechi], Theory and practice of speech research (APCO-99) [Teoriya i praktika rechevyh issledovanij (APCO-99)], Moscow, pp. 87–103.

[8] Krivnova O.F. (2007), Factor of speech respiration in intonational-pausal articulation of speech [Faktor rechevogo dyhaniya v intonacionno-pauzal'nom chlenenii rechi], Linguistic polyphony [Lingvisticheskaya polifoniya], Languages of Slavic cultures [Yazyki slavyanskih kul'tur], Moscow, pp. 424 – 444.

[9] Krivnova O.F. (2009), The general phonetic picture of respiratory pauses in the reproduced speech (on the reading material) [Obshchaya foneticheskaya kartina dyhatel'nyh pauz v reproducirovannoj rechi (na materiale chteniya)], Phonetics and Grammar: present, past, future [Fonetika i grammatika: nastoyashchee, proshedshee, budushchee], Vol. 11, pp. 61–71.

[10] Krivnova O.F. (2010), Duration of respiratory pauses with different text localization [Dlitel'nost' dyhatel'nyh pauz s raznoj tekstovoj lokalizaciej], Phonetics today: Materials of reports and messages of the VI International Scientific Conference [Fonetika segodnya: Materialy dokladov i soobshchenij VI Mezhdunarodnoj nauchnoj konferencii], Moscow, pp. 84-86.

[11] Krivnova O.F. (2016), Prosodic phrasing in spoken text: localization of breathing pauses [Prosodicheskoe chlenenie zvuchashchego teksta: tekstovaya lokalizaciya dyhatel'nyh pauz], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2016” [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam mezhdunarodnoj ezhegodnoj konferencii «Dialogue 2016»], Moscow, pp. 340–354.

[12] Krivnova O.F. (2017), Phonetic characteristics of breathing pauses with dIfferent text localization [Foneticheskie harakteristiki dyhatel'nyh pauz s raznoj tekstovoj lokalizaciej], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2017” [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam mezhdunarodnoj ezhegodnoj konferencii «Dialogue 2017»], Moscow, pp. 207–220.

[13] Sanderman A.A. (1996), Prosodic phrasing: production, perception, acceptability and comprehension, Thesis Technische Universiteit Eindhoven, Eindhoven.

[14] Testelets Ya.G. (2001), Introduction to the general syntax [Vvedenie v obshchij sintaksis], Russian State University for the Humanities [Rossijskij gosudarstvennyj gumanitarnyj universitet], Moscow.

[15] Zarifyan A.G. (2013), Physiology of respiration: a textbook [Fiziologiya dyhaniya: uchebnoe posobie], KRSU, Bishkek.

[16] Zellner B. (1994), Pauses and the temporal structure of speech, Fundamentals of speech synthesis and speech recognition, Chichester, pp. 41 - 62.

[17] Zinder L.R. (1979), General phonetics: Textbook. Manual [Obshchaya fonetika: Ucheb. Posobie], Higher School [Vysshaya shkola], Moscow.

Appendix

Appendix 1. In Table 1, column 1 shows the text units of different strength, taking into account their hierarchical boundaries. In column 2 (without the parentheses) the number of units of hierarchical boundary that are not final unit in a boundary of higher strength is indicated, and in parentheses – the number of units that complete a higher-strength boundary text unit. The data is presented in Table 1.

Text units	Quantity in the text	Quantity of text units in lower strength units	Quantity in phonetic words	End punctuation mark
Paragraph	7	from 1 to 10	from 2 to 84	. ? !
		Avg. 3,8	Avg. 28,6	
Sentence inside a paragraph	22 (+5) – endpoints in a higher strength unit		from 2 to 14	. ? !
		Avg. 2,5	Avg. 7,4	
Clause inside the sentence	41 (+ 27)	-	from 1 to 7	. , - ? ! :
			Avg. 2,9	
Utterance *Эх*	1	-	1	.

Table 1: Compositional and syntactic structure of the experimental text

Appendix 2

Experimental text with the distribution of pauses in accordance with the reading of Participant e-1 (during NB). The paragraph is separated by an interval, / is the sentence boundary inside the paragraph, ** is the clause boundary inside the sentence, the utterance is italicized. The pause duration in seconds is indicated in parentheses. After the parentheses there is the character of the pause (BP or NBP). The length of the syntagma in seconds is given as superscript:

Кеша всегда приходил в сапогах$^{1.680235}$ / (0.298) NBP – любил их, видно.$^{0.811328}$ / ** (0.455391) BP Никогда не разувался:$^{1.177735}$ / (0.177969) BP в сапогах руки мыл,$^{1.151563}$ / (0.173) BP ел,$^{0.355938}$ / (0.113) NBP пил, / спать ложился.$^{1.230078}$ / ** (0.392578) BP Однажды Кеша пропал.$^{1.188203}$ /** (0.389961) BP Три дня его дома не было,$^{1.350469}$ / (0.123) NBP отец с матерью обыскались.$^{1.463008}$ / ** (0.259102) BP На четвертый – пришел,$^{1.206524}$ / (0.102070) NBP и как пришел-то$^{0.902930}$ / (0.264336) BP – босиком!$^{0.591484}$ / ** (0.358555) BP Рубашка вся помята,$^{1.358321}$ / (0.253867) BP у куртки пуговицы оторваны, / стоит, / переминается.$^{2.648594}$ / ** (0.269570) NBP Ну, отец его давай расспрашивать.$^{1.583399}$ / ** (0.410899) BP А Кеша только зубами стучит.$^{1.570313}$ / ** (0.298359) BP Папа допытывается:$^{0.926485}$ / (0.303594) NBP «Колись, зараза, / где тебя носило?!»$^{1.845118}$ - / BP (0.259102) а по радио поэт Левитанский читает «Как показать зиму».$^{2.973126}$ / ** (0.562695) BP Да вот же лучшая иллюстрация / - замерзший и растерянный блудный Кеша!$^{3.690235}$ / ** (0.429219) BP «Слушай, / – говорю я$^{0.889844}$ / (0.319) BP – пойдем, / ну, пойдем, / умоешься»$^{1.570313}$ / (0.088984) - еле увел.$^{0.722344}$ / ** (0.262) BP Сидит Кеша над тазиком, / щеки трет:$^{2.072813}$ / (0.408281) BP

- Пошел в кафе, / – говорит$^{1.240547}$ / (0.246) BP – хотел спокойно посидеть,$^{1.455157}$ / (0.177969) BP заказать что-нибудь.$^{1.073047}$ / ** (0.337617) Заказать заказал,$^{1.002383}$ / (0.102070) а в рот ни куска не попало$^{1.436836}$ / (0.222461) BP – это все Пашка, гад:$^{1.096602}$ / (0.264336) BP «Сыграй кружочек!»$^{0.970977}$ / (0.183203) - ага, сыграл!$^{0.926485}$ / ** (0.476328) NBP

Я ему говорю:$^{0.680469}$ / (0.319297) BP «Так ты что, / в карты что ли сапоги продул?»$^{2.114688}$ / ** (0.350703) NBP

Продул.[0.554844] / ** (0.144) NBP Еще как продул.[1.436836] / ** (0.337617) BP Он же не только сапоги, / он и получку просадил, за два месяца.[3.030704] / ** (0.471094) BP *Эх*.[0.366406] (0.565313) BP Тут Кеша как начнет слезы утирать.[1.805860] / ** (0.209375) NBP А мне его жалко страшно, дурня этакого![2.093750] / ** (0.423984) BP Он человек-то хороший.[1.141094] / ** (0.272) BP Знаете как он папе с мамой помог, / когда им квартиру не давали?[2.931251] / ** (0.330) BP Это ведь он бегал, / все бумажки собирал, / письма писал.[2.653829] / ** (0.403047) BP А как он со мной нянчился,[1.397578] / (0.690938 - сглатывание) NBP когда мама заболела.[1.224844] / ** (0.408281) BP Однажды ночью разбудил меня и шепчет:[2.088516] / (0.413516) BP

- Давай просыпайся, / тут такое![1.863438] / ** (0.465859) NBP

- Что случилось?[0.884610] / ** (0.188438) NBP

- Показать хочу![0.916016] / (0.264336) BP – и протягивает мне какой-то странный сверток,[2.046641] (0.196289) BP тяжелый и теплый.[1.012852] / ** (0.389961) BP Это дядя Кеша ежика во дворе нашел[1.965508] / (0.256484) BP и в полотенце его завернул,[1.394961] / (0.177969) BP чтобы я посмотрел.[1.023321] / ** (0.434453) BP

Experimental text with the distribution of pauses in accordance with the reading of Participant e-2 (during FB). The paragraph is separated by an interval, / is the sentence boundary inside the paragraph, ** is the clause boundary inside the sentence, the utterance is italicized. The pause duration in seconds is indicated in parentheses. After the parentheses there is the character of the pause (BP or NBP). The length of the syntagma in seconds is given as superscript:

Кеша всегда приходил в сапогах[1.509348] / (0.450671) BP – любил их, видно. / ** Никогда не разувал(ся):[2.024020] / (0.448004) BP в сапогах руки мыл, / ел, / пил,[1.717350] / (0.597339) BP спать ложил(ся). / ** Однажды Кеша пропал.[2.048020] / ** (0.605339) BP Три дня его дома не было[0.636007] / (0.552005) BP отец с матерью обыскались. / ** На четвертый – при(шел),[2.704026] / (0.384004) BP и как пришел-то / – босиком![1.434680] / ** (0.626673) BP Рубашка вся помята,[1.194678] / (1.317346) BP у куртки пуговицы оторваны,[1.610682]/ (0.437338) BP стоит, / переминается.[1.184011] / ** (0.637339) BP Ну, отец его давай расспрашивать.[1.573348] / ** (0.546672) BP А Кеша[0.538672] (0.224002) BP только зубами стучит.[1.040010] / ** (0.509338) BP Папа допытывается:[0.944009] / (0.474671) BP «Колись, зараза, / где тебя носило?!»[1.725350] / (0.525338) BP - а по радио поэт Левитанский читает «Как показать зиму».[2.981362] / ** (0.522672) BP Да вот же лучшая иллюстрация - / замерзший и растерянный блудный Кеша![3.258698] / ** (0.432004) BP «Слушай, / – говорю (я)[0.874675] / (0.418671) BP – пойдем, / ну, пойдем, / умоешься»[1.533348] / (0.208002) BP - еле увел.[0,624006] / ** (0.464004) BP Сидит Кеша над тазиком, / щеки трет:[1.989353] / (0.402671) BP

- Пошел в кафе, / – говорит[1.224012] / (0.336003) BP – хотел спокойно посидеть, / заказать что-ни(будь).[2.368023] / ** (0.312003) BP Заказать заказал, / а в рот ни куска не попало[2.421357] / (0.448004) BP – это все Пашка, гад:[1.120011] / (0.341337) BP «Сыграй кружочек!» / - ага, сыграл![1.858685] / ** (0.528005) BP

Я ему говорю:[0.618673] / (0.469338) BP «Так ты что, / в карты что ли сапоги продул?»[1.821357] / ** (0.461338) BP

Продул. / ** Еще как продул.[1.298679] / ** (0.474671) BP Он же не только сапоги,[1.056010] / (0.394670) BP он и получку просадил, за два месяца.[1.968019] / ** (0.490671) BP *Эх*.[0.453338] (0.389337) BP Тут Кеша как начнет слезы утирать.[1.925352] / ** (0.922676) – BP А мне его жалко страшно,[1.210678] (0.458671) BP дурня этакого![0.752007] / ** (0.554672) BP Он человек-то хороший.[1.077344] / ** (0.538672) BP Знаете как он папе с мамой помог, / когда им квартиру не давали?[3.322699] / ** (0.378670) BP Это ведь он бегал, / все бумажки собирал, / письма писал.[2.730693] / ** (0.320003) BP А как он со мной нянчился, / когда мама заболела.[2.506691] / ** (0.373337) BP Однажды ночью разбудил меня и шепчет:[1.994686] / (0.490671) BP

- Давай просыпайся, / тут такое![1.733350] / ** (0.389337) BP

- Что случилось?[0.714674] / ** (0.266669) BP

- Показать хочу![0.789341] / (0.224002) BP – и протягивает мне какой-то странный сверток,[1.786684] (0.346670) BP тяжелый и теплый.[0.930676] / ** (0.642673) BP Это дядя Кеша ежика во дворе нашел[2.002686] / (0.362670) BP и в полотенце его завернул, / чтобы я посмотрел.[2.213355] / ** (0.410671) BP

Appendix 3

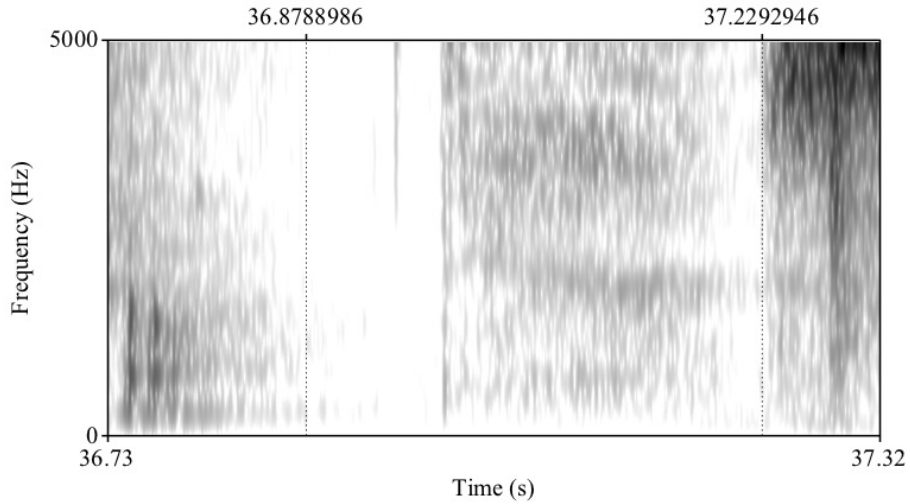


Figure 1-a: Dynamic spectrogram of the BP after the sentence during NB

In Figure 1-a, the cursors highlight the BP after the sentence during NB. The pause is located between the sounds [ə] and [s]. The duration of the segmented fragment was 350 ms in this case.

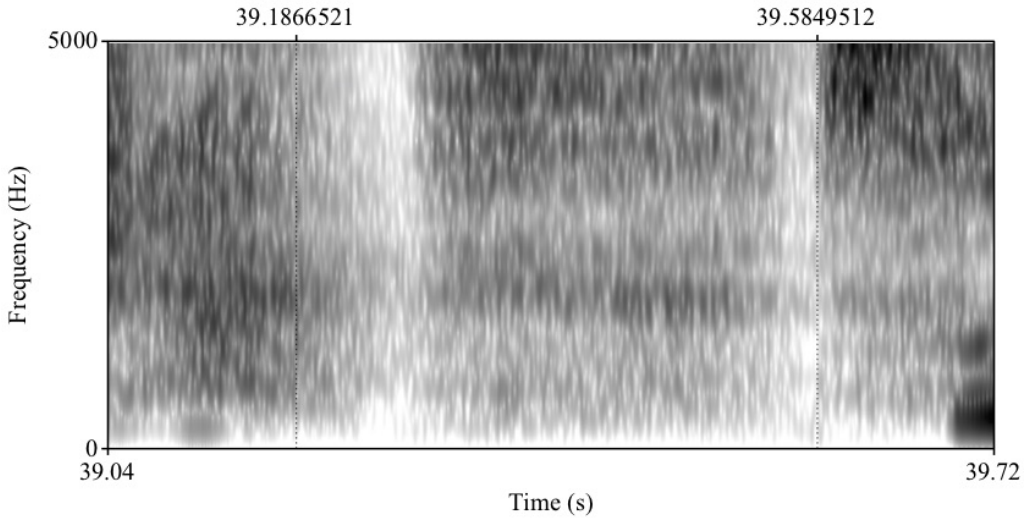


Figure 2-a: Dynamic spectrogram of the BP after the sentence during FB

In Figure 2-a, the cursors highlight the BP after the sentence during FB. The pause is located between the sounds [ə] and [s]. The duration of the segmented fragment was 398 ms in this case.

Appendix 4

Participants	Time of the voiced text	Total time of TPs	Pause frequency p. m.	IPs	BPs as a % of IPs	Quantity of IPs	The length of the BG in phonetic words		The average duration of a syllable p. ms.	Quantity of syllables p.s.	The frequency of BPs p.m.
							Avg.	Diapason			
a-1	76.1	17,1	37,5	47	57,5	27	7,4	2 - 12	159	6,3	21,4
a-2	75,8	23,3	40	52	88,5	46	4,3	2 - 9	159	6,3	37,5
b-1	69,8	11,7	33,3	38	60,5	23	8,7	1 - 16	147	6,8	20
b-2	67	15,7	21,4	24	100	24	8,7	3 - 18	142	7,1	21,4
c-1	65,1	16,1	42,9	48	54,2	26	7,6	2 - 16	137	7,3	24
c-2	64,2	21,1	37,5	40	100	40	5	2 - 12	135	7,4	37,5
d-1	76,8	12,4	31,6	40	47,5	19	10,4	5 – 21	161	6,2	15
d-2	76,2	26,6	27,3	34	94,1	32	6,25	2 - 13	159	6,3	25
e-1	78,4	16,7	40	54	68,5	37	5,4	1 - 12	164	6,1	28,6
e-2	76	22,1	37,5	47	100	47	4,2	1 - 10	159	6,3	37,5
f-1	74,9	17,5	31,6	40	77,5	31	6,5	1 - 12	156	6,4	25
f-2	72,7	22,8	33,3	40	92,5	37	5,4	1 - 12	153	6,6	30
g-1	78	15,3	37,5	48	56,3	27	7,4	1 - 16	164	6,1	20,7
g-2	77,5	20,8	40	52	86,5	45	4,4	1 - 12	161	6,2	35,3
h-1	63	11,4	31,6	33	90,9	30	6,6	2 - 12	132	7,6	28,6
h-2	66,8	22	33,3	37	100	37	5,4	1 - 12	139	7,2	33,3
Avg.-1	72,8	14,8	35,8	43,5	64,1	27,5	7,5	1,9-14,6	152,5	6,6	22,9
Avg.-2	72	21,8	33,9	40,8	95,2	38,5	5,6	1,6-12,3	150,9	6,7	32,2

Table 2: General characteristics of speech breathing and the tempo of reading the experimental text by different participants during NB and FB

Appendix 5

Participants	BPs at the borders of the clauses (as a % of the total number of BPs)	BPs after text units of different syntax strength (as a % of the total number of units of the corresponding type)				IPs after clauses within a sentence as a % of the total number of clausal boundaries	BPs after the clause within the sentence as a % of the total number of pauses	IPs within a clause from the total number of clausal boundaries	BPs inside the clause (as a % of the total number of BPs)
		After paragraphs inside the text	After the sentences inside the paragraph	After the clause inside the sentence	After the utterance				
a-1	96	57	64	22	100	41	54	10	4
a-2	93	57	95	49	0	59	83	7	7
b-1	96	57	73	7	100	24	29	2	4
b-2	100	43	91	5	0	5	100	0	0
c-1	96	71	73	15	0	41	36	7	4
c-2	98	86	100	32	0	41	78	2	2
d-1	100	71	50	12	0	29	41	5	0
d-2	97	86	77	22	0	27	81	2	3
e-1	97	43	82	41	100	63	65	2	3
e-2	91	100	82	49	100	49	100	10	9
f-1	97	100	82	17	100	32	53	2	3
f-2	97	86	91	29	100	34	85	2	3
g-1	96	57	68	22	100	54	41	2	4
g-2	98	100	95	44	100	59	75	5	2
h-1	100	86	73	24	0	27	89	0	0
h-2	97	86	91	29	0	29	100	2	3
Avg.-1	97	68	71	20	63	39	51	4	3
Avg.-2	96	81	90	32	38	38	88	4	4

Table 3: BPs at the boundaries of text fragments of different syntax strength during NB and FB

Appendix 6

	IPs between						IP inside a clause		IP after utterance	
	paragraphs		sentences		clauses					
	BP	NBP	BP	NBP	BP	NBP	BP	NBP	BP	NBP
a-1	0,524	0,308	0,486	0,287	0,401	0,174	0,457	0,172	0,605	-
a-2	0,528	0,232	0,538	-	0,451	0,131	0,317	-	-	-
IP	0,463/0,429		0,419/0,538		0,288/0,398		0,314/0,317		0,605/-	
b-1	0,321	0,124	0,436	0,213	0,303	0,116	0,501	-	0,501	-
b-2	0,496	-	0,702	0,090	0,393	-	-	-	-	-
IP	0,237/0,496		0,376/0,673		0,172/0,393		0,501/-		0,501/-	
c-1	0,464	0,332	0,359	0,376	0,329	0,186	0,317	0,317	-	0,615
c-2	0,506	0,197	0,572	-	0,491	-	0,536	-	-	-
IP	0,426/0,462		0,363/0,572		0,237/0,491		0,317/0,536		0,615/-	
d-1	0,519	0,216	0,441	0,218	0,418	0,121	-	0,244	-	0,362
d-2	0,490	0,101	0,896	-	0,818	0,138	1,240	-	-	-
IP	0,432/0,434		0,329/0,896		0,245/0,695		0,244/1,240		0,362/-	
e-1	0,419	0,370	0,380	0,240	0,275	0,223	0,566	-	0,566	-
e-2	0,421	-	0,520	-	0,464	-	0,355	-	0,389	-
IP	0,391/0,421		0,355/0,520		0,256/0,464		0,566/0,355		0,566/0,389	
f-1	0,481	-	0,490	0,441	0,405	0,208	0,805	-	0,805	-
f-2	0,681	0,144	0,623	-	0,539	-	0,822	-	0,822	-
IP	0,481/0,604		0,483/0,623		0,314/0,539		0,805/0,822		0,805/0,822	
g-1	0,451	0,458	0,367	0,350	0,333	0,175	0,368	-	0,368	-
g-2	0,507	-	0,463	-	0,370	0,214	0,579	0,104	0,579	-
IP	0,454/0,507		0,362/0,463		0,243/0,331		0,368/0,342		0,368/0,579	
h-1	0,387	0,273	0,376	0,161	0,329	0,074	-	-	-	-
h-2	0,645	-	0,617	-	0,579	-	0,694	-	-	-
IP	0,371/0,645		0,364/ 0,617		0,306/0,579		-/0,694		-	
Avg.-1	0,446	0,297	0,417	0,286	0,349	0,160	0,502	0,244	0,569	0,486
Avg.-2	0,534	0,169	0,616	0,090	0,513	0,161	0,649	0,104	0,597	-
Avg.IP	0,406/0,500		0,381/0,607		0,257/0,486		0,445/0,615		0,546/0,597	
407/561										

Table 4: Average duration of BPs and NBPs with different text localization during NB and FB. The first column contains the participants (n-1 - reading in NB, n-2 – in FB). In the IP line, the average duration of the IPs during NB and FB (NB / FB) is indicated by a "slash" (/). The hyphen (-) indicates the absence of an IP. The duration of the pauses was determined in seconds

Examining the role of linguistic context in aspectual competition: a statistical study

Beatrice Bernasconi
Roma Tre University – Sapienza University, Rome, Italy
beatrice.bernasconi@uniroma3.it

Valentina Noseda
Catholic University of the Sacred Heart Milan, Italy
valentina.noseda@unicatt.it

Abstract

This paper aims to show the results of a quantitative study on verbal aspect in modern Russian. Adopting a corpus-based approach, we investigate the phenomenon known as 'aspectual competition', which can take place when the imperfective aspect (ipf) is used instead of perfective to designate a single and complete event in the past. In particular, we investigate the interaction between the choice of aspect and co-textual factors in overlapping situations. In this study the attention is focused on one aspectual pair, namely *pokupat'*$_{ipf}$ - *kupit'*$_{pf}$, 'to buy'. The work consists of two parts: in Phase 1 data were collected from the spoken subcorpus of the Russian National Corpus and the web-corpus RuTenTen11, annotated for several morpho-syntactic factors, and then examined. In Phase 2 a questionnaire was submitted to native speakers in order to collect more empirical evidence on aspect choice and verify the results obtained from the corpus study. In both phases, statistical methods were used to analyse the data. Results show that the aspect of the target verb mainly interacts with two factors: the presence of a contiguous verbs in the linguistic context and the presence of an object modifier.

Keywords: verbal aspect, aspectual competition, general-factual imperfective, corpus linguistics, quantitative methods, statistical models

DOI: 10.28995/2075-7182-2021-20-110-118

Изучение роли языкового контекста в конкуренции видов: статистическое исследование

Беатриче Бернаскони
Третий Университет Рима – Сапиенца Римский университет
Рим, Италия
beatrice.bernasconi@uniroma3.it

Валентина Нозеда
Католический Университет Святого Сердца
Милан, Италия
valentina.noseda@unicatt.it

1 Introduction[1]

In this paper, we report the results of a double experiment on the choice of verbal aspect in competing[2] situations, namely when the Russian imperfective (ipf) can be used, instead of perfective (pf), to denote complete events in the past:

(1) *Он показывал мне ее фотографию.* [Padučeva 1996: 10]

[1] This work is the result of the close collaboration between the two authors, but Valentina Noseda is responsible for sections 1, 2, 5 and Beatrice Bernasconi for sections 3, 4.

[2] We distinguish between "competition" (examined here) and "opposition", e.g. when pf is opposed to the habitual or processual readings of ipf [Grønn, 2004: 30-35].

'He showed me her picture.'

Russian aspectologists usually refer to this phenomenon as *obščefaktičeskoe resul'tativnoe značenie* (in English *general-factual meaning*, henceforth ipf OR)[3], which, in turn, can be divided into two or three different sub-types (depending on the school of thought): in [1996] Padučeva proposes a threefold classification including *ekzistencial'noe* (existential), *konkretnoe* (concrete) and *akcional'noe* (actional) ipf OR, while Grønn [2004] mentions two types: *existential* and *presuppositional*[4].

A countless number of scholars have tried to grasp the motivations that lie behind this particular use of Russian ipf, and even though a point of encounter is far from being reached, several interesting and stimulating insights have enriched the literature on this topic [see, e.g., Forsyth, 1970; Gebert, 2014a, 2014b; Glovinskaja, 1982; Grønn, 2004; Israeli, 1996, 2001; Mehlig, 2001, 2013; Padučeva, 1996; Rassudova, 1982; Šatunovskij, 2009].

As far as the semantic differences between ipf OR and pf are concerned, we can summarise the main findings as follows[5]:

- with ipf OR the result of the action is topicalised, while the focus is represented by the action itself (existential meaning) or by another salient element of the sentence (actional/presuppositional meaning). With pf, the speaker focuses on the result [Padučeva, 1996: 37; Gebert, 2014a: 6];
- ipf OR is characterised either by the absence of the result at the moment of speech [Padučeva, 1996: 37], or by the uncertainty about the maintenance of the result at the moment of speech [Glovinskaja, 1982: 118]. Kreisberg [2007: 217] points out that the result could be present, but with ipf OR it is entirely irrelevant;
- while pf refers to a single and specific action, ipf refers to an action that could have occurred more than once (existential meaning) or is potentially replicable[6] [Padučeva 1996: 47-48];
- ipf is characterised by temporal indefiniteness, while with pf it is clear when the action reached its limit [Padučeva, 1996: 41]. That is why pf occurs instead of ipf in narrative progression [Grønn, 2004: 141];
- pf may signal the presence of a "pragmatic contract" [Israeli, 1996] between the speaker and their interlocutor, or a sort of "expectation" on the part of the speaker [Padučeva 1996];
- with pf "feeling is deliberately suppressed", while ipf is more emotional, conveying an implicit evaluative component [Forsyth, 1970: 89-91] (actional/presuppositional meaning).

As we can see, pragmatic factors can be extremely significant when choosing ipf OR over pf, and most of the times the pragmatic meaning of a sentence is not expressed by any evident linguistic cue in the text[7]. Nevertheless, none of the above-cited studies has addressed the matter adopting a fully usage-based and quantitative approach, which "is quite surprising considering the important role corpora could play for instance when comparing the frequency or preference for Ipf vs. Pf in specific syntactic environments" [Grønn, 2004: 12].

This work aims to examine the subject precisely from this perspective, filling the gap in the lack of quantitative studies on Russian general-factual imperfective. By collecting a sufficient amount of authentic linguistic data and examining them with statistical modelling, we try to determine the role of linguistic context in the choice of aspect.

[3] Most scholars agree on distinguishing, apart from resultative (*rezul'tativnoe*), three other general-factual uses, that will not be taken into account in the present study: atelic (*nepredel'noe*), bidirectional (*dvunapravlenoe*), non-resultative (*nerezul'tativnoe*) [Glovinskaja, 1982; Padučeva, 1996; Zaliznjak, Šmelev, 2000].

[4] Another interesting distinction, that does not exclude Padučeva's classification, is proposed by Mehlig [2001] and involves "actual" and "non-actual" predicates.

[5] Note that, to a great extent, such differences depend on the type of ipf OR we are dealing with.

[6] According to Padučeva [1996: 48], a sentence like *ty segodnja pokupal*$_{ipf}$ *kožanuju kurtky*? (Did you buy a/the leather jacket today?) does not admit a resultative reading, due to its 'unique' character; therefore, this should not be treated as a case of aspectual competition.

[7] Thanks to a corpus study conducted in the Russian National Corpus, Reynolds [2016: 103], demonstrated that verbal aspect in Russian seems to be "predominantly determined suprasententially, with lexical cues playing only a very minor role". Reynolds, however, investigated all the main uses of pf and ipf, including examples in all tenses, rather than just general-factual ipf vs pf.

In particular, we concentrated our attention on a single telic[8] aspectual pair: *pokupat'*$_{ipf}$ - *kupit'*$_{pf}$, 'to buy'. Such decision was determined by the idea that a verb, or a group of verbs sharing a main semantic trait, behaves in a unique way in competing situations [see Israeli, 1996; 2001]. As Gebert [2004: 202] points out "by now it is a universally accepted triviality that in languages aspect depends on the meaning of the verb".

2 Methodology

The study consists of a corpus analysis and an experiment with native speakers. In Phase 1 we extracted 600 examples of *pokupat'*$_{ipf}$ and *kupit'*$_{pf}$, in the past tense (300 for each form) from the spoken corpus of the Russian National Corpus (RNC)[9] and RuTenTen2011, a web corpus accessed through Sketch Engine[10]. This choice was motivated by the fact that "obščefaktičeskoe značenie charakterno prežde vsego dlja rečevogo, a ne narrativnogo režima"[11] [Sičinava, 2013].

All the examples were annotated for several factors, namely one dependent variable (ASPECT: *IPF* or *PF*) and eight independent variables[12]:

- object (OBJ), which could be singular (*sg*), plural (*pl*), a pronoun (*pron*) or absent (*no*);
- the presence of object modifiers (OBJMOD);
- object position in relation to the verb (OBJPOS), which could be *before*, *after* or *NA* (if the object was not expressed);
- time-measure complements (TIME), which could be definite (*def*), indefinite (*indef*) or absent (*no*);
- locative complement (LOC), which could be definite (*def*), indefinite (*indef*) or absent (*no*);
- the presence of other complements (OTHER)[13];
- sentence type, namely whether the sentence is a question or not (QUESTION);
- the presence of a contiguous verb in the sentence and, if so, its aspect (CONTVERB)[14];

Statistical tests were run on the data to determine if any of the above-listed factors interact[15] with the aspect of the main verb (*pokupat'*$_{ipf}$ - *kupit'*$_{pf}$) (see Section 3). The first model chosen for this purpose is Classification And Regression Trees ("CART") [Strobl et al., 2009], in which the algorithm makes recursive binary splits in the data, according to the independent variables that are associated with the dependent variable in a statistically significant way. This process yields a tree that shows the best way of separating the values according to the dependent variable. The algorithm also returns *p-values* for each split, showing their significance. A second test, namely a random forest model, was run on the dataset in order to verify the importance of the independent variables and to validate the results obtained through the CART model.

Phase 2 (Section 4) consisted of an experiment conducted with native speakers. A questionnaire was designed and submitted to 102 native speakers of Russian in order to give further evidence to the results obtained from the corpus study. The answers were then subjected to statistical analysis. In this case a

[8] Linguists do not always agree on which verbal predicates can be subjected to a factual-resultative interpretation [Grønn, 2004: 66], but most of them share the idea (which we support) that only telic predicates should be taken into account when talking about ipf OR [Grønn 2004; Gebert, 2014a; Kreisberg, 2007].

[9] https://ruscorpora.ru/new/search-spoken.html

[10] https://www.sketchengine.eu

[11] 'The general-factual meaning characterises above all spoken language rather than (written) narration'.

[12] Although we agree with those who put great emphasis on the distinction between different types of ipf OR, e.g. [Grønn, 2004] (see Section 1), the dataset was not annotated according to this parameter, as it does not apply to pf.

[13] Mainly a beneficiary in the dative case.

[14] With the label 'contiguous verb' we refer to a verbal predicate denoting a past action that preceded or followed the action conveyed by *pokupat'*$_{ipf}$ or *kupit'*$_{pf}$, like *slomat'sja*$_{pf}$ - 'to break' in the following example: *Купил я электродрель на магазин.ру, а она сломалась у меня в тот же день* (RuTenTen11) [I bought an electric drill on magazin.ru and it broke the same day]. Note, though, that we annotated a contiguous verb only when both predicates were deictic, excluding therefore all the examples that displayed a deictic past along with a relative past, e.g.: *Мне просто рассказывали, что на барахоловке где-то покупали* (RNC). (…) [I was just told that they had bought it at a flea market somewhere].

[15] We prefer to use the term 'interaction' since we cannot prove that the authors of the examples based their choice between *pokupat'*$_{ipf}$ or *kupit'*$_{pf}$ on the varying of such factors. In this sense 'interaction' can be intended as a synonym of 'collocation'.

mixed effects logistic regression model[16] was run to detect how participants' answers vary according to the parameters considered.

3 Corpus study: looking for statistical evidence

Firstly, data were submitted to a CART test. The model was run considering the dependent variable (ASPECT) and all the independent variables for which the database was annotated (see Section 2). Results are shown in Figure 1.

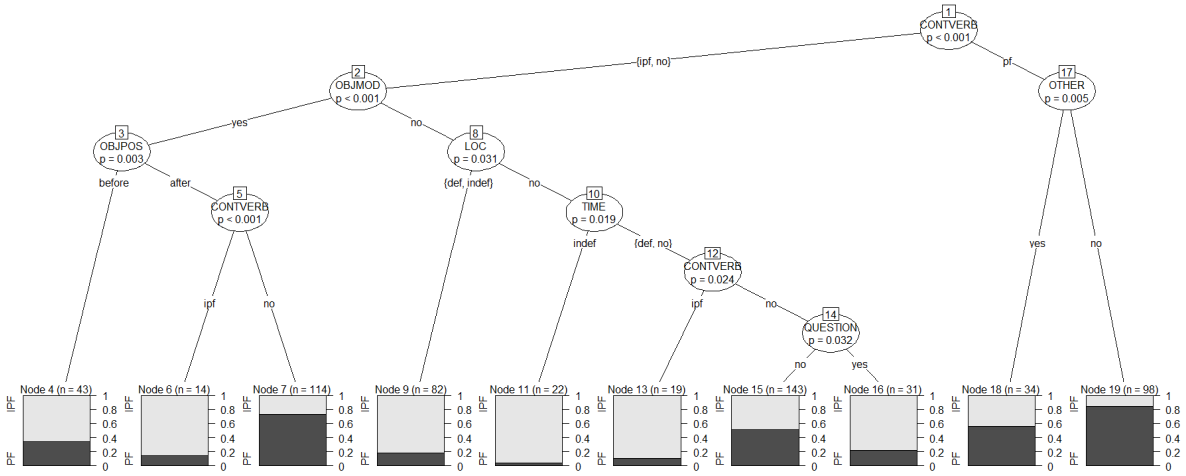


Figure 1: CART model for *pokupat'*$_{ipf}$-*kupit'*$_{pf}$

As illustrated in the plot, the model split our data into nineteen nodes, using the most significant factors as separators. All splits are statistically significant since *p-values* are always lower than 0.05. The most significant ones, however, are those in Nodes 1, 2, and 5 with a contiguous verb (CONTVERB) and an object modifier (OBJMOD) as factors and *p-values* < 0.001. The nodes at the bottom show the number of examples in each of them ("n=") and how the values of the dependent variable ASPECT are distributed.

The first split (Node 1) divides the data according to the factor CONTVERB. When the values for this factor are "ipf" or "no" (i.e., when the contiguous verb is either imperfective or absent), the data are split according to other factors (Nodes 2 to 16). When the contiguous verb is pf, the data are split only according to the factor OTHER (Node 17). When the contiguous verb is either ipf or absent, the data are divided according to the presence of a modifier of the object. If the object is modified, the examples are then separated depending on the position of the object (OBJPOS) and then again on the contiguous verb. These last two splits result in Node 4, 6, and 7. In Node 4, forty-three examples with an object modifier and a preposed object are contained. The distribution of the values for ASPECT in this node is 65% ipf – 35% pf, showing a predominance of ipf. When the object has a modifier and it is postposed, examples are divided according to the contiguous verb. The fourteen examples with a contiguous ipf are grouped in Node 6, with a predominance of ipf of the main verb *pokupat'*$_{ipf}$-*kupit'*$_{pf}$ (86% ipf – 14% pf). In Node 7 one hundred and fourteen examples are gathered. In these examples there is no contiguous verb, the object has a modifier and it is postposed. The distribution of ipf and pf is respectively 27% – 73%.

Nodes from 8 to 16 have in common the absence of an object modifier and show a predominance of ipf. Only in Node 15 ipfs and pfs are balanced with a distribution of, respectively, 49% – 51%. This node contains the highest number of examples (n=143) but is also the least significant (*p-value* = 0.032). Node 17 splits all the examples with a pf as a contiguous verb into nodes 18 and 19, according to the presence of an additional complement. In both cases, *kupit'*$_{pf}$ is more frequent (56% in Node 18 and 85% in Node 19).

[16] Using a mixed effects model we could handle the individuality of each participant. We assigned them a random ID (e.g. A, B, C …CZ) and included it in the model as a random variable.

A second test, namely a random forest model, was then run on the dataset. The dot chart in Figure 2 shows its results.

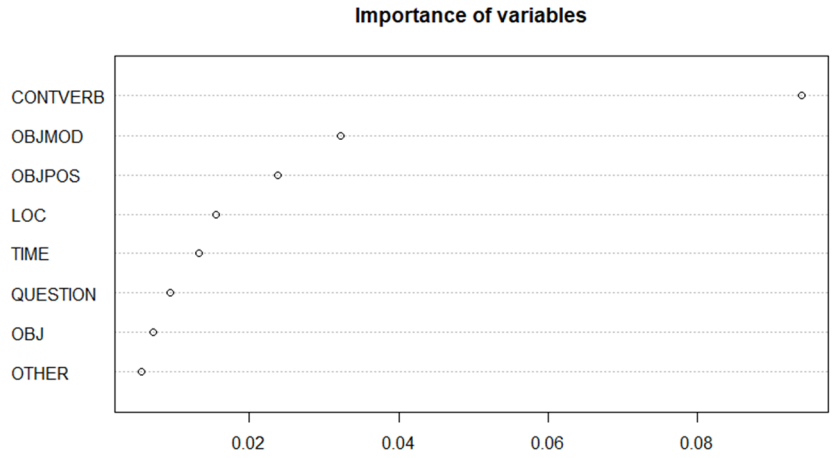


Figure 2: Variable importance of variables for *pokupat'*$_{ipf}$*-kupit'*$_{pf}$

The most significant variable in predicting the outcome ASPECT is CONTVERB, followed by OBJMOD and OBJPOS. These last two, however, present a slightly low coefficient of importance (< 0.05), which leads us to conclude that the real predictor is only the contiguous verb. Although the CART displayed many other significant splits, the actual importance of the other variables is irrelevant. This can also be explained by looking at the size of the nodes in Figure 1 and the distribution of the two aspects in each of them. Table 1 shows the values of importance for each variable considered.

Values of variable importance

CONTVERB	OBJMOD	OBJPOS	LOC	TIME	QUESTION	OBJ	OTHER
0.094	0.032	0.024	0.015	0.013	0.009	0.007	0.006

Table 1: Values of variable importance

4 Questionnaire

The results obtained so far were further verified by conducting an experiment with Russian native speakers.

4.1 Stimuli and procedure

We designed a questionnaire with 28 authentic examples taken from our dataset in which the main verb (*pokupat'*$_{ipf}$ *-kupit'*$_{pf}$) was replaced by a blank space. Each example represented the combination of features of every node obtained from the CART (see Figure 1, Section 3)[17]. Selecting the extracts, we aimed at providing the fullest context possible. Examples were not edited, therefore they maintained speech-related characteristics, including some grammatical inaccuracies. Example (2) is representative of how the stimuli were submitted to participants[18].

(2a) *[А.] Это киевские конфеты. Мы в Киеве/_____ с Галей. [Б.] Как вы э-э/ в Киеве/ время провели? Хорошо? [А.] Хорошо*. [НКРЯ]
'[A.] These are candies from Kiev. Galja and I (bought$_{ipf}$) them in Kiev. [B.] How was your time in Kiev? Was it good? [A.] It was.' [RNC]

[17] To avoid a time and effort consuming task, we decided to extrapolate four examples from nodes with more than fifty entries (taking one entry with ipf and one with pf from each of the two corpora considered) and only two examples from smaller nodes.
[18] The letters in square brackets indicate turns of speech in the dialogue.

The experiment was conducted as an online survey using the platform Survio[19]. Participants were recruited by sharing the direct link to the survey to individuals and on social networks. In one week, 102 answers were successfully collected[20]. For each example, participants were asked to choose the verb that better fitted the context according to their perception. A facultative question, in which they could explain their choice, was also included.

4.2. Results and Statistical analysis

Answers showed that participants chose pf in most of the cases (58,6%). If looking at the original aspect of each example, native speakers were more consistent in the answer when the original text presented a perfective verb, while in examples with an original ipf, the distribution of answers was almost equal (see Figure 3):

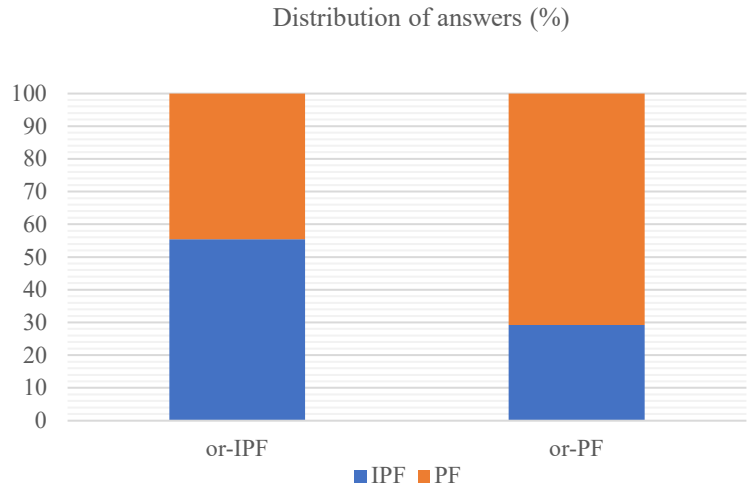


Figure 3: Distribution of answers according to the original aspect

When the original aspect was ipf (left column), participants' answers matched the original in 55,4% of the cases, while pf was chosen in 44,6% of the cases. When the original aspect was pf (right column), only 29,2% of the surveyed people answered with an ipf, while most answers (70,8%) matched the original.

By looking at the answers to the facultative question, it appears that often contextual 'cues' do not raise the same reaction in every native speaker. In (3), for example, some of them were triggered to choose the ipf (which was also the original aspect) because of the word *vse*, 'everything', while others chose pf because of the influence of the contiguous verb *vzjal*, 'took$_{pf}$'. The distribution of answers for this example is 22,5% ipf vs. 77,5% pf.

(3) *Он это _____ а потом взял всё и [пауза] не пропил/ да/ в казино не проиграл/ то есть за границу не увёз/ а оставил всё в России.* [НКРЯ]

'He (bought$_{ipf}$) this and then he took everything and [pause] didn't drink it all nor lost it in a casino, that is, he didn't' take it abroad but left it all in Russia.' [RNC]

A mixed effects logistic regression model was then run on the answers. A random ID was assigned to each participant and was included as the random variable PERSONID. The answers (ANSWER) were coded according to the aspect that each participant chose for each example. The independent variables CONTVERB and OBJMOD were considered as predictors for the outcome ANSWER. Since they resulted to be the most significant variables from the corpus study (see Section 3), we decided to test how

[19] www.survio.com

[20] Before starting the survey, the subjects were told that their participation was anonymous and voluntary. By participating they confirmed that they grew up in the Russian Federation and their educational level reached high-school level. Finally, they were asked to state their age and their gender identity. No IP address or any other identifying information was collected.

native speakers' responses varied depending on these two factors. A summary of the model is shown in Table 2.

Parametric coefficients	***Estimate***	***Odds ratios***	***Std. Error***	***p-value***
(Intercept)	0.25451		0.11714	0.0298
CONTVERB-*no*	-0.59246	0.55296	0.11306	1.60e-07
CONTVERB-*PF*	0.58278	1.79101	0.14317	4.69e-05
OBJMOD-*yes*	0.94596	2.57528	0.08831	<2e-16
Random effects	***Variance***		***Std. Dev.***	
PERSONID *(Intercept)*	0.1814		0.4259	

Table 2: Mixed effects logistic regression model

The intercept is defined by the following reference values for each parameter: ANSWER-*IPF*, CONTVERB-*IPF* and OBJMOD-*no*. The simple odds of ANSWER-*PF* vs. ANSWER-*IPF* at the reference level are greater than 1 (approximately 1.29), which means that the probability of ANSWER-*IPF* is greater when CONTVERB is *IPF* and there is no object modifier. As we can see from the estimates of the predictors, the value CONTVERB-*no* reduces the odds of pf answers and boosts the odds of ipf answers. On the contrary, both CONTVERB-*PF* and OBJMOD-*yes* boost the odds of ANSWER-*PF*. When the contiguous verb is pf the probability that the outcome is ANSWER-*PF* is 1.79 times higher than for CONTVERB-*IPF*; when the direct object has a modifier the odds of having ANSWER-*PF* are 2.58 times higher than with OBJMOD-*no*. We can therefore claim that the presence of a contiguous perfective verb and/or an object modifier encouraged the speakers to choose *kupit'*$_{\text{pf}}$. The absence of such factors or a contiguous ipf, on the other hand, more often resulted in the choice of *pokupat'*$_{\text{ipf}}$.

5 Conclusion

The first conclusion that can be drawn from this study is that the linguistic context (or "syntactic environment" [Grønn, 2004]) plays in most cases a marginal role in the choice of ipf OR, since only two out of the eight factors considered proved to be statistically noteworthy.

The statistical tests presented in Section 3 showed that only the factor CONTVERB (contiguous verb) significantly interacts with the other predictors. Its significance was confirmed by the logistic regression run on the questionnaire data (Section 4): the participants' answers depended on the presence of a contiguous verb in the sentence. In particular, native speakers tended to choose *kupit'*$_{\text{pf}}$ if the contiguous verb was pf, while a contextual ipf or the absence of a contiguous verb influenced the choice towards ipf. This is due to the fact that a series of pfs usually expresses narrative progression (as already stated in Section 1), whereas ipf is frequently used when the action cannot be precisely located on the time axis [Plungjan, 2004: 208].

The object modifier resulted as the second most significant factor from both the CART and the random forest models (see Section 3), while in the questionnaire answers it seemed to influence the participants' choice even more than contiguous verbs. To explain the preference towards ipf when the object is not modified, we can refer to the distinction between type and token reference. It has been claimed [Hedin, 2000] that by choosing ipf the speaker does not refer to a particular token, i.e. to a specific and concrete object or event, but to a 'type'. Even though such correlation is not always confirmed by empirical evidence [Mehlig, 2001], in our questionnaire pf appeared to be the most probable choice when the purchased item was more specific.

Finally, as far as the pair *pokupat'*$_{\text{ipf}}$-*kupit'*$_{\text{pf}}$ is concerned, the overall preference towards pf when denoting a complete action in the past (as emerged from the survey) could be due to the fact that the

perfective past tense is the preferred form to express telic actions [Gebert, 2014a], therefore *pokupat'*$_{ipf}$ is the marked choice in such contexts [Grønn, 2004].

Clearly, further research is needed to confirm our claims with respect to other aspectual pairs.

References

[1] Forsyth John. A Grammar of Aspect: Usage and Meaning in the Russian Verb. — Cambridge: Cambridge University Press, 1970.

[2] Gebert Lucyna. Linguistica slava tra slavistica e linguistica generale [Slavic Linguistics between Slavic studies and General Linguistics]. — Studi Slavistici [Slavic Studies], 2004. — Vol. 33(2). — P. 91–110.

[3] Gebert Lucyna. L'imperfettivo fattivo slavo e l'imperfetto narrativo romanzo: un confronto [Slavic factual imperfective and Romance narrative imperfect: a comparison] // L'architettura del testo. Studi contrastivi slavo-romanzi [The architecture of the text. Slavic-Romance contrastive studies]. — Alessandria: Edizioni dell'Orso, 2014a. — P. 3–17.

[4] Gebert Lucyna. Scelta aspettuale 'oggettiva' e 'soggettiva' e l'imperfettivo fattivo ['Objective' and 'subjective' aspectual choice and factual imperfective] // Studi italiani di linguistica slava. Strutture, uso e acquisizione [Italian studies of Slavic linguistics. Structures, use and acquisition]. — Firenze: Firenze University Press, 2014b. — P. 319–331.

[5] Glovinskaja Marina Ja. Semantičeskie tipy vidovych protivopostavlenij russkogo glagola [Semantic types of aspectual oppositions of the Russian verb]. — Moskva: Nauka, 1982.

[6] Grønn Atle. The semantics and pragmatics of the Russian Factual Imperfective (Doctoral thesis, University of Oslo). — 2004. —Access mode: https://www.duo.uio.no/handle/10852/76860.

[7] Hedin Eva. The type-referring function of the imperfective // Tense and aspect in the languages of Europe. – – Berlin: Moutin de Gruyter, 2000. — P. 227–264.

[8] Israeli Alina. Discourse analysis of Russian aspect: accent on creativity. — Journal of Slavic Linguistics, 1996. — Vol. 4(1). — P. 8–49.

[9] Israeli Alina. The choice of aspect in Russian verbs of communication: pragmatic contract. — Journal of Slavic Linguistics, 2001. — Vol. 9(1). — P. 49–98.

[10] Kreisberg Alina. Risultato e conseguenza nella semantica delle predicazioni [Result and consequence in the semantics of predications]. — Studi Slavistici [Slavic Studies], 2007. — Vol. 4. — P. 215–235.

[11] Mehlig Hans R. Verbal aspect and the referential status of verbal predicates: On aspect usage in Russian who-questions. — Journal of Slavic Linguistics, 2001. — Vol. 9(1) — P. 99–125.

[12] Mehlig Hans R. Obščefaktičeskoe i edinično-faktičeskoe značenija nesoveršennogo vida v russkom jazyke [General-factual and single-factual meanings of imperfective aspect in Russian]. — Vestnik Moskovskogo universiteta. Ser. 9. Filologija, 2013. — Vol. 4. — P. 19–46.

[13] Padučeva Elena V. Semantičeskie issledovanija [Semantic studies]. — Moskva: Jazyki slavjanskoj kul'tury, 1996.

[14] Plungjan Vladimir A. K diskursivnomu opisaniju aspektual'nych pokazatelej [Towards a discursive description of aspectual markers] // Tipologičeskie obosnovanija v grammatike. K 70-letiju professora V. S. Chrakovskogo [Typological reasoning in grammar. On the 70th birthday of Professor V. S. Chrakovskogo]. — Moskva: Znak, 2004. — P. 390–411.

[15] Rassudova Ol'ga P. Upotreblenie vidov glagola v sovremennom russkom jazyke [The use of verbal aspect in contemporary Russian]. — Moskva: Russkij jazyk, 1982.

[16] Reynolds Robert. Russian natural language processing for computer-assisted language learning: capturing the benefits of deep morphological analysis in real-life applications (Doctoral thesis, UiT The Arctic University of Norway). — 2016. —Access mode: https://hdl.handle.net/10037/9685.

[17] Šatunovskij Il'ja B. Problemy russkogo vida [Problems of Russian aspect]. — Moskva: Jazyki slavjanskich kul'tur, 2009.

[18] Sičinava Dmitrij V. Nesoveršennyj vid [The imperfective Aspect] // Materialy dlja proekta korpusnogo opisanija russkoj grammatiki [Materials for a corpus description of Russian grammar]. — 2013. — Access mode: http://rusgram.ru.

[19] Strobl Carolin, Malley James, Tutz Gherhard. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. — Psychological methods, 2009. — Vol. 14. — P. 323–348.

[20] Zaliznjak Anna A., Šmelev Aleksej D. Vvedenie v russkuju aspektologiju [An introduction to Russian aspectology]. — Moskva: Jazyki russkoj kul'tury, 2000.

Pragmatic Markers of Russian Everyday Speech: Quantitative Data

Bogdanova-Beglarian N. V.
Saint Petersburg State University
Saint Petersburg, Russia
n.bogdanova@spbu.ru

Blinova O. V.
Saint Petersburg State University,
HSE University, Saint Petersburg
Saint Petersburg, Russia
o.blinova@spbu.ru,
ovblinova@hse.ru

Sherstinova T. Ju.
HSE University, Saint Petersburg
Saint Petersburg State University,
Saint Petersburg, Russia
t.sherstinova@spbu.ru,
tsherstinova@hse.ru

Troshchenkova E. V.
Saint Petersburg State University
Saint Petersburg, Russia
e.troshchenkova@spbu.ru

Gorbunova D. A.
Saint Petersburg State University
Saint Petersburg, Russia
dgorbunova2@gmail.com

Zajdes K. D.
Saint Petersburg State University
Saint Petersburg, Russia
kristina.zaides@student.spbu.ru

Popova T. I.
Saint Petersburg State University
Saint Petersburg, Russia
tipopova13@gmail.com

Sulimova T. S.
Saint Petersburg State University
Saint Petersburg, Russia
spb.sulimova@gmail.com

Abstract

The article summarizes the results of a large research project dedicated to investigation of pragmatic markers (PM) in Russian everyday speech. Pragmatic markers are essential in spontaneous spoken discourse; thus, the quantitative data on their usage are necessary for solving both theoretical and practical issues related to the study of spoken communication. New results were obtained on the data of two speech corpora: "One Day of Speech" (ORD; mostly dialogues; the annotated subcorpus contains 321 504 tokens) and "Balanced Annotated Text Library" (SAT; monologues; the annotated subcorpus includes 50 128 tokens). Statistical data were calculated for PM in dialogic and monologic speech, pragmatic markers common in both types of speech (e. g., hesitative markers like vot, tam, tak) are identified, as well as PM that are the most typical for monologues (e. g., boundary markers like znachit, nu, vot, vs'o) or dialogue (e. g., 'xeno'-markers such as takoi, grit and metacommunicative markers like vidish', (ja) ne znaju). Special attention is given to the pragmatic markers usage in different communicative situations.

Keywords: Russian everyday speech; speech corpus; pragmatic marker; corpus annotation; monologue; dialogue

DOI: 10.28995/2075-7182-2021-20-119-126

Прагматические маркеры русской повседневной речи: количественные данные

Богданова-Бегларян Н. В.
СПбГУ
Санкт-Петербург, Россия
n.bogdanova@spbu.ru

Блинова О. В.
СПбГУ, НИУ ВШЭ СПб
Санкт-Петербург, Россия
o.blinova@spbu.ru,
ovblinova@hse.ru

Шерстинова Т. Ю.
НИУ ВШЭ СПб, СПбГУ,
Санкт-Петербург, Россия
tsherstinova@hse.ru
t.sherstinova@spbu.ru

Трощенкова Е. В.
СПбГУ
Санкт-Петербург, Россия
e.troshchenkova@spbu.ru

Горбунова Д. А.
СПбГУ
Санкт-Петербург, Россия
dgorbunova2@gmail.com

Зайдес К. Д.
СПбГУ
Санкт-Петербург, Россия
kristina.zaides@student.spbu.ru

Попова Т. И.
СПбГУ
Санкт-Петербург, Россия
tipopova13@gmail.com

Сулимова Т. С.
СПбГУ
Санкт-Петербург, Россия
spb.sulimova@gmail.com

Аннотация

В статье подводятся итоги большого исследовательского проекта, посвященного изучению прагматических маркеров (ПМ) русской повседневной речи. Прагматические маркеры являются неотъемлемой частью спонтанного устного дискурса, поэтому количественные данные об их реализации необходимы для решения как теоретических, так и практических задач, связанных с изучением речевой коммуникации. Новые результаты получены на материале двух речевых корпусов: «Один речевой день» (ОРД; преимущественно диалоги; аннотированный подкорпус включает 321 504 токена) и «Сбалансированная аннотированная текстотека» (САТ; монологи; в аннотированном подкорпусе 50 128 токенов). Статистические данные посчитаны для ПМ диалогической и монологической речи, выявлены маркеры, общие для обоих типов речи (хезитативы типа *вот*, *там, так*), а также те, что свойственны в большей степени монологу (разграничители типа *значит, ну вот, всё*) или диалогу (ксенопоказатели типа *такой, грит* и метакоммуникативы типа *видишь, (я) не знаю*). Особое внимание в работе уделяется употреблению прагматических маркеров в разных условиях коммуникации.

Ключевые слова: русская повседневная речь; речевой корпус; прагматический маркер; корпусная разметка; монолог; диалог

1 Введение

Под прагматическими маркерами (ПМ) в работе понимаются те единицы устной речи, которые прошли процесс прагматикализации, в результате практически утратили свое исходное лексическое и/или грамматическое значение и приобрели прагматическое, выполняя в дискурсе лишь определенные функции: маркируют границы реплик (*значит*, *ну вот*, *всё*) или ввод чужой речи (*такой/ая/ие*, *типа* (*того что*), *грит/грю/грят* и под.), вербализуют хезитацию говорящего (*это самое*, *как его* (*её, их*), *там*), его рефлексию по поводу сказанного либо готовящегося к произнесению (*или как там*? *или как сказать*? *или что*?) или самокоррекцию (*это*, *это самое*), выражают метакоммуникацию (*знаешь*, *понимаешь*, *да*, (*я*) *не знаю*, *что ещё*?) и т. п.

Отличия прагматических маркеров от дискурсивных слов, или дискурсивных маркеров (ДМ), под которыми и в зарубежной, и в отечественной лингвистике понимается очень широкий класс функциональных единиц (см., например: [1; 15; 5; 6; 9]), изложены в работе: [12]. Анализу ПМ в настоящем исследовании предшествовало тщательное ручное аннотирование корпус-

ного материала с разграничением ПМ и омонимичных им значимых единиц языка (см., например: [4]). Пилотная разметка корпусных данных выполнялась четырьмя независимыми экспертами, что позволило оценить состоятельность и объективность такой разметки.

Принимая во внимание вариативность ряда форм ПМ, для описания системы ПМ были введены понятия *базового варианта* и *структурных вариантов* (реальных употреблений) ПМ. Объектом количественного анализа в настоящей статье стали именно реальные употребления ПМ, с качественной стороны они описаны в специальном словаре прагматических маркеров[1].

2 Материал и методика исследования

В основу словника прагматических маркеров, включающего 60 единиц, была положена типология ПМ русской устной речи, описанная в [4; 13]. Типология разрабатывалась на основе эмпирических данных двух речевых корпусов: корпуса монологической речи «Сбалансированная аннотированная текстотека» (САТ) и корпуса русской повседневной речи «Один речевой день» (ОРД) (см.: о них: [8; 2; 3; 10; 11]). Были выделены и описаны следующие типы ПМ: хезитативы (Х), рефлексивы (Ф), метакоммуникативы (М), разграничители (Г), ксенопоказатели (К), аппроксиматоры (А), дейктические (Д) и ритмообразующие (Р) маркеры, маркеры самокоррекции (С) и заместители (З).

На двух аннотированных выборках (ОРД, 321 504 токена; и САТ, 50 128 токенов) получены количественные характеристики реальных употреблений ПМ и установлены корреляции между появлением ПМ в повседневной речи и различными факторами: тип речи (диалог – монолог), место и тип коммуникации, социальная роль говорящего в конкретном коммуникативном макроэпизоде, а также его социальные и психологические характеристики. В Таблице 1 представлены данные об объеме проанализированных подвыборок в разных социолектах.

Гендер				
мужчины		женщины		
171 497		158 390		
Возрастные группы				
младшая		средняя		старшая
143 805		67 089		118 993
Образование				
неоконч. высшее / среднее		среднее специальное		высшее
47 634		36 313		215 540
УРК (уровень речевой компетенции)				
низкий		средний		высокий
19 983		245 586		54 825
Профессиональные группы				
ГУМ	ЕСТ	ИНЖ	ИТ	ОБР
30 493	11 988	30 897	29 105	29 738
ОФ	РАБ	СИЛ	СО	ТВОР
8 989	13 290	1 587	27 041	17 581

Таблица 1: Объем подвыборок с учетом различных социальных параметров (в токенах)

Рассмотрим основные полученные результаты.

3 Общие частоты ПМ (реальные словоупотребления)

Анализ корпусного материала позволил получить статистику конкретных словоупотреблений ПМ (370 реализаций 60-ти базовых ПМ) (о базовых ПМ и их вариантах см.: [13]). На аннотиро-

[1] Мы отдаем себе отчет в том, что с момента записи наших корпусов прошло уже довольно много времени и полученные данные нельзя считать отражением *сегодняшней* устной коммуникации, но когда речь идет о корпусных массивах данных, которые требуют больших трудозатрат по сбору, систематизации, расшифровке, разметке и разнообразной обработке, иначе, по-видимому, и быть не может. Это близко к ситуации создания любого словаря, который к моменту выхода отчасти уже утрачивает адекватность реальному состоянию нашего языка и тем более нашей повседневной речи.

ванных подвыборках было выделено 370 и 133 ПМ соответственно. В среднем ПМ показали частоту 29 611 ipm (2,96 %) в диалогической речи и 19 300 ipm (1,93 %) в монологах. Видно, что ПМ в диалоге используются на треть чаще.

Верхние десять позиций общего частотного словника (топ-10) заняли следующие единицы: ВОТ, ТАМ, ДА, ТАК, КАК БЫ, ГОВОРИТ (ГРИТ), ЗНАЧИТ, ЗНАЕШЬ, НУ ВОТ, СЛУШАЙ. Абсолютным «лидером» во всех социолектах является маркер ВОТ – его частотность составляет около четверти от общего количества ПМ в речи всех говорящих (1827/20,7 %) (здесь и везде далее в скобках через слеш показаны абсолютные и относительные величины). Достаточно частотен также маркер ТАМ, имеющий ранг 2 и в общем словнике (976/11,1 %), и в большинстве социолектов. Частотность остальных маркеров верхней зоны частотного списка существенно ниже: от 3 до 5 % (от 243 до 462 употреблений).

Интересно, что это общее распределение употреблений ПМ полностью (хотя и не всегда в одинаковом порядке) повторилось в речи женщин, людей среднего возраста и носителей языка с высоким уровнем речевой компетенции (УРК). Верхние зоны частотных списков речи мужчин, а также носителей языка среднего возраста отличаются отсутствием в них контактоустанавливающих маркеров ЗНАЕШЬ и СЛУШАЙ, которые в общем списке имеют ранги, соответственно, 8 и 10 и входят практически во все другие частотные списки. Только в речи старших говорящих место маркера СЛУШАЙ занял контактный глагол ПОНИМАЕШЬ. В речи молодежи заметен высокий ранг маркеров КОРОЧЕ (5) и ТИПА (9).

Прагматический маркер КОРОЧЕ, именно в таком, редуцированном, варианте, а не в полном (базовом) КОРОЧЕ (ГОВОРЯ), в целом маркирует речь мужчин (ранг 8), младших говорящих (ранг 5), в том числе не имеющих высшего образования и высокого УРК, а также речь детей (ранг 10). Из профессиональных групп говорящих этот ПМ имеет высокий ранг 3 в речи рабочих и представителей силовых структур.

Дети, которые попали в число говорящих только как коммуниканты в корпусе ОРД, имеют свой собственный частотный список ПМ (топ-10), не совпадающий с данными по взрослой речи: ТАМ, КАК БЫ, ВОТ, НЕ ЗНАЮ, ДА, ТАК, ТИПА, ВОТ ТАК ВОТ, ЗНАЕТЕ, КОРОЧЕ.

Маркер ТИПА можно признать специфическим показателем речи младших говорящих (ранг 9), в том числе детей (ранг 7), а также речи представителей силовых структур (ранг 6) и творческой интеллигенции (ранг 10).

При рассмотрении условий коммуникации наибольшие отличия демонстрирует общение в кафе и ресторанах. В целом по корпусу (преимущественно это общение *дома, в офисе, на улице, в казарме, в медицинских учреждениях*) ранг 1 имеет маркер ВОТ, ранг 2, за исключением казармы, – маркер ТАМ. В «казарменном» общении на втором месте в частотном списке оказался маркер КОРОЧЕ, совсем немного уступающий по употребительности маркеру ВОТ (31/19,4 vs 32/20 %). После ТАМ в этом списке следуют маркеры ТИПА и КАК БЫ, также весьма типичные в разговорах между курсантами. *Кафе* и *рестораны* дают совсем другую картину: на первом месте в соответствующих словниках – с одинаковой частотой – находятся маркеры ТАМ и ДА (по 18/17,4 %), затем идет ПМ СЛУШАЙ (13/12,4 %) и лишь затем ВОТ (9/8,6 %).

«Лидирующее» положение маркера ВОТ (ранг 1) сохраняется во всех ***социальных ролях*** говорящих. Особенно высока доля этого ПМ в речи «родителей» (132/40 %). Ранг 2 почти во всех социальных ролях имеет маркер ТАМ. Единственным исключением стала роль «однокурсника», в которой ТАМ отошло на четвертую позицию, уступив место маркерам ДА и КОРОЧЕ. Очень распространенный в нашей повседневной речи ПМ КОРОЧЕ встретился еще в верхней зоне частотного списка социальной роли «друга» (ранг 4), а позицию 7 в большинстве ролей («друг», «коллега», «муж»/«жена», «однокурсник» и «родители») на удивление устойчиво занимает маркер КАК БЫ. Самый высокий его ранг 3 зафиксирован в роли «ребенка» («сын»/«дочь»), в социальной роли «подруги» этот маркер имеет ранг 4, а в роли «клиент – сервис» он вообще не попал в верхнюю зону частотного списка прагматических маркеров.

Метакоммуникативы вошли в верхнюю зону (топ-10) частотных списков ПМ для всех социальных ролей, особенно их много в роли «родителей» (ПОНИМАЕШЬ, ЗНАЕШЬ, ПРЕДСТАВЛЯЕШЬ, СЛУШАЙ) и «подруги» (ЗНАЕШЬ, ПОНИМАЕШЬ, СЛУШАЙ). Видимо, именно в этих ролях говорящий в наибольшей степени стремится установить, а затем и удержать контакт с собеседником.

Интересно было также проанализировать распределение в разных социальных ролях маркеров-*ксенопоказателей* ГОВОРИТ (ГРИТ) и ГОВОРЮ (ГРЮ), свидетельствующих о том, что говорящие в этих ролях часто пересказывают чужие (или свои) слова и мнения. Оба эти ПМ обнаружились в социальной роли «коллеги» (ранги 3 и 5 соответственно), в ролях «мужа»/«жены» и «клиент – сервис» в топ-10 вошел только маркер ГОВОРИТ (ранги 8 и 3 соответственно). В других ролях эти ПМ в верхних зонах частотных списков не отмечены.

Распределение употреблений ПМ с учетом ***психотипа*** говорящего (экстраверты – интроверты) показало, что первые две позиции в частотных списках маркеров в речи обоих психотипов занимают единицы ВОТ и ТАМ, причем первых в 2-3 раза больше, чем вторых: 342/21,5 vs 166/10,5 % у экстравертов, 210/26,9 vs 70/8,95 % у интровертов. Речь экстравертов отличает также наличие в зоне топ-5 ксенопоказателя ГОВОРИТ (ранг 3), а речь интровертов – наличие в той же зоне метакоммуникативов ДА и ПОНИМАЕШЬ (ранги, соответственно, 3 и 4).

Подключение к анализу материала данных о ***темпераменте*** и ***уровне невротизма (нейротизма)*** говорящего не поколебало высокого первого ранга маркера ВОТ во всех случаях, хотя доля его оказалась весьма низкой (по сравнению с другими группами) в речи говорящих с низким уровнем невротизма (65/12,8 %) и в смешанной группе холериков-флегматиков (14/12,4 %). И эти же две группы говорящих «любят» маркер КОРОЧЕ: его ранг 3 в обоих случаях. Высокий ранг 2 маркера ТАМ сохраняется в речи говорящих с низким и средним уровнем невротизма, а также в речи сангвиников, флегматиков и смешанной группы холериков-флегматиков. В других группах говорящих на эту позицию выдвинулся маркер ДА (говорящие с высоким уровнем невротизма, меланхолики и холерики).

4 Словарь прагматических маркеров

Наличие различных функций ПМ в устном тексте, часто совмещенных в одной единице, – при размытом, ослабленном или вовсе отсутствующем семантическом их наполнении – вынудило поставить задачу создания специального словаря таких единиц (далее – Словарь ПМ), который должен отражать богатство и разнообразие как самого списка ПМ, так и функций, выполняемых ими в устной речи. Такой ресурс может быть исключительно полезен специалистам самых разных направлений.

К настоящему времени словарь ПМ создан, он включает в себя перечень наиболее частотных, регулярно используемых в русской устной речи, прагматических маркеров, с указанием их типа, функции и примеров употребления. Словарь подготовлен в двух версиях – бумажном варианте [7] и электронной версии с аудиопримерами. Словарные статьи в Словаре ПМ построены как лексикографические эссе, что обусловливается спецификой самого материала (ср.: «Путеводитель по дискурсивным словам русского языка» [1]). В таких эссе дается описание функций (прагматических значений) и особенностей функционирования всех выявленных ПМ. При этом разные функции ПМ в спонтанном тексте часто совмещаются в одной единице – при размытом, повторим, ослабленном или вовсе отсутствующем семантическом ее наполнении, т. е. ПМ оказались почти всегда полифункциональны и очень зависимы в своем статусе от контекста, что еще более укрепило нас в осознании необходимости такого словаря.

Словарь ПМ может быть полезен специалистам самого разного толка: собственно лингвистам, исследователям повседневной русской речи (коллоквиалистика, когнитивистика, социо- и психолингвистика); создателям грамматики речи (или модели языка, основанной на употреблении – *usage-based theory*, см.: [14]), которая, вне всякого сомнения, отличается от грамматики языка; специалистам по корпусной лингвистике, разрабатывающим системы автоматического аннотирования и анализа корпусного материала на разных уровнях; переводчикам спонтанных текстов на другие языки, хотя бы в рамках художественного произведения, при передаче речи персонажей; преподавателям русского языка иностранцам, которые вынуждены учиться воспринимать и правильно понимать нашу спонтанную речь как устно, так и письменно, при чтении русскоязычных текстов. В ряду других словарей, построенных на материале устных корпусов, Словарь ПМ отражает лексическое и дискурсивное своеобразие повседневной русской речи.

Электронная версия словаря доступна онлайн на сайте https://www.ord-multimedia-dict.com/.

5 Заключение

В статье представлены количественные данные о частоте использования прагматических маркеров устной речи, полученные на основе двух проаннотированных речевых корпусов (ОРД и САТ).

Основой для анализа стали ранжированные частотные списки ПМ с информацией об относительной частоте их употребления в диалоге (выборка ОРД) и монологе (выборка САТ); а также частотные списки функциональных типов прагматических маркеров с привязкой к типу речи (диалог vs монолог). Кроме того, получены частотные списки ПМ для разных условий коммуникации (т. е. с учетом типа разговора, социальной роли говорящего и др.) и для разных групп говорящих (социолектов). Все эти данные суммированы в словаре ПМ, имеющем, кроме бумажной, электронную версию, дающую возможность прослушать аудиопримеры на каждый тип ПМ.

Качественный анализ конкретных употреблений ПМ вынужденно оставлен за рамками настоящей статьи, равно как и количественные данные по распределению базовых вариантов ПМ и конкретных их функциональных типов. Это может стать предметом отдельных статей.

В целом анализ корпусного материала показал, что ПМ действительно регулярно встречаются в повседневной речи говорящих всех социальных групп и во всех без исключения коммуникативных ситуациях. Однако частота использования определенных базовых ПМ и их вариантов меняется в зависимости от характеристик как самих говорящих, так и условий коммуникации. Наиболее типичные тренды использования ПМ описаны в настоящей статье. Следует принимать во внимание, что в речи отдельных говорящих могут наблюдаться определенные «выбросы», отличающие идиостиль от «среднего по группе». Индивидуальная вариативность ПМ и ее мера заслуживает специального рассмотрения.

Полученные данные расширяют теоретические представления об употреблении ПМ в реальном повседневном общении и могут быть использованы в разнообразных практических приложениях – от разработки систем в области речевых технологий до задач лингвистической экспертизы, практики перевода и преподавания русского как иностранного.

Благодарность

Исследование проведено при финансовой поддержке гранта Санкт-Петербургского государственного университета (проект № 75254082 «Моделирование коммуникативного поведения жителей российского мегаполиса в социально-речевом и прагматическом аспектах с привлечением методов искусственного интеллекта»).

References

[1] Baranov A. N., Plungyan V. A., Rakhilina E. V. (1993), Russian Discourse Words Guide [Putevoditel’ po diskursivnym slovam russkogo jazyka]. Moscow.

[2] Bogdanova-Beglarian N.V., Blinova O.V., Sherstinova T.Yu., Martynenko G.Ya. (2019a), Corpus of Russian Everyday Speech “One Day of Speech”: present state and prospects [Korpus russkogo jazyka povsednevnogo obshchenia «Odin rechevoj den’»: tekushchee sostojanie i perspektivy] // Proceedings of the V V. Vinogradov Russian Language Institute. Vol. 21. Russian National Corpus: Research and Development [Trudy In-ta russkogo jazyka im. V. V. Vinogradova. Vyp. 21. Nacional’nyj korpus russkogo jazyka: issledovania i razrabotki], Moscow, pp. 101–110.

[3] Bogdanova-Beglarian N.V., Blinova O.V., Zaides K.D., Sherstinova, T.Yu. (2019b), “Balanced Annotated Text Library” (SAT; monologues): Studying the Specifics of Russian Monological Speech [“Sbalansirovannaya annotirovannaya tekstoteka” (SAT): izuchenie spetsifiki russkoj monologicheskoj rechi] // Proceedings of the V. V. Vinogradov Russian Language Institute. Vol. 21. Russian National Corpus: Research and Development [Trudy In-ta russkogo jazyka im. V. V. Vinogradova. Vyp. 21. Nacional’nyj korpus russkogo jazyka: issledovania i razrabotki]. Moscow, pp. 111–126.

[4] Bogdanova-Beglarian N.V., Blinova O.V., Martynenko G.Ya., Sherstinova, T.Yu., Zaides K.D., Popova T.I. (2019c), Annotation of Pragmatic Markers in the Russian Speech Corpus: Problems, Searches, Solutions, Results [Annotirovanie pragmaticheskikh markerov v russkom rechevom korpuse: problemy, poiski, reshenia, rezul’taty] // Computational Linguistics and Intellectual Technologies [Kompjuternaya lingvistika i intellektual’nye tekhnologii]: Proceedings of the International Conference “Dialogue 2019”. Vol. 18 (25). Moscow, pp. 72–85.

[5] Discursive Words of the Russian language: An Experience of Context-Semantic Description [Diskursivnye slova russkogo jazyka: Opyt kontekstno-semanticheskogo opisani].(1998), Moscow.
[6] Discursive Words of the Russian language: Variation and Semantic Unity. Collection of articles [Diskursivnye slova russkogo jazyka: varjirovanie i semanticheskoe jedinstvo. Sb. Statej] (2003), Moscow, 207 p.
[7] Pragmatic Markers of Russian Everyday Speech: Dictionary-Monograph Pragmaticheskie [Markery russkoj povsednevnoj rechi: slovar'-monografia] (2021), St. Petersburg. In Print.
[8] Asinovsky A.S., Bogdanova N.V., Rusakova M.V., Ryko A.I., Stepanova S.B., Sherstinova T.Yu. (2009), The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation // Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 5729. Springer, — Berlin-Heidelberg, 2009, pp. 250–257.
[9] Beliao Julie, Lacheret Anne. Disfluency and Discursive Markers: when Prosody and Syntax Plan Discourse // DiSS 2013: The 6th Workshop on Disfluency in Spontaneous Speech, — Stockholm, Sweden. № 54 (1), 2013, pp. 5–9.
[10] Bogdanova N.V., Sherstinova T.Yu., Blinova O.V., Martynenko G.Yu. An Exploratory Study on Sociolinguistic Variation of Spoken Russian // SPECOM 2016. Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. — Springer, Switzerland,2016. — pp. 100–107.
[11] Bogdanova N.V., Sherstinova T.Yu., Blinova O.V., Baeva E.M., Martynenko G.Ya, Ryko A.I. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. — Springer, Switzerland, 2016b — pp. 659 666.
[12] Bogdanova N.V., Filyasova Yu.A. Discourse vs Pragmatic Markers: A Contrastive Terminological Study // 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts, SGEM 2018. — Vienna ART Conference Proceedings. Vol. 5, Iss. 3.1, 2018. — pp. 123–130.
[13] Bogdanova N.V., Blinova O.V., Sherstinova T.Yu., Troshchenkova E.V., Gorbunova D.A., Zaides K.D. Pragmatic Markers of Russian Everyday Speech: the Revised Typology and Corpus-Based Study // Proceedings of the 25th Conference of Open Innovations Association, FRUCT. — Helsinki, Finland,2018. — pp. 57–63.
[14] Boye Kasper, Harder Peter. A Usage-based Theory of Grammatical Status and Grammaticalization // Language, 88 (1), 2012. — pp. 1–44.
[15] Shiffrin Deborah. Discourse Markers. Cambridge University Press, Cambridge, UK, 1996.

Список литературы

[1] Баранов А. Н., Плунгян В. А., Рахилина Е. В. Путеводитель по дискурсивным словам русского языка. — М.: Помовский и партнеры, 1993.
[2] Богданова-Бегларян Н. В., Блинова О. В., Мартыненко Г. Я., Шерстинова Т. Ю. Корпус русского языка повседневного общения «Один речевой день»: текущее состояние и перспективы // Труды Ин-та русского языка им. В. В. Виноградова. — Вып. 21. Национальный корпус русского языка: исследования и разработки / Гл. ред. А. М. Молдован. Отв. ред. вып. В. А. Плунгян. — М., 2019а, сс. 101–110.
[3] Богданова-Бегларян Н. В., Блинова О. В., Зайдес К. Д., Шерстинова Т. Ю. Корпус «Сбалансированная аннотированная текстотека» (САТ): изучение специфики русской монологической речи // Труды Ин-та русского языка им. В. В. Виноградова. — Вып. 21. Национальный корпус русского языка: исследования и разработки / Гл. ред. А. М. Молдован. Отв. ред. вып. В. А. Плунгян. — М., 2019б, сс. 111–126.
[4] Богданова-Бегларян Н. В., Блинова О. В., Мартыненко Г. Я., Шерстинова Т. Ю., Зайдес К. Д., Попова Т. И. Аннотирование прагматических маркеров в русском речевом корпусе: проблемы, поиски, решения, результаты // Компьютерная лингвистика и интеллектуальные технологии: По м-лам ежегодной международной конференции «Диалог» / Гл. ред. В. П. Селегей. — Вып. 18 (25), 2019в, сс. 72–85.
[5] Дискурсивные слова русского языка: Опыт контекстно-семантического описания / Под ред. К. Л. Киселевой, Д. Пайара. — М.: Метатекст, 1998.
[6] Дискурсивные слова русского языка: варьирование и семантическое единство. Сб. статей / Под ред. К. Л. Киселевой, Д. Пайара. — М.: Азбуковник, 2003.
[7] Прагматические маркеры русской повседневной речи: Словарь-монография / Сост., отв. ред. и автор предисловия Н. В. Богданова-Бегларян. — СПб.: Нестор-История, 2021. — В печати.
[8] Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., Sherstinova, T. The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation // Matoušek, V., Mautner, P. (eds.) TSD 2009. — LNAI, Vol. 5729. — Springer, Berlin-Heidelberg, 2009. — P. 250–257.

[9] Beliao Julie, Lacheret Anne. Disfluency and Discursive Markers: when Prosody and Syntax Plan Discourse // DiSS 2013: The 6th Workshop on Disfluency in Spontaneous Speech, Stockholm, Sweden. — № 54 (1), 2013. — P. 5–9.
[10] Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G. An Exploratory Study on Sociolinguistic Variation of Spoken Russian // SPECOM 2016. Lecture Notes in Artificial Intelligence. — LNAI, Vol. 9811. — Springer, Switzerland, 2016a. — P. 100–107.
[11] Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Baeva, E., Martynenko, G., Ryko, A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016, Lecture Notes in Artificial Intelligence. — LNAI, vol. 9811. — Springer, Switzerland, 2016b. — P. 659–666.
[12] Bogdanova-Beglarian, N. V., Filyasova, Yu. A. Discourse vs Pragmatic Markers: A Contrastive Terminological Study // 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts, SGEM 2018. Vienna ART Conference Proceedings, 19-21 March, 2018. — Vol. 5, Iss. 3.1. — P. 123–130.
[13] Bogdanova-Beglarian, N. V., Blinova, O. V., Sherstinova, T. Yu., Troshchenkova, E. V., Gorbunova, D. A., Zaides, K. D. Pragmatic Markers of Russian Everyday Speech: the Revised Typology and Corpus-Based Study // Proceedings of the 25th Conference of Open Innovations Association FRUCT / S. Balandin, V. Niemi, T. Tuytina (eds.). — Helsinki, Finland, 2019. — P. 57–63.
[14] Boye Kasper, Harder Peter. A Usage-based Theory of Grammatical Status and Grammaticalization // Language. — 88 (1) — 2012. — P. 1–44.
[15] Shiffrin Deborah. Discourse markers. — Cambridge University Press, Cambridge, UK, 1996.

Semantic Representations in Computational and Theoretical Linguistics: the Potential for Mutual Enrichment[1]

Igor M. Boguslavsky
A. A. Kharkevich Institute for Information Transmission Problems, B. Karetnyj 15, Moscow, 103051, Moscow, Russia;
Universidad Politécnica de Madrid, 28040 Madrid, Spain
bogus@iitp.ru

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

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

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

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

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

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

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

Abstract

Research in semantics is actively conducted both in theoretical and computational linguistics, but the formulation of tasks, objectives and results of semantic research in the two communities are usually largely different. As a step towards reducing this gap and increasing the awareness of theoretical linguists about what computational linguists are doing, we examine meaning representation approaches in computational linguistics and contrast them with how this is done within one of the best-known theoretical approaches – the Meaning ⇔Text Theory.

Keywords: semantic representation, computational linguistics, theoretical linguistics, Meaning ⇔ Text Theory
DOI: 10.28995/2075-7182-2021-20-127-141

[1] Публикуется по специальному решению Редсовета

Семантические представления в компьютерной и теоретической лингвистике: потенциал взаимного обогащения

Богуславский И. М., Диконов В. Г., Иншакова Е. С., Иомдин Л. Л., Лазурский А. В., Рыгаев И. П., Тимошенко С. П., Фролова Т. И.

Аннотация

Семантические исследования активно ведутся как в теоретической, так и в компьютерной лингвистике, но постановки задач, цели и результаты таких исследований пересекаются слабо. В качестве шага, направленного на то, чтобы сократить этот разрыв и сделать более понятным для теоретических лингвистов то, что делают их коллеги в компьютерной лингвистике, мы рассматриваем некоторые способы представления значения предложения, принятые в компьютерной лингвистике, и сопоставляем их с тем, как представляются соответствующие явления в рамках одной из известных теоретических моделей – в модели "Смысл ⇔ Текст".

Ключевые слова: семантическое представление, компьютерная лингвистика, теоретическая лингвистика, модель «Смысл ⇔Текст»

1 Introductory remarks

From the start, the Dialogue conference has been a platform destined to bring together the two parts of our broad linguistic community – computational linguists (including mathematicians and engineers working in the field of computer natural language processing) and, so to speak, "linguistic" linguists engaged in theoretical and descriptive linguistics. Sadly, however, we have to state that the two communities have little understanding of, and little interest in, each other. If, for the sake of definiteness, we confine ourselves to semantics, we will see that even though semantics is actively developed by both communities, the tasks set by them, the objectives, and the results hardly ever come together. A mere look at the topics presented to the linguistic and computational linguistic sections of Dialogue – and to other similar forums – will suffice to conclude that the presenters speak different scientific languages and have little in common.

This cannot be accepted as normal: ultimately, we have one and the same object of research, the natural language. So, we should not simply acknowledge the difference in objectives, approaches and methods between computational linguistics (CL) and theoretical linguistics (TL) but, rather, endeavor to bridge this gap and raise the awareness of the “opposing sides” about these issues. Both communities will win if they get better acquainted with the practices and solutions of their colleagues from the other group.

The state-of-affair, as we see it, is that TL papers oriented at computational linguistics do appear from time to time, while there is virtually no movement in the opposite direction. We are unaware of any paper by CL workers targeted toward the TL community, attempting to compare the respective approaches. With this paper, we are making a step in this direction. It should be emphasized that the paper is primarily intended for non-CL linguists. Computational linguists will find no new results and no answers about the potential of the phenomena under discussion for the purposes of natural language processing. We hope, however, that the paper will offer something new to TL researchers. We are striving to compare the different approaches to an object which both communities consider as relating to them. We will talk to theoretical linguists about the objects that are relevant to them from the viewpoint which is close to them. Specifically, we will discuss the object very familiar to linguists interested in formal methods of linguistic description – the semantic representation of the sentence, as well as the information which this representation permits to express. Methods of semantic representations are of great interest to CL, too, because more and more applications require that text meaning be taken into account. Yet, as TL and CL have rather different objectives, whatever is interesting and important for one domain may be irrelevant for the other domain.

We will briefly present several types of semantic representations used in CL, comment on their similarities and differences and compare them with the approaches accepted in TL (exemplified by the Meaning ⇔ Text theory). A clear formulation of these similarities and differences seems important in order for theoretical linguists to better understand what are their neighbors engaged in, and to acquire a more stereoscopic view of their object of research. The readers interested in a more detailed overview of semantic representations made in CL perspective are referred to Abend, Rappaport 2017 and Bos, Abzianidze 2019.

2 Semantic Representation Requirements

We have to emphasize from the start that different research groups use different names for representations with which they work but we will call all of them s e m a n t i c r e p r e s e n t a t i o n s, or SemR, for consistency. The objective of any SemR is to reflect the meaning of a sentence but the amount of this reflection may vary depending on the particular purpose for which a given SemR is primarily intended. In particular, if SemR is to be used for inferences of any kind, it is highly desirable that it include certain logical information, e.g. negation and quantifiers, as well as some information on the lexical meaning, such as information on implicative predicates or presuppositions. For other purposes such information may be irrelevant and is not included into SemR.

This orientation to the purpose reflects certain important differences between CL and TL approaches. Normally, a theoretical linguist is not faced with the question why a phenomenon should be dealt with: if a phenomenon exists then it needs to be described. Contrariwise, CL takes into account two issues before tackling a phenomenon. First, it is considered to be of great importance whether the phenomenon is essential for applications. If application operation is not largely affected by the phenomenon, it is likely to be ignored rather than represented in SemR. Second, SemRs are often discussed in CL from the standpoint of corpus annotation, since the developers expect SemRs to be managed by machine learning techniques, which implies the need for a large corpus marked up with such representations.

In such discussions, the focus is often placed on what information, potentially useful for applications, can be quickly and uniformly marked up by annotators, rather than on what semantic information is conveyed by natural language sentences and therefore should be reflected in the SemR, as is customary for TL. Accordingly, the simplicity of SemR is often considered to be a vital advantage, because it affects the required level of annotators and the expected markup speed. To give an example, the Prague tectogrammatical corpus (Hajič 2002, Hajič et al. 2001) was developed by specially trained annotators, while semantic markup of the UCCA (Abend, Rappaport 2013a, b) and UDS corpora (White et al. 2017) was outsourced to much less skilled personnel. The GBM project took a middle stand as it used both qualified experts and unskilled annotators (Bos et al. 2017).

The volume of this paper does not allow us to review all types of SemR presented in the literature: there are quite a few of them, see e.g. AMR (Banarescu et al., 2013), Bridge (Bobrow et al. 2007), Compreno (Anisimovich et al. 2012), FrameNet (Baker et al. 1998), GMB (Bos et al., 2017), MRS (Copestake et al. 2005), OntoNotes (Hovy et al. 2006), PDC (Hajič 2002, Hajič et al. 2001), UNL (Uchida et al. 2005), OntoAgent (McShane, Nirenburg 2012), UCCA (Abend, Rappaport 2013a, b), UDS (White et al. 2017), SemETAP (Boguslavsky et al. 2020, Boguslavsky 2021), and de Salvo Braz et al. 2015. In section 4 below we will illustrate some representative approaches.

Before we proceed with the discussion, one important remark has to be made. Recently, an approach to semantics has gained popularity in CL, which avoids presenting the meaning of a linguistic object in the form of an explicit structure understandable by a human. This approach is based on the distributive hypothesis, which maintains that units occurring in similar contexts have similar meanings. In this approach, the meanings of words and even larger linguistic units are presented as vectors (ordered sequences of figures) built by analyzing the distribution of a given unit in a big collection of texts. Such vectors allow a quantitative assessment of the semantic proximity between the different units and are instrumental in the solution of certain other tasks (Lenci 2008), but they cannot help one obtain an exact idea of what a unit really means. Consequently, they give no clear answer to the essential questions of linguistic semantics: What does the word A mean? How does the meaning of A differ from the meaning of B? Therefore, vector methods of meaning representation cannot be considered transparent. As the aim of this paper is to discuss the compatible approaches to semantics in CL and TL, we will not touch on distributive models here.

3 Types of information represented in SemR

It is much more difficult to decide what information should be present in the semantic structure of a sentence than to answer a similar question about the syntactic structure. Indeed, in the case of syntactic structure, it is at least clear of what building blocks it should be made. The goal of the syntactic structure is to link **the words of the sentence**, which are observed directly. Of course, it is not always easy to construct an adequate structure but we know what units should be used. It is not at all obvious for the

semantic structure: what are the semantic elements that constitute a SemR of a sentence? How are they related to its words? How are these elements linked with each other? What information should appear in a SemR?

Below, we will comment on certain aspects essential in the comparison of different CL approaches to the construction of the SemR.

3.1 Which semantics is reflected by a SemR?

All SemRs referred to in Section 2 strive to abstract away from grammatical and syntactic idiosyncrasies inherent in natural languages. In particular, this is manifested in the fact that grammar words (auxiliary and support verbs, strongly governed prepositions and conjunctions, or articles) are removed from the sentence, passive constructions are replaced by active ones, nouns derived from verbs are reduced to the base verbs, etc. In many cases, such techniques help produce similar representations for different but synonymous constructions, and different representations for syntactically close but semantically diverging constructions. In this respect, many approaches view their semantic constructions as something close to deep syntactic constructions as understood in the Meaning ⇔ Text theory or the Prague school.

Such transformations are primarily confined to morphological and syntactic phenomena. The word semantics is often taken into account only partially: the words that are close in meaning receive the same representation or are reduced to one group of synonyms (WordNet synsets), or one frame of the FrameNet, or a different group of meanings. For example, the AMR approach builds one and the same SemR for sentences like *It may rain, It might rain, Rain is possible, It's possible that it will rain* (Banarescu *et al*. 2019). The attempts to represent the word meanings explicitly are very limited. AMR uses transparent lexical derivational models like *-able* or *-full*. The noun group *an edible sandwich* receives the same structure as *a sandwich that can be eaten.* To some extent, syntactic derivation is considered: *an attractive man* is represented in the same way as *a man who attracts.* In the structure used by Compreno, a word is not only referred to a class of the semantic hierarchy but may also be assigned a semantic feature, or semanteme, which explicates some component of the word meaning (Anisimovich et al. 2012).

A more detailed description of word semantics is represented in structures used in Bridge (Bobrow et al. 2007), OntoAgent (McShane, Nirenburg 2012) and SemETAP (Boguslavsky 2017, Boguslavsky et al. 2020). Among other things, Bridge takes account of implicative components of the word meaning, which specify the implications presumed by the word (as in *John managed to leave* ⟹ *John left*). OntoAgent and SemETAP decompose the word meaning into smaller semantic elements when necessary.

3.2 SemR nodes

There are two major approaches to the selection of units to be used as SemR nodes. Within the first approach the nodes are natural language words. In this case, the dictionary of a given natural language is frequently linked to a special lexical resource (which may have different names – ontology, dictionary of predicates, concepts, semantic classes or frames, WordNet) used to refer the words to a more general taxonomic category. For example, FrameNet refers the words *give*, *donate*, *gift* to the Giving frame, or concept. SemRs of the Bridge system refer each word to a set of WordNet synsets, in which at least one of the word's meanings is represented. Compreno associates the words with a specially designed hierarchy of semantic classes. Other versions of this approach, however, do not link the words as SemR units with any abstract conceptual entities. These are the cases of tectogrammatical structures of Prague Dependency Treebank (PDT) or Discourse Representation Structures of the GMB corpus.

Within the second approach, the nodes of SemR are elements of a semantic metalanguge (ontology). This can be exemplified by SemRs projects OntoAgent, SemETAP, and UNL. AMR structures occupy an intermediate position. The major part of semantic elements is composed of English words but there are several specially designed predicates, such as street-address, which has a list of arguments including house number, street, city, state, and zip code.

It is worth noting that the ontology (or another similar resource) plays different roles in these two approaches. In the first approach, the reference to a frame simply supplements the word in the SemR of the sentence but does not supersede it. Let us come back to the Giving frame above. The fact that it is referred to by the verbs *give, donate,* or *gift*, permits us to see and describe in a compact way the common features of the three verbs: they describe similar, though not identical situations and have the same sets of slots (frame elements). However, the semantic differences between the verbs remain unexplained. So

if a SemR contained the frame Giving instead of the verb *donate*, we would lose a part of the meaning because there is no full semantic identity between giving and donating. The approaches using FrameNet frames do not do that: the structure retains the initial verb (*donate*) beside the reference to the frame. In OntoAgent or SemETAP approaches, the ontology also contains concepts of the Giving type, but in the SemR such a concept will replace the verb *donate.* However, no meaning loss will occur, since the semantic representative of the verb *donate* is not the concept Giving alone but a certain construction composed of several concepts, which will explicitly express the semantic difference of *donate* from Giving.

3.3 Relations between SemR nodes

SemR elements are linked by relations. Of special importance are the relations between the predicates and their arguments. It is essential to show "who is doing what with which to whom", even though the boundaries between the argument (actant, or core) relations and non-argument ones (circumstantial, peripheral, or non-core) may be drawn variously.

All SemRs reflect verbal arguments. In many cases, arguments of some nouns and adjectives are present. We did not observe any SemRs (with the notable exception of SemETAP) that provided arguments of adverbs, despite the fact that such arguments are commonplace, see e.g. *far (from), independently (of), similarly (to), comparably (with), more (than), up (the hill)* etc. The relations between the predicates and their arguments are often marked with very general semantic roles, such as Agent, Theme, Patient etc. or using asemantic tags like ARG0, ARG1, ARG2 (as is common in PropBank or AMR). FrameNet takes a special stand, since many of the frame elements are specific for individual frames. For instance, the frame Arrest introduces the following specific argument relations (core frame elements): Authorities, Charges, Offence, Suspect.

To indicate non-argument semantic links between SemR elements, varied sets of semantic roles are used. One of the most popular sets of roles is proposed by the VerbNet project (Kipper et al. 2006).

3.4 Other types of information in SemR

In addition to the semantic relations between its elements, SemR may contain other types of data. We mentioned in Section 3.2 above that the SemR may contain a reference from a sentence word to a semantic element of a higher level of abstraction: a frame, a WordNet synset, or a class in the semantic hierarchy. Other sorts of data that can be marked in a SemR include information on anaphora or coreference (AMR, GMB, PDC, Compreno, OntoAgent, SemETAP, UNL). Such information may involve finding the antecedents of anaphoric pronouns (*When Mary woke up, she* [Mary or another person] *felt a sore throat*) or restoring the syntactically conditioned zero anaphora (*Having received* [Pete] *a bad mark, Pete decided to start* [Pete] *working* [Pete] *hard*).

In logically oriented SemRs, a logical structure is marked, which may include the negation, quantifiers and their scopes (GMB). The tectogrammatical structures of PDT are marked for thematic-rhematic articulation and the deep word order. In AMR structures, named entities are referred to the respective Wikipedia article:

```
(s / ship
  :wiki "RMS_Titanic"
  :name (n / name
        :op1 "Titanic"))
```

On the other hand, some SemRs are left with unmarked grammatical meanings, such as the number, tense, or definiteness/indefiniteness expressed by articles (AMR).

3.5 Reliance on a specific linguistic theory

Certain SemRs are built in accordance with a specific linguistic theory. So, PDT structures are oriented to Functional Generative Description, developed by the Prague School of Linguistics (Sgall, Hajičova and Panevová, 1986). Minimal Recursion Semantics (MRS) structures are closely connected with the Head-driven Phrase Structure Grammar (Pollard, Sag 1987). SemRs of GMB rely on the Discourse Representation Theory, or DRT (Kamp and Reyle 1993). SemETAP has been conceived within the

framework of the Meaning ⇔ Text theory (Mel'čuk 1974, 2012, 2013, 2014). By reasons of space, we cannot discuss this topic in more detail.

4 Examples of semantic structures

Having illustrated some of the general approaches to SemR construction we will now give a brief account of other approaches that are relatively rarely discussed.

4.1 Bridge (Bobrow et al. 2007)

The Bridge system, developed in Palo Alto Research Center (PARC), is designed to convert sentences into abstract knowledge representations (AKR). An AKR consists of three main parts: the conceptual structure, the contextual structure, and the temporal structure. The conceptual structure describes objects, their properties and the events in which they take part. The contextual structure relates this information to the real word, communicating whether the propositions mentioned in the conceptual structure exist in reality and reflecting the presuppositions of objects' existence. The temporal structure refers the time of the events mentioned to the moment of speech.

The content words are referred to the ontology, which in this case is WordNet. Every word receives references to all synsets in which they are present (if any).

The following example of a complete SemR was built for the sentence *John Smith discovered that three men had died.*

Conceptual Structure:

subconcept(discover:2, [detect-1, . . ., identify-5])
role(Theme, discover:2, ctx(die:5))
role(Agent, discover:2, Smith:1)
subconcept(Smith:1, [male-2])
alias(Smith:1, [John, Smith, John Smith])
role(cardinality restriction, Smith:1, sg)
subconcept(die:5, [die-1, die-2, . . ., die-11])
role(Theme, die:5, man:4)
subconcept(man:4, [man-1, . . ., world-8])
role(cardinality restriction, man:4, 3)

Contextual Structure:

context(t)
context(ctx(die:5))
top context(t)
context lifting relation(veridical, t, ctx(die:5))
context relation(t, ctx(die:5), crel(Theme, discover:2))
instantiable(Smith:1, t)
instantiable(discover:2, t)
instantiable(die:5, ctx(die:5))
instantiable(man:4, ctx(die:5))

Temporal Structure:

temporalRel(startsAfterEndingOf, Now, discover:2)
temporalRel(startsAfterEndingOf, Now, die:5)
Comments.

The conceptual structure introduces the concepts *discover*, *die* and *man* and specifies the WordNet synsets to which they belong. Smith is said to be male (this information is derived from knowledge that John is a name of a male).

The contextual structure contains two contexts: an upper level context **t**, which describes what the speaker communicates as a true statement about the world, and the internal context ctx(die:5), describing what John Smith has found – the fact that three men had died.

The verb *discover* has the presupposition that the subordinate predication is true. This component of the lexical meaning is described in the contextual structure with the relation called context lifting relation (veridical, t, ctx(die:5)). Thanks to this relation, an inference will be made that the predicate die:5 (instantiable(die:5, ctx(die:5)) is true.

The temporal structure reports that *discover* and *die* events took place prior to the moment of speech.

4.2 Tectogrammatical structure of PDT (Hajič 2002, Hajič et al. 2001)

The tectogrammatical level is the deepest level envisaged within the framework of Functional Generative Description, developed by the Prague linguistic school (Sgall, Hajičova and Panevová, 1986). This is the level of the semantic structure representation of a sentence, or, as the authors call it, the "linguistic meaning". In all, Functional Generative Description involves three levels of sentence representation. Beside the tectogrammatical level, there is an analytical level at which the surface syntactic structure is presented, and a morphological level at which the sentence is viewed as a sequence of lemmas supplied with morphological features. The Prague Dependency Treebank contains the structures of all three levels for every treebank sentence. We will discuss the deepest, tectogrammatical structure. Its major features are as follows.

- A tectogrammatical structure is a tree whose nodes are content words of a sentence. All grammatical words, including prepositions, conjunctions etc. are not represented, and the information conveyed by them is transferred to the attributes of content words.
- The nodes are linked by dependency relations, or functors (deep syntactic relations). There are five actant relations (Actor, Patient, Addressee, Origin, Effect), a large group of circumstantial relations and several technical relations marking the negation, coordination, apposition, foreign-language expressions etc. In all, 72 functors are used (Hajič 2002).
- The nodes are supplied with a set of attributes that convey varied information which enables one to synthesize the original sentence, or a sentence synonymous with it, from the tectogrammatical structure.
- The structure restores certain types of ellipsis and establishes certain types of coreference relations.
- The structure shows thematic/rhematic articulation, together with the so-called deep word order which reflects the position of a word in the old/new scale (the communicative dynamism).

Fig. 1 presents an example of a tectogrammatical structure.

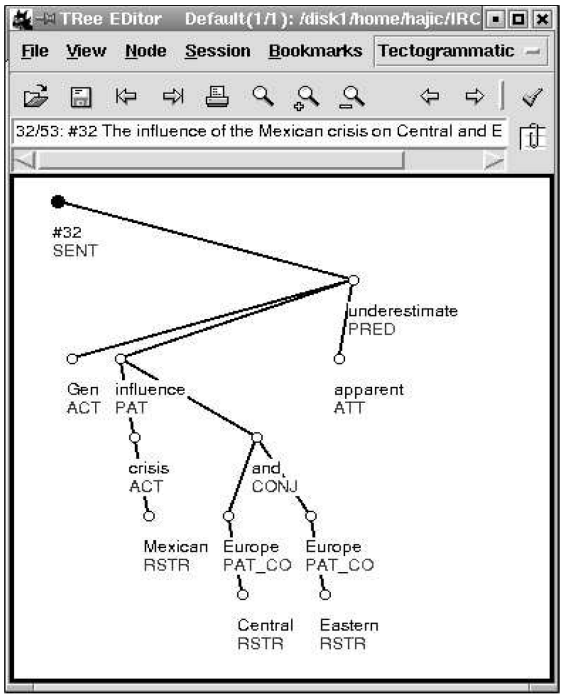


Fig. 1. The tectogrammatical structure for the sentence *The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated* (Hajic et al. 2001)

4.3 DRS (Bos et al. 2017)

SemRs which constitute the Groningen Meaning Bank (GMB) semantic corpus are built on the basis of Discourse Representation Theory and are called Discourse Representation Structures (DRS). In contrast to most SemRs appearing in semantic corpora, DRS embraces many sorts of information. A DRS is a multilayer record that contains the predicate-argument structure, thematic roles, verbal tense, coreference, quantifier scopes, rhetorical relations, and presuppositions. Importantly, DRS are built for whole texts rather than individual sentences and reflect not only intrasentential but intersentential factors (coreference, rhetorical relations). Events are represented in the neo-Davidsonian style (Parsons 1990): every event is marked by an individual variable referring to this event.

For instance, the record **administer(e18)** denotes that the event **administer** has received the name **e18**, which represents this event in various propositions, e.g. **Agent(e18,x15)** – 'the agent of the event e18 is the entity x15' (see Fig.2).

A DRS consists of two parts. The upper part lists the entities participating in the situation being depicted, while the lower part represents their properties and interrelations. Fig.2 provides an example of SemR in the form of DRS.

x2 e18 x15 x19 t12 t20
named(x2, cayman_islands, org) administer(e18) Theme(e18, x2) named(x15, jamaica, loc) Agent(e18, x15) timex(x19,+1863XXXX) after(e18, x19) now(t12) e18 ⊆ t20 t20 < t12

Fig. 2. A DRS for the sentence *The Cayman Islands were administered by Jamaica after 1863* (Bos et al. 2017: 9).

Omitting some details, this DRS can be read as follows: «the event 'administer' has the named entity Jamaica as Agent and Cayman Islands as Theme. The event had place in the past, starting from 1863».

Words appearing in DRS are supplied with three classes of markup: named entities, such as Person, Location, Organization (in all, 7 varieties), indicators of Animacy degree, such as Human, Organization, Animal, Machine etc. (9 varieties), and WordNet synsets.

Semantic elements are linked with thematic roles borrowed from VerbNet. For certain kinds of expressions, such as two-noun compounds, possessive and temporal constructions, an implicit relation is generated in the form of a preposition. For example, the sentence *The Apple spokesman announced Wednesday that its new products will be released this week* contained 4 implicit relations: *the Apple spokesman* = '(spokesman) of Apple', *announced Wednesday = '*(announced) on Wednesday', *its (Apple) products = '*(products) by Apple', *released this week* = '(released) in this week' (Bos et al. 2017: 16-17). We see that the generated relations are not very semantic. Apparently, the expressions *announced Wednesday* and *released this week* contain the same semantic relation but as the relations are generated from words requiring different prepositions, the representations turn out different. Besides, the generated prepositions are normally polysemic, which is not accounted for in any way.

DRS can be directly translated into formulas of first order predicate logic, which allows one to use the available inference software.

4.4 Compreno (Anisimovich et al. 2012, Stepanova et al. 2016)

The integral syntactico-semantic structure of Compreno is a non-tree graph of dependencies, supplied with grammatical and semantic information. The nodes of the graph correspond to the words of the sentence and are connected by syntactic relations and semantic roles (Stepanova et al. 2016). Some types of ellipsis are restored. Each word is associated with some element of the semantic hierarchy. For

instance, the Russian word *бутерброд* is viewed as SANDWICH_AS_FOOD, and the word *говорить* as TO_SAY_SPEAK_TELL_TALK.

Besides the reference to a semantic or lexical class, the word may have a semantic feature (semanteme) which is a component of the word's lexical meaning. The semanteme facilitates the choice of a translational equivalent in a different language.

Fig. 3 gives an example of a Compreno SemR. Note that the SemR shows the restored subject *ученик* 'student' (one of the students ⇒ one student of the students').

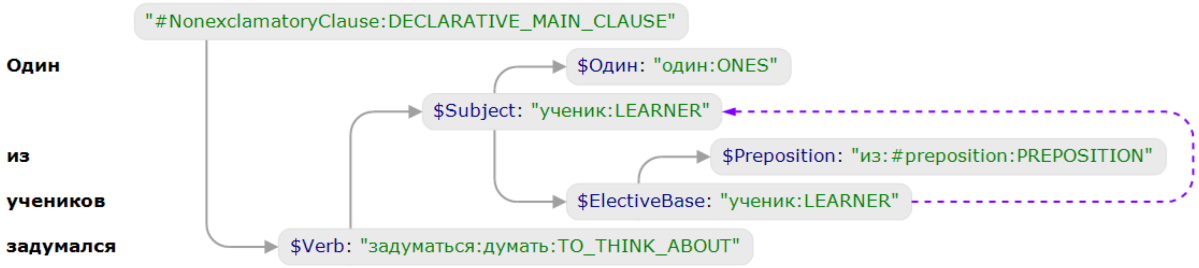


Fig. 3. Compreno SemR for the sentence
Один из учеников задумался 'One of the students started to ponder'

4.5 UNL (Uchida et al. 2005, Boguslavsky et al. 2005)

The Universal Networking Language (UNL) project was proposed by H. Uchida and developed by an international consortium for a number of years. The central idea is to develop a universal interlingual semantic representation (interlingua), which could serve as basis for the construction of a system of multilingual communication. Certain features of UNL are inherent in many SemR projects. In the amount of meaning, the semantic elements correspond to natural language lexemes. This means that on the one hand they match one lexical meaning of a word, rather than the whole vocable, while on the other hand they do not resort to decomposition of lexical meanings into smaller elements.

The semantic elements are linked to each other with dozens of binary relations resembling conventional semantic roles. The elements may be assigned additional features corresponding to modal and grammatical meanings. Semantic elements are organized into a hierarchy. A specific feature of UNL is a mechanism of dealing with lexical-semantic discrepancies between close though non-identical words of different languages. These discrepancies are described by the so-called constraints that are part of semantic elements. For example, semantic elements for the verb *marry* and its Russian counterpart *жениться* have specific constraints saying that the agent of the first one is a human (of either sex) and that of the second one is a male.

4.6 SemETAP (Boguslavsky et al. 2020, Boguslavsky 2021)

SemETAP is the semantic component of the functional model of language, ETAP-4, which implements the basic linguistic competences of humans – text understanding and text production. One of the key operations in text understanding is the extraction of all possible inferences. The model of understanding builds sequentially two semantic structures: the basic SemR and the enhanced SemR. The basic structure presents the direct meaning of the sentence, while the enhanced structure enriches it with a number of inferences which are construed on the basis of linguistic and extralinguistic knowledge accessible to the model. Both structures are built from the elements of a language-independent ontology, which thereby can be seen as a metalanguage of semantic description. All essential linguistic units are described in terms of ontological elements. The semantic description of many ontological concepts includes a decomposition of their meaning in the enhanced semantic structure into smaller components, which helps achieve a deeper text understanding, extract inferences and answer questions. For instance, the semantic structure of the concept 'envy' allows the model of understanding which processes a sentence like *Петя завидует Коле, что он нравится девушкам* 'Pete envies Nick because girls like him' and the question *Кто не нравится девушкам?* 'Whom don't girls like' provide the answer *Петя* 'Pete'.

Fig. 4 presents an example of a basic structure in SemETAP.

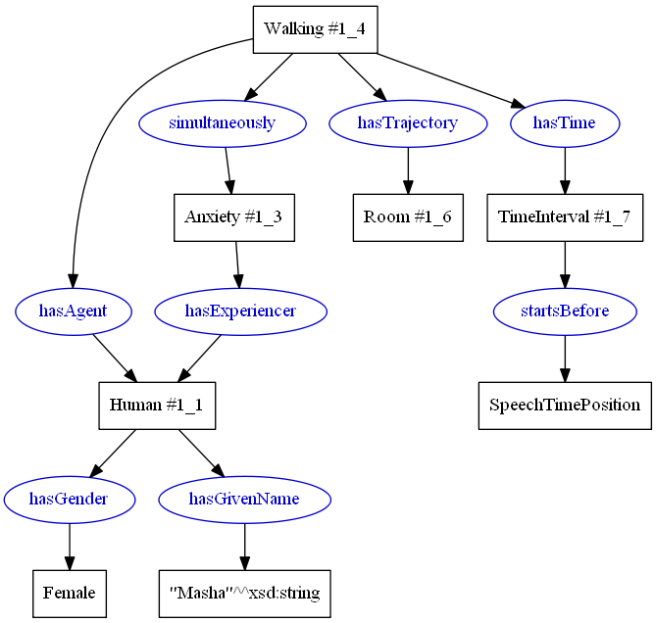


Fig. 4. The basic semantic structure for the sentence *Маша в волнении ходила по комнате* 'Masha walked about the room in anxiety'.

The structure can be read as follows: "A person of female gender, whose name is Masha, walks about the room and simultaneously experiences anxiety; this happened prior to the moment of speech".

5 Representation of Semantics in the Meaning ⇔ Text model

Of the whole range of theoretical semantic approaches, we chose the Meaning ⇔ Text model (MTT), developed by Igor A. Mel'čuk. It is a complete and very detailed theoretical model of language that describes how to represent a sentence at every level, including syntactic and semantic. This is why it is convenient to compare it with the semantic representations described above. Henceforth, we will call the semantic representation format of MTT — SemR-MTT for brevity, and semantic representations used in computational linguistics, SemR-CL.

We assume that the reader has sufficient knowledge of MTT and mention only two relevant publications: the first monograph describing the model (Mel'čuk 1974), and a comparatively recent three-volume work "Semantics: From meaning to text" (Mel'čuk 2012, 2013, 2015). We will present some characteristic features of MTT in comparison with SemRs-CL.

MTT is a strictly stratificational model. Every sentence receives formal representations at multiple levels: phonetic, morphologic, shallow-syntactic, deep-syntactic, semantic and conceptual. Most SemRs-CL are not viewed as parts of stratificational models and in many cases do not provide separate syntactic and semantic representations. Typical examples are AMR, which does not contain any dedicated representation of syntax, and Compreno, which has an integral syntactico-semantic structure. A notable exception is the tectogrammatical structure in PDT. It corresponds to the level of "linguistic meaning" (Hajič 2002) in the Functional Generative Description model and is contrasted with the morphological and analytical (surface-syntactic) structure. Among the levels of representation stipulated by MTT two may be seen as relevant counterparts of SemR-CL: deep-syntactic level (DSyntR-MTT) and semantic level (SemR-MTT).

The main purpose of (DSyntR-MTT) is to represent the syntactic structure of a sentence in a fashion that is abstracted away from any syntactic peculiarities of the sentence. Any syntactically conditioned grammemes, strongly governed prepositions and conjunctions, auxiliary verbs and other conventional words required by the grammar are eliminated from it. Syntactically synonymous constructions receive identical representation. The nodes of DSyntR-MTT are (almost) always content words, which may be supplemented with MTT lexical functions and a small number of special «fictitious» lexemes.

Most of the reviewed SemRs-CL are in many respects closer to the deep syntactic level of MTT than to its semantic level. First, the nodes of most SemRs-CL are content words of a natural language rather

than special semantic elements. Second, SemRs-CL closely reflect the syntactic skeleton of the source sentence. The best example here is AMR.

The types of relations between nodes of both DSyntR-MTT and SemR-CL are different from the relations between nodes of a purely syntactic (surface syntactic) structure, such as subject, direct object, indirect object, modifier, etc.

However, DSyntR-MTT uses universal deep-syntactic dependencies (numbered relations 1-7) to attach actants and **coordinative, quasi-coordinative, appenditive, attributive, descriptive-attributive relations. Semantic roles (agent, theme, experiencer, etc.) are not used. SemRs-CL use semantic roles in most cases. Some SemRs-CL (OntoNotes, AMR) reject the semantic roles to attach arguments to predicates and use indices instead: ARG0, ARG1, ARG2, etc.**

As noted above, DSyntR-MTT makes use of lexical functions in addition to content words. SemRs-CL completely ignore this large group of lexical means of expression, or, at best, omit support verbs corresponding to the lexical functions Oper/Func/Labor, as is the case in AMR. In the latter case, SemRs replace combinations like *make adjustments* with the single verb *adjust.* Meanwhile, dropping the support verb is not always possible. In particular, such omissions may produce anomalous results when a noun has a modifier. Compare *I had a feeling of relief – I felt relief* (the support verb *have* is dropped and the noun *feeling* is replaced with the corresponding verb) vs. *I had an unsettling feeling – *I felt unsettlingly* (the omission is not possible). This is one of the reasons why support verbs receive special treatment in MTT.

In certain SemRs-CL, including OntoAgent, SemETAP and UNL, the nodes may be special semantic elements instead of natural language words. Such structures bear more semblance to SemRs-MTT semantic structures and are farther away from the deep-syntactic level.

SemR-MTT is designed to represent the meaning of a source sentence irrespective of what words and syntactic structures were chosen to form it. The elements of SemR-MTT are not words but semantic elements — semantemes.

All content lexical units of the language receive definitions (semantic decompositions) consisting of semantemes. SemR-MTT can use different degrees of decomposition — from the minimal degree, when semantemes correspond to lexical senses of the corresponding words, to maximal degree, when decomposition reaches the level of semantic primitives. Minimal decomposition produces compact semantic structures, but in some cases it is not an adequate solution because certain semantic links cannot be revealed.

For example, the sentence *The green party won the majority at municipal elections* can be represented with a minimal decomposition of the semanteme ‘majority’. However, to adequately represent the sentence *The green party won a marginal majority at municipal elections* we need a deeper decomposition of ‘majority’. The modifier *marginal* does not refer to the whole meaning ‘majority’ (= ‘the number of votes «for» which is greater than the number of votes «against»’). Indeed, the number of votes secured by the greens can be very large. The modifier *marginal* concerns one of the inner semantic components of the word *majority*: the numbers of «for» votes is insignificantly greater than the number of «against» votes. It is impossible to show what the modifier’s contribution actually is without decomposing the meaning of ‘majority’ to the level, when its component ‘be greater than’ becomes explicit.

SemRs-CL almost never decompose lexical meanings. We know of only three projects in which the decomposition is performed in a substantial degree: SemETAP, OntoAgent, and Bridge.

Our comparison of DSyntR-MTT and SemR-MTT with SemRs-CL will not be complete without mentioning two “negative” facts. What features of semantic approaches used in computational linguistics are lacking in MTT?

First, MTT disregards all considerations of relevance of specific semantic components for computer applications. A semantic component is added to SemR-MTT of an expression if it is recognized as part of the meaning of that expression. Second, SemR-MTT never includes components that are produced by extralinguistic mechanisms: logical inference, world knowledge, common sense axioms, etc. All such phenomena are considered by MTT to belong to a level deeper than semantics, namely the level of conceptual representation. This distinction is accepted in computational linguistics too, but it is not drawn with the same rigor. There are approaches (DRT, OntoAgent and SemETAP) that make use of both linguistic and extralinguistic knowledge and aim to derive a wide range of logical inferences from texts.

6 Overview of Semantic Representation Features Considered

For the convenience of the readership, we summarize the main features of SemRs discussed above in Table 1. The cells for which we have no information are left blank. SemRs are classified by the following parameters:

- SemR nodes

[1] SemR nodes are NL words (+) or elements of a metalanguage (ontology, hierarchy of semantic classes, etc.) (-)
[2] If SemR nodes are NL words, they are linked to a higher level resource (WordNet, FrameNet, etc.)
[3] If SemR nodes are NL words, they are represented by a canonical variant (*may ⇒ possible, construction ⇒ construct*)

- SemR relations

[4] Predicate-argument relations are represented by semantic roles - Agent, Patient, etc. (+) and not by asemantic labels - ARG0, ARG1, etc. (-).

- Other information

[5] Anaphora/coreference is represented; ellipsis is restored.
[6] Information structure, Topic/Focus articulation is represented.
[7] Logical structure (quantifiers and their scope) is represented.
[8] Lexical meanings are decomposed (meaning postulates, lexical entailments, etc.)

- Levels of representation

[9] SemR is opposed to a syntactic structure.
[10] SemR is opposed to knowledge representation, which explicates different kinds of reasoning, e.g. logical entailment or common sense reasoning.

- Theory neutrality

[11] SemR is based on a specific linguistic theory.

- Existence of corpus

[12] There exists a corpus annotated by this type of SemR.
AMR: https://amr.isi.edu/
GMB: https://gmb.let.rug.nl/
PDC: https://ufal.mff.cuni.cz/pdt2.0/
UNL: http://www.unlweb.net/wiki/List_of_UNL_Corpora

	AMR	Bridge/PARC	GMB	PDC	Compreno	UNL	OntoAgent	Sem-ETAP	**MTT**
[1]	+	+	+	+	+	-	-	-	-
[2]	-	+	+	-	+	N/A	N/A	N/A	N/A
[3]	+	+		-	-	N/A	N/A	N/A	N/A
[4]	-	+	+	+	+	+	+	+	-
[5]	+		+	+	+	+	+	+	+
[6]	-			+	-	-	-	-	+
[7]	-		+	-	-	-		-	+
[8]	-	+	-	-	-	-	+	+	+
[9]	-	+	+	+	-	-	+	+	+
[10]	-	+	+	-	-	-	+	+	-
[11]	-	-	+	+	-	-	-	+	+
[12]	+	-	+	+	-	+	-	-	-

Table 1. Classification of SemRs

7 Conclusion

A wide range of SemRs are used in computational linguistics. Their main common feature is that they abstract away from grammatical and syntactic variety and try to represent different but synonymous constructions in similar fashion while contrasting syntactically similar constructions that differ in meaning. Another common feature is representation of predicate-argument relations, first of all for verbs, but

sometimes also for nouns and adjectives. Various sets of semantic roles are used to represent semantic relations between elements of SemR-CL. Many SemRs-CL also reflect other semantic phenomena: named entities (*Washington* as a human, city or US state), semantic derivates (*teacher* – 'someone who teaches'), coreference links, temporality, certain types of ellipsis.

Some SemRs-CL provide means to express presuppositions, rhetorical and discourse relations, scopes of quantifiers, implicit semantic relations. In rare cases of SemRs, partial decomposition of lexical meanings can be observed.

Often, the SemRs of sentences are produced on a large scale to create a semantic treebank. In this case the SemRs are mostly built by hand. A common way to facilitate this process is to ignore phenomena that are hard to tag for inexperienced annotators. Another method of building a semantic treebank is using an automatic semantic analyzer with subsequent manual editing of the output by experts (e.g. Bos et al. 2017).

Comparison of SemRs used in computational linguistics with the ways to express the meaning of language expressions in theoretical linguistics (illustrated by the Meaning ⇔ Text theory) showed that the two paths of semantic research have much in common but also reveal significant differences. As expected, the differences are mostly caused by different goal setting. Theoretical linguistics aims to embrace the full scope and complexity of linguistic phenomena and build explanatory models. Computational linguistics sees one of its primary goals in the creation of semantically annotated corpora (in order to be able to apply the full power of machine learning methods to work with SemRs).

If we agree that SemRs developed by theoretical linguists adequately represent the semantic phenomena encountered in the natural language, we should conclude that such SemRs can serve as a convenient source, from which the computational linguists can borrow methods of formal representation for certain new phenomena, as need be.

On the other hand, theoretical linguists can greatly benefit from the approaches to SemR construction developed in the field of CL. As noted before, most kinds of SemR-CL are embodied in text corpora annotated with such structures. Linguists have long learned to use annotated corpora in their research. The more diverse are the annotated data, the greater the value of the corpus for theoretical linguistics. A good example is the well-known FrameNet corpus, which provides vast information of ways in which semantic frames are expressed in English and some other languages. Another example is SynTagRus, which provides rich annotation for Russian texts, including morphological, syntactic, lexical-semantic, lexical-functional, anaphoric and some other types of annotation. Such corpora with deep annotation facilitate targeted search for data and statistical processing of the results.

We are sure that the wide range of semantic treebanks, already developed or being developed in CL, will be used fruitfully by theoretical and descriptive linguists to select, research and quantitatively assess the data.

Acknowledgments

This work was done with the financial support of a grant (No. 19-07-00842) from the Russian Foundation for Basic Research.

References

[1] Abend O., Rappoport A. UCCA: A semantic-based grammatical annotation scheme. // Proc. of IWCS. — 2013a. — p. 1–12.

[2] Abend O., Rappoport A. Universal Conceptual Cognitive Annotation (UCCA). // Proc. of ACL. — 2013b. — p. 228–238.

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

[4] Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo. — 2012. — p. 90–103.

[5] Baker C. F., Fillmore, Ch., Lowe J. The Berkeley FrameNet project. // COLING-ACL '98. Proceedings of the Conference. — Montreal, Canada. — 1998.

[6] Banarescu L., Bonial C., Shu Cai, Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn Ph., Palmer M., Schneider N. Abstract meaning representation for sembanking. // Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. — Sofia, Bulgaria. — 2013. — p. 178–186.
[7] Banarescu L., Bonial C., Shu Cai, Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn Ph., Palmer M., Schneider N. Abstract Meaning Representation (AMR) 1.2.6 Specification. — 2019. — https://github.com/amrisi/amr-guidelines/blob/master/amr.md#special-frames-for-roles
[8] Bobrow D., Cheslow B., Condoravdi C., Karttunen L., Holloway King T., Nairn R., de Paiva V., Price Ch., Zaenen A. PARC's Bridge and Question Answering System. // Proceedings of the GEAF 2007 Workshop Tracy Holloway King and Emily M. Bender (Editors). CSLI Studies in Computational Linguistics. Ann Copestake (Series Editor). — 2007.
[9] Boguslavsky I. Semantic Descriptions for a Text Understanding System. Computational Linguistics and Intellectual Technologies. // Proceedings of the International Conference "Dialog" [Komp'yuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]. — 2017. — p. 14–28.
[10] Boguslavsky I. Semantic analysis supported by inference in a functional model of language [Semanticheskij analiz s oporoj na umozakljuchenija v funktsional'noj modeli jazyka]. // Problems of linguistics [Voprosy jazykoznanija], № 1 — 2021. — pp. 29-56.
[11] Boguslavsky I., Cardeñosa J., Gallardo C., Iraola L. The UNL Initiative: An Overview. Lecture Notes in Computer Science. // Volume 3406, Springer Berlin / Heidelberg. — 2005. — p. 377 – 387.
[12] Boguslavsky I.M., Dikonov V.G., Frolova T.I., Iomdin L.L., Lazursky A.V., Rygaev I.P., Timoshenko S.P. Full-fledged Semantic Analysis as a Tool for Resolving Triangle-Copa Social Scenarios. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'yuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], issue. 19 (26), — 2020 — p. 106-118.
[13] Bos J., Abzianidze L. Thirty Musts for Meaning Banking. // Proceedings of the First International Workshop on Designing Meaning Representations. — Florence, Italy, August 1st. — 2019. — p. 15–27.
[14] Bos J., Basile V., Evang K., Venhuizen N.J., Bjerva J. The Groningen Meaning Bank. // Ide N., Pustejovsky J. (eds). Handbook of Linguistic Annotation. — Springer, Dordrecht. — 2017.
[15] Copestake A., Flickinger D., Pollard C., Sag I. Minimal recursion semantics: An introduction. Research on Language and Computation 3:281–332. — 2005.
[16] Hajič J. Tectogrammatical Representation: Towards a Minimal Transfer In Machine Translation. // Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks. Association for Computational Linguistics. — 2002. — p. 216–226.
[17] Hajič J., Hladká B., Pajas P. The Prague Dependency Treebank: Annotation Structure and Support. // IRCS Workshop on Linguistic Databases. — 2001. — p. 105--114.
[18] Homola P. Neo-Davidsonian Semantics in Lexicalized Grammars. // Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013). — 2013. — p.134-140.
[19] Hovy E., Marcus M., Palmer M., Ramshaw L., Weischedel R. OntoNotes: the 90% solution. // Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. — Stroudsburg, PA, USA. — 2006. — p. 57–60.
[20] Kamp H., Reyle U. From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT. — Kluwer, Dordrecht. — 1993.
[21] Kipper K., Korhonen A, Ryant N., Palmer M. Extending verbnet with novel verb classes. // Proceedings of LREC, volume 2006. — 2006. — p. 1
[22] Lenci A. Distributional semantics in linguistic and cognitive research. // Special issue of the Italian Journal of Linguistics, Rivista di Linguistica 20.1. — 2008. — p. 1-31.
[23] McShane M., Nirenburg S. A knowledge representation language for natural language processing, simulation and reasoning. // International Journal of Semantic Computing Vol. 6, No. 1, 3_23. — 2012.
[24] Mel'čuk I.A. An essay of the theory of linguistic "Meaning ⇔ Text"models. Semantics. Syntax. [Opyt teorii lingvističeskix modelej "Smysl ⇔ Tekst". Semantika, Sintaksis.]. — Science [Nauka], Moscow. — 1974.
[25] Mel'čuk I. Semantics*: From Meaning to Text.* Vol. 1. — Amsterdam/Philadelphia: John Benjamins. — 2012.
[26] Mel'čuk I. Semantics*: From Meaning to Text.* Vol. 2. — Amsterdam/Philadelphia: John Benjamins. — 2013.
[27] Mel'čuk I. Semantics*: From Meaning to Text.* Vol. 3. — Amsterdam/Philadelphia: John Benjamins. — 2015.
[28] Parsons T. Events in the semantics of English: A study in subatomic semantics. — Cambridge, MA: The MIT Press. — 1990.
[29] Pollard, C., Sag I. Information-based syntax and semantics. Volume 1. Fundamentals. // CLSI Lecture Notes 13. — 1987.
[30] de Salvo Braz R., Girju R., Punyakanok V., Roth D., Sammons, M. Knowledge Representation for Semantic Entailment and Question-Answering. // IJCAI'05: Workshop on Knowledge and Reasoning for Question Answering. — 2005.

[31] Sgall, P., Hajičová E., Panevová J. The Meaning of a Sentence in its Semantic and Pragmatic Aspects. — Prague - Amsterdam: Academia – North-Holland. — 1986.
[32] Stepanova M. E., Budnikov E. A., Chelombeeva A. N., Matavina P. V., Skorinkin D. A. Information Extraction Based on Deep Syntactic-Semantic Analysis. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow, June 1–4. — 2016.
[33] Uchida H., Zhu M., Della Senta T. Universal Networking Language. Edition 2. — UNDL Foundation. Geneva. — 2005.
[34] White A., Reisinger D., Sakaguchi K., Vieira T., Sheng Zhang, Rudinger R., Rawlins K., Van Durme B. Universal decompositional semantics on universal dependencies. // Proc. of EMNLP. — 2016. — pp. 1713–1723.

Semantic features and valency properties of the Russian verb *podoždat'* 'wait'

Igor M. Boguslavsky
A.A.Kharkevich Institute for Information Transmission Problems, B.Karetnyj 15, Moscow, 103051, Moscow, Russia;
Universidad Politécnica de Madrid, 28040 Madrid, Spain
bogus@iitp.ru

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

Abstract

The paper presents a detailed account of the semantics of the Russian perfective verb подождать (≈ 'wait some time'), which belongs to the family of words focused around the verb ждать 'wait'. The verb, much like the whole family, has a set of unique and non-trivial semantic properties that have not been so far adequately represented either in traditional and computer dictionaries of the Russian language or in scientific descriptions. The main features of this verb include its peculiar morphological and semantic relationship with the dominant word of the family, the verb ждать, as well as a ramified valence frame, characterized by rarely occurred means of implementing semantic valencies and unusual conditions of cooccurrence.

Keywords: Russian language, lexical semantics, lexicographic definition, family of words, valence structure, valence cooccurrence

DOI: 10.28995/2075-7182-2021-20-142-153

Семантические особенности и валентные свойства русского глагола *подождать*

И.М.Богуславский
Институт проблем передачи информации им А.А.Харкевича РАН Б.Каретный пер, 15,
Москва, 103051, Россия;
Universidad Politécnica de Madrid, 28040 Мадрид, Испания
bogus@iitp.ru

Л.Л.Иомдин
Институт проблем передачи информации им А.А.Харкевича РАН Б.Каретный пер, 15,
Москва, 103051, Россия
iomdin@gmail.com

Аннотация

Подробно рассматривается семантика русского глагола подождать, входящего в словообразовательное гнездо ждать. Этот глагол, как и все гнездо в целом, обладает набором неповторимых и нетривиальных семантических свойств, которые до сих пор не были адекватным образом представлены ни в традиционных и тем более компьютерных словарях русского языка, ни в научных описаниях. Главными особенностями этого глагола является его своеобразное морфологическое и семантическое соотношение с доминантой словообразовательного гнезда – глаголом ждать а также сложная валентная структура, отличающаяся редкими способами реализации семантических валентностей и необычными условиями совместной встречаемости.

Ключевые слова: Русский язык, лексическая семантика, словарное толкование, словообразовательное гнездо, валентная структура слова, совместная встречаемость валентностей

1 Вводные замечания

Словообразовательное гнездо глагола *ждать* обладает редкими свойствами, совокупность которых представляется уникальной и, по-видимому, не встречается в русском языке за пределами этого гнезда.

Главными из этих свойств выступают, во-первых, своеобразная словообразовательная парадигма этого гнезда, а, во-вторых, разветвленная валентная структура принадлежащих к гнезду *ждать* глаголов, к тому же характеризующаяся необычными средствами заполнения валентностей.

В настоящей статье мы коснемся сравнительно небольшой части особенностей упомянутой группы глаголов: сначала коротко рассмотрим их словообразовательную парадигму, а затем сосредоточимся на валентных свойствах одного из ее элементов, а именно, глагола *подождать*.

2 Ждательный падеж

Одной из конституирующих особенностей словообразовательного гнезда *ждать* является вариативность падежа, оформляющего прямое дополнение при глаголах этой группы – это может быть винительный падеж (*жду маму*$_{\text{вин}}$) или родительный падеж (*жду приема*$_{\text{род}}$ *у директора).* Эта вариативность, обнаруженная в свое время А.А.Зализняком у глагола *ждать* и нескольких других глаголов (*остерегаться*, *опасаться, стесняться*), была названа им «ждательным падежом» (см. Зализняк 1967: 49 и сл.). Показательно, что данной вариативностью обладают практически все глаголы нашего словообразовательного гнезда – не только переходные *ожидать, поджидать*, *подождать, обождать, выжидать, выждать* и *прождать,* но и глаголы с формантом *–ся*, которые обычно с большим трудом допускают дополнение в винительном падеже, - такие как *дожидаться, дождаться, заждаться.* Тот факт, что эта вариативность весьма сложна, поразному проявляется для разных слов рассматриваемой группы и для разных лексических значений этих слов[1] и не сводится ни к свободному варьированию, ни к дополнительной дистрибуции, никак не уменьшает ее принципиальную роль в идентификации данного словообразовательного гнезда.

Приведем некоторые примеры.

Ожидать

(1) *А Деточкин пешком потопал на станцию и стал ожидать электричку*$_{\text{вин}}$ (Э.Рязанов, Э.Брагинский);

(2) *Ожидая электрички*$_{\text{род}}$*, сделал несколько затяжек и, рачительно притушив и спрятав окурок, сразу почувствовал себя всемогущим* (П. Акимов).

Поджидать

(3) Меня сильно знобило, и я лег у огня, поджидая кипяток$_{\text{вин}}$ *(А. Ферсман);*

(4) *Однако дед не спешил обнародовать свои успехи, а поджидал конца*$_{\text{род}}$ *лета, когда доказательства успехов станут многочисленнее и крупнее* (Н. Дубов).

Подождать

(5) *Нет, — думаю, — пускай она едет, а я подожду следующий поезд*$_{\text{вин}}$ (С. Алексиевич);

(6) *Я вышел на ближайшей станции, подождал следующего поезда*$_{\text{род}}$ *и сел в него* (Ф. Искандер).

---

[1] Ср., например, глаголы *ждать* и *ожидать* в значении 'предстоять', которые допускают только винительный падеж прямого дополнения, начисто исключая родительный: *Что ждет нашу страну (*нашей страны), какие песни будет петь она завтра?* (А. Городницкий); *Третий вопрос: какое будущее ожидает нашу армию (*нашей армии)?* (М. Е. Салтыков-Щедрин). Стоит добавить, что такое дополнение (выражающее не объект, а экспериенцера) не может принимать родительного падежа даже при наличии отрицания: *Россию <*России> не ждет в ближайший год наплыв иностранных туристов.* Любопытным образом, у автоконверсивов глаголов *ждать* и *ожидать* в этом значении (по существу, тоже выражающих значение 'предстоять', хотя и с некоторой семантической добавкой антропоморфности) прямое дополнение спокойно выражается родительным падежом: *Россия (не) ждет в ближайший год наплыва иностранных туристов.*

Выжидать

(7) *Я спустился по дорожке в аллею, стал посредине аллеи и выжидал баронессу*$_{вин}$ *и барона* (Ф. М. Достоевский);

(8) *Иван Матвеич, слышавший вопрос, с любопытством выжидал ответа*$_{род}$ (Ф. М. Достоевский).

(9) *Несколько дней выжидал он — сигнала, знамения, трансцендента, беспарашютного падения вниз, росчерка молнии, подсказки заболтавшихся каморников* (А.Азольский).

Выждать

(10) *Если в районе ледокола будет туман, мы сможем опуститься на островах Карла у нашей базы и оттуда, выждав погоду*$_{вин}$ *и пополнив запасы бензина, вернуться на «Малыгин»* (М. С. Бабушкин);

(11) *Тогда больной отстал от толпы и, став позади сторожа, выждал удобного мгновения*$_{род}$ (В.М.Гаршин) ;

(12) *Пользуясь тихой погодой, в день — два переплывем на Котельный; там у нас есть запас корма для всех, и там можно выждать зиму*$_{вин}$*, если погода переменится; если нет, поплывем и дальше вдоль берега Котельного и к материку*; (В. А. Обручев).

(13) *Выждав длинную паузу*$_{вин}$*, священник снова поднял руку* (Б. Васильев).

Любопытно, что глагол *выждать* демонстрирует нетривиальную осцилляцию валентности объекта и валентности длительности: в (10) и (11) выражается валентность объекта (‘ожидается наступление (хорошей) погоды или удобного момента’), а в (12) выражается продолжительность ожидания (конца зимы). Аналогично, валентность длительности выражается и в (13) – словом *пауза* (тот факт, что паузу организует сам субъект ожидания, не влияет на статус валентности, хотя и обусловливает некомпозициальность выражения *выждать паузу*). В отличие от других глаголов нашего словообразовательного гнезда (в том числе и от видовой пары *выжидать*; ср. выше (9)), эти валентности не могут выражаться одновременно: ср. *Он ждал <выжидал> удобного момента целый год,* но не **Он выждал удобного момента целый год.*

Пережидать

Этот глагол обычно присоединяет в качестве прямого дополнения винительный падеж:

(14) *Я сидел за кухонным столом*, *пил чай и пережидал зиму*$_{вин}$, *как лодочники пережидают ледоход*$_{вин}$ *перед тем*, *как переплыть реку* (А. Курков)*;*

Изредка при *пережидать* встречается и родительный падеж дополнения, однако, как представляется авторам, в таких случаях этот глагол выступает не в своем прототипическом значении 'бездействовать до завершения нежелательной ситуации или процесса, после чего начать действовать’, а просто в значении ‘ожидать’, без указания на начало новой активности:

(15) *Ссадив хозяина, он отъезжал на полверсты от станции и там, завязив в снегу в стороне от дороги сани и лошадь, пережидал отхода*$_{род}$ *поезда* (Л. Н. Андреев);

(16) *Я им что, бессовестный генерал КГБ или бывший вороватый министр финансов, который за границей пережидает окончания*$_{род}$ *срока давности?* (С. Есин).

Переждать

Этот глагол сов. вида можно считать строгой видовой парой к пережидать, и он тоже принимает в качестве прямого дополнения в основном винительный падеж:
(17) *Немногие остаются здесь, чтобы переждать долгое холодное время*$_{вин}$*…* (Ю. Рытхэу);

(18) *— Через этот подарок, — сказал дядя Сандро, переждав тетю*$_{вин}$ *Катю, как некий стихийный шум, — он хотел показать, что выселение абхазцев отменяется* (Ф. Искандер).

Обратим внимание на метафоричность прямого дополнения в (18): *тетя Катя* предстает здесь как нежелательный процесс. Примеры на родительный падеж единичны и тоже представляют непрототипическое значение глагола:

(19) *Терпѣливо переждавъ окончанія*$_{род}$ *оживленной бесѣды, я спросилъ Османа о результатѣ его переговоровъ* (А. В. Елисеев, 1886).

Добавим, что родительный падеж при глаголах *пережидать* и *переждать* (как в примерах 15-16 и 19) предстает как несовременный или по крайней мере уходящий**.**

Прождать

(20) *Прождав автобуса*$_{род}$ *минут десять, Рашид с Митей решили добираться до метро пешком, благо дворами здесь идти было недалеко* (А. Житков);

(21) *Там прождал часа два автобус*$_{вин}$*, пока не узнал, что автобусов не будет «из-за распутицы»* (М. Харитонов).

Дожидаться

(22) *Еще когда она дожидалась машину*$_{вин}$ *в Анастасовке, к ней подошел какой-то парень и спросил по-русски*: *— Ты из Чегема?* (Ф. Искандер).

(23) *Я вышел на улицу и стал дожидаться дежурной машины*$_{род}$ (А. и Г. Вайнеры).

Дождаться

(24) *Поддержанный в тубдиспансере лекарствами и питанием, я настолько окреп, что, дождавшись жену*$_{вин}$ *домой, ринулся искать работу* (В.Астафьев);

(25) *Дождусь маму*$_{вин}$ *с хлебушком и буду дожидаться рассыльной*$_{род}$ (В. Чивилихин);

(26) *Сбежала я на четвертый день, дождавшись отбоя*$_{род}$ (Д. Рубина).

Заждаться

(27) *Наверняка преданные спартаковские болельщики заждались любимую команду*$_{вин}$ *дома и с нетерпением ждут субботнего матча* (сайт sportrbc.ru);

(28) *Друг! Тебя*$_{вин}$ *заждались дома, — / Да и мне мешаешь пить!..*(Саша Черный).[2]

(29) *Скворцы / погоды*$_{род}$ *вешней / заждались за лесами, — / и в жданьи безутешном / ребята / по скворешням / расположились сами* (В.В. Маяковский).

(30*) Из бесед с ними у меня сложилось впечатление, что они заждались приказа*$_{род}$ (А. Костюков).

«Ждательный падеж» является отличительной особенностью и для других глаголов рассматриваемого словообразовательного гнезда, даже периферийных и окказиональных – типа *поожидать, повыждать, наожидаться, изождаться* и, возможно, еще нескольких; в целях экономии места мы не станем приводить конкретных примеров на такие глаголы. Добавим, что в нашу задачу не входило рассматривать теоретические аспекты введения «ждательного» падежа в научный оборот, которые в последнее время детально обсуждаются в грамматических исследованиях (ср., например, Тестелец 2011); приведенные нами наблюдения служат лишь для большей глубины лексикографического портретирования описываемого здесь гнезда.

[2] Отметим попутно любопытный семантический эффект: в (27) и (28) наречие *дома* (заполняющее валентность места при глаголе *заждаться)* указывает на то, что у субъекта и объекта ожидания «дом» один и тот же. Аналогичную картину можно увидеть также в предложении (22) и в строчке известной песни «Раскинулось море широко» (*Напрасно старушка ждет сына домой*), где валентность места для глаголов *дождаться* и *ждать* несколько неожиданно выражается направительным наречием *домой*, которое также указывает на общность дома у субъекта и объекта ожидания.

3 *Подождать* в роли делимитатива для *ждать*

У глагола *ждать* – доминанты рассматриваемого словообразовательного гнезда – фактически нет обычной делимитативной пары с приставкой *по-*: глагол *пождать,* если и существует в современном языке, то лишь в виде реликтовых осколков. В разных подкорпусах НКРЯ можно обнаружить пару устаревших или стилизованных примеров типа
(31) *Ну если б я был бедняк? Разве два месяца пождать ничего не значит?* (М. Ю. Лермонтов);

(32) *Удивился гость, покачал головой / И пошел на Садовую улицу / Ждать трамвая номер тринадцатый. / Ждет он час, ждет другой, — не идет трамвай. / А прохожие только посмеиваются: / «Ишь нашелся какой избалованный. / Что ж, пожди, потерпи, коли время есть, / Долго ли до второго пришествия?» / И прождал бы он так до вечера, / Да терпение аглицкое лопнуло. / И побрел он пешком к Покрову, домой* (З.Гиппиус).

К этим примерам можно присовокупить и конструкцию фольклорного жанра *ждет-пождёт* (в которой, разумеется, глагол *пождёт* не является делимитативом и которая стоит в одном ряду с другими редупликативными фольклорными конструкциями с приставкой *по-* типа *тянет-потянет, стук-постук, скок-поскок*), ср.

(33) *И царица у окна / Села ждать его одна. / Ждет-пождет с утра до ночи, / Смотрит в поле, инда очи / Разболелись глядючи / С белой зори до ночи* (А.С.Пушкин).

Пожалуй, данными явлениями функционирование глагола *пождать* в современном русском языке и ограничивается. Фактическое отсутствие делимитатива для *ждать* являет собой уникальную ситуацию. На наш взгляд, она не имеет семантических оснований и может быть расценена как случайная флуктуация словообразовательной картины (или, если воспользоваться метафорой Е.В.Падучевой, как лексический пробел).

Обратим внимание, что у других глаголов несовершенного вида, принадлежащих к словообразовательному гнезду «ждать», делимитативные пары с приставкой *по-* в языке присутствуют, хотя их частотность и невелика: *ожидать – поожидать*, *выжидать – повыжидать*, *дожидаться – подожидаться, пережидать - попережидать* (ср, например, *Попивая кофе, можно поожидать своего самолета и подумать о чём-нибудь*; *Можно будет повыжидать дальнейшего развития*; *«Постоял на обочине, позаводил пару раз, подожидался загорания лампочки», «так мне и не удалось попережидать схватки в вертикальном положении»* - корпус Aranea Russicum).

Показательно в этой связи и то, что в близкородственных славянских языках делимитативы на *по-* для эквивалентов глагола *ждать* не обнаруживают никакой нерегулярности; ср. укр. *Чекати – почекати,* бел. *чакаць - пачакаць*, пол. *czekać – poczekać,* чеш. *čekat – počkat*, болг. *чакам – почакам.*

В этих условиях роль делимитатива для *ждать,* по нашему убеждению, полностью принимает на себя глагол *подождать.* Идея о делимитативности этого глагола неоднократно высказывалась. Так, Е.В. Падучева, обсуждая семантику делимитативных глаголов, имплицитно относила к ним глагол *подождать*: “делимитативный показатель имеет предпосылкой потенциально неограниченную продолжительность деятельности и не присоединяется к деятельностям, имеющим предел, в том числе и внешний (исключения – *поискать, подождать*); ср. Падучева 1996:147.

В книге Анны А. Зализняк, А.Д. Шмелёва и И.Б. Левонтиной (2012:350) указывается, что "глагол п*одождать* отчасти выполняет функцию отсутствующего делимитатива **пождать*”.

По нашему мнению, *подождать* – в современном языке вполне полноценный, а не частичный, делимитатив для *ждать,* соотносящийся с последним так же, как *поспать* со *спать* или *поработать* с *работать.*

Как и другие делимитативные глаголы, *подождать* особенно активно используется в ситуациях, которые характеризуются небольшой длительностью (точнее, длительностью, которая расценивается говорящим как небольшая). Приведем несколько цифр, полученных в результате анализа материала основного подкорпуса НКРЯ, которые, как нам представляется, подтверждают этот тезис достаточно красноречиво.

Мы сравнили встречаемость глаголов *ждать* и *подождать* с существительными, обозначающими отрезок времени – нейтральными и уменьшительными (*год-годик, день-денёк, час-часок/часик*, *минута-минутка-минуточка, секунда-секундочка*), а также с количественным наречием *немного* и его уменьшительным вариантом *немножко*.

Так, в основном подкорпусе НКРЯ[3] , как со снятой, так и с неснятой омонимией, глагол *ждать* встречается примерно 93 тыс. раз, а глагол *подождать* – 14 тыс. раз. Слово *год* встречается 901 тыс. раз, а *годик* – 790 раз, слово *день* встречается 412 тыс. раз, а *денёк* – около 1700 раз; слово *час* – 174 тыс. раз, а слова *часик и часок*, вместе взятые, – 2000 раз, слово *минута* – 116 тыс. раз, а слова *минутка* и *минуточка* – 4300 раз; слово *секунда* – 22 тыс. раз, а слово *секундочка* – 270 раз. Наконец, слово *немного* представлено 60 тысячами вхождений, а слово *немножко* – 11 тысячами.

Очевидно, что центральный глагол *ждать* и неуменьшительные слова отличаются гораздо большей частотностью, чем производный глагол *подождать* и уменьшительные слова.

Сочетания же этих слов друг с другом (рассматривались случаи, когда глагол непосредственно предшествует существительному или наречию) обнаруживают очевидно непропорциональное общей частотности высокое распространение уменьшительных слов в сочетаниях с глаголом *подождать* и весьма низкую сочетаемость глагола *ждать* с уменьшительными словами и с существительными, обозначающими малую длительность типа *секунда, минута, немного* (см. табл.1).

Табл. 1. Сочетаемость глаголов *ждать* и *подождать* с показателями длительности

	ждать	*подождать*
год	76	33
годик	0	8
день	96	17
денёк	0	15
час	130	9
часик/часок	2	3
минута	99	99
минутка/минуточка	0	118
секунда	6	24
секундочка	0	8
немного	7	368
немножко	0	39

Еще одно свойство делимитативов, полностью применимое к *подождать,* состоит в том, что последние довольно неохотно, сравнительно с нейтральными глаголами, присоединяют отрицание. По статистике того же подкорпуса НКРЯ (в той же старой версии), нейтральные глаголы обладают значительно более высокой частотой встречаемости, чем их делимитативные пары; так, глагол *играть* встречается 80 тыс. раз, а глагол *поиграть* – 3 тыс. раз (в 26 раз реже); глагол *летать* встречается 11500 раз, а *полетать* – 300 раз (в 38 раз реже), глагол *ждать* – 93 тыс. раз, а *подождать* 14000 раз (в 6,6 раз реже). Между тем в присутствии отрицания разница в частотности нейтрального и делимитативного глаголов оказывается еще более резкой: сочетание *не* + *играть* встречается 3660 раз, в то время как сочетание *не* + *поиграть* всего 36 раз (в 100 раз меньше), *не* + *летать* встречается 586 раз, а *не полетать – 4 раза* (в 145 раз меньше), *не* + *ждать* – 5500 раз, а *не* + *подождать* – 100 раз (в 55 раз меньше).

[3] Все цифры касаются старой версии НКРЯ, дата обращения – 5.09.2020. Никакие предварительные фильтры и постобработка (например, чтобы исключить омонимию типа *день* как существительное или как императив от *деть, час* и *часы, часик* и *часики*) не использовались. Большие числа даются округленно. Во всех случаях учитывались любые формы соответствующих глаголов.

Добавим еще типичные контексты (например, диалоги и контрасты), в которых встречается нейтральный и делимитативный глагол, выступающий в качестве перфективной пары к нейтральному; ср. *Почитай! – Не хочу я читать; Полетай! – Я уже летал; Отец сказал ему: «Бараны тебя подождут, а машины ждать не станут* (Ю.Трифонов) и т.п. Поведение глагола *подождать* в таких контекстах не отличается от поведения стандартных делимитативов на *по-*.

Следует отметить, что морфологическое устройство глагола *подождать* представляется нам загадочным. С одной стороны, можно предположить, что он содержит приставку *под-* в варианте *подо-*, определяемом скоплением согласных в корне (как в *подоткнуть*, *подорвать*, *подобрать*). Однако трудно понять, какова семантика такой приставки: мы не видим здесь ни намека на низкую позицию (как в *подлезть* или *подобрать*), ни идеи приближения (как в *подъехать*), ни идеи неполного количества (как в *подправить, подработать* или *подстраховать*).

Невозможно считать также, что *подождать* является видовой парой к *поджидать*: семантика этих двух глаголов весьма различна. Зато делимитативный статус, свойственный приставке *по-,* в глаголе *подождать* ощущается в полной мере. В частности, к ней вполне приложима характеристика делимитатива, предложенная недавно Е.В. Падучевой и опубликованная в ее посмертно изданной статье (Падучева 2020): «Делимитативный способ действия ограничивает ситуацию в ее временной протяженности, выявляя “порцию” ситуации – обычно небольшую и неопределенной длины».

Нам представляется возможным произошедшее в процессе эволюции данной глагольной формы случайное приращение к приставке *по-* из *пождать* форманта *до-,* который, однако утратил всякую связь с соответствующей приставкой. Этот гипотетический процесс (который мы, к сожалению, не можем подтвердить этимологически) несколько напоминает картину в квазисинонимичных парах глаголов *подвинуть(ся) – пододвинуть(ся),* в которой у вторых глаголов также трудно усмотреть семантические основания для присутствия приставки *под-*[4].

В любом случае, загадка формы *подождать* еще ждет надежного решения.

Добавим к сказанному, что в составе рассматриваемого словообразовательного гнезда содержится еще один делимитативный глагол – *обождать*, практически полностью синонимичный *подождать* и отличающийся от него только разговорно-просторечным статусом (приобретенным, судя по всему, в последние десятилетия)[5]. Заметим, что *обождать* не имеет даже следа делимитативной приставки *по-* .

4 Валентная структура *подождать*

Обратимся теперь к валентной структуре глагола *подождать*. Разумеется, эта структура в целом наследует валентную структуру глагола *ждать* – стержня словообразовательного гнезда, равно как и других глаголов этого гнезда, однако обладает и заметными особенностями.

4.1 Толкование

ПОДОЖДАТЬ

‘А1 не делает А2 (*подожди звонить*) или нечто, связанное с А2 (*подождем с квартирой, с этими планами*), до того, как начнет иметь место ситуация А3 (*подождать прихода гостей*), или появится объект А3 (*подождать гостей*), или пройдет период А4 (*два часа*)’.

[4] Неожиданным образом, глагол *подвинуться* вообще обладает странными семантическими особенностями. *Подвинься* с насыщенной валентностью (*подвинься ко мне)* обозначает приближение. *Подвинься* без насыщенной валентности обозначает отдаление. Между тем валентность исходной точки у этого глагола совсем не выражается: **Подвинулся от него.*

[5] Ср. нейтральный речевой статус *обождать* в классических текстах середины XX века: *Онъ отвѣтилъ, что такъ и сдѣлаетъ, но еще обождетъ денька два, авось деньги вернутся* (В. В. Набоков. Отчаяние, 1936); *В изысканных выражениях извинившись перед первосвященником, он попросил его присесть на скамью в тени магнолии и обождать, пока он вызовет остальных лиц, нужных для последнего краткого совещания* (М. А. Булгаков. Мастер и Маргарита, 1929-1940).

4.2 Варианты заполнения валентностей

a. A1: валентность субъекта ожидания. Прототипически выражается существительным в именительном падеже со значением человека или (значительно реже) разумно действующего агента; ср.

(34) *Мельник подождал, пока тот коснулся маслянистой шевелящейся поверхности, и спустил курок* (Д. Глуховский);

(35) *Пусть Европа подождет, пока русский царь рыбу ловит», — бросил он через плечо* (О.Гриневский; апелляция к якобы имевшей место реплике Александра III).

b. A2: валентность содержания, отражающая отложенное действие, субъектом которого выступает обычно субъект ожидания. Она может выражаться несколькими способами:
- инфинитивом;
- предложной группой *с* + твор. падеж;
- конструкцией *не* + пов.

Рассмотрим их по порядку.

Инфинитивное заполнение валентности отложенного действия обнаруживается в следующих примерах.

(36) *Он, Петенька Скоробогатов*, *поэтому подождет подписывать два договора, которые ему давеча прислали* (А. и Б.Стругацкие);

(37) *Ты подожди уходить. У меня к тебе дело* (А. Геласимов).

(38) *Подожди причитать раньше времени, дедушка Лих, — обнял его большую голову Тим* (И.Краева);

(39) *Да тут ничего и навязывать нельзя, — но вот с высоты моих лет и опыта: я бы подождал терять голову — посмотрел бы, как и что* (Ю. Домбровский)

В большинстве случаев инфинитив, выражающий отложенное действие, выступает при императиве глагола *подождать* – грамматическом, как в (37) и (38), или смысловом, как в (39). При этом семантика выражения, содержащего заполненную таким образом валентность, часто сводится просто к призыву не совершать действие и не предполагает призыва отложить его: вряд ли в (37-38) говорящий предлагает собеседнику *причитать* позднее *или потерять голову* через некоторое время. Этим же объясняется возможность инфинитивного заполнения валентности глаголом, обозначающим ненамеренное действие или процесс:

(40) *Подожди нервничать <переживать, паниковать>*;

(41) *Стоп, Миша, подождем умирать, сейчас будет вода* (А.Шумилов);

(42) *— Сражение выиграно, и в пленении Мюрата нет ничего необыкновенного. Но лучше подождать радоваться* (Л.Н.Толстой).

Предложная группа «*с* + твор. падеж» как способ реализации валентности отложенного действия иллюстрируется следующими примерами:

(43) *Поэтому подожди с ответом, пока я не напишу тебе из Европы* (В.В. Набоков);

(44) *Ну, Алеша, мы еще подождем с поцелуями, потому что мы этого еще оба не умеем, а ждать нам еще очень долго, — заключила она вдруг* (Ф.М. Достоевский).

(45) *Помню, я однажды кость проглотил… — Подожди ты со своей костью! — перебил его Миша* (А.Рыбаков).

В (43)-(44) группа «с+S, твор.пад.» выражает отложенное действие (ответ, поцелуи) субъект которого совпадает с субъектом ожидания, а в (45) отложенное действие субъекта ожидания хотя и не вербализуется, но как-то связано с предметом, обозначенном этим S (можно предположить, что здесь это – рассказ о случае с проглоченной костью).

Неожиданным образом в ряде случаев S в такой группе может выражать действие, у которого с субъектом ожидания совпадает не субъект, а объект; ср.

(46) *А лавочник, в ответ на мою просьбу подождать с уплатою долга, протянул ко мне масленую, пухлую, как оладья, руку и сказал: — Поцелуй — подожду!* (М.Горький);

(47) *Энергетики согласились подождать с возвратом долгов, но на всякий случай оставили без ГВС 30 домов* (сайт regnum.ru);

(48) *Ты пару дней сможешь подождать с моим переездом?* (Л.Фандеева).

Очевидно, что в (46)-(47) платить по долгам должен не лавочник и не энергетики, а их контрагенты, а в (48) ждать будешь ты, а переезжать я. В то же время при реализации валентности отложенного ожидания *у подождать* посредством инфинитива последний может быть только субъектным.

Достаточно нестандартный способ заполнения валентности отложенного действия – императив с отрицанием – наблюдается во фразах типа

(49) *— Девочки… подождите… не бранитесь, — говорил он, перемежая каждое слово вздохами, происходившими от давнишней одышки* (А. И. Куприн);

(50) *Подожди, не уходи, мне еще нужно столько рассказать* (Д. Гранин).

В этом случае в императиве оказывается как сам глагол *подождать,* так и глагол, заполняющий его валентность; при этом последний обязан стоять в несовершенном виде.[6] Как и в случае инфинитивного заполнения валентности отложенного действия, здесь тоже может идти речь о призыве отказаться от планируемого действия, а не просто о его переносе на потом.

Строго говоря, этот способ заполнения данной валентности, предусматривающий согласование форм главного и зависимого глаголов, изредка наблюдается не только при императиве *подождать*; ср.

(51) *Подождем, не будем ничего менять?* (Е. Завершнева);

(52) *Подожду, не стану плакать. И не стану кликать смерть. Кто же я – земная мякоть? Или неземная твердь?* (В. Долина).

Здесь заполняющий рассматриваемую валентность глагол согласуется с *подождать* по лицу и времени, а в будущем времени обязан стоять в аналитической форме.

c. А3 – валентность, которую можно с некоторой степенью условности назвать валентностью ситуации, прекращающей ожидание, также выражается несколькими способами:
   - именная группа в родительном или винительном падежах (т.е., в ждательном падеже, о котором речь шла выше); ср. примеры (5) и (6), а также

(53) *Павел сказал, что надо подождать его жену* (Д. Самойлов);

(54) *Я хотел подождать возвращения Машки, чтобы принести ей свою расплющенную школьную булочку* (А. Лиханов).

В выражениях типа (5), (6) и (53), когда валентность выражается существительным со значением человека или транспортного средства А3, речь идет о прибытии А3 в место, где находится субъект ожидания. Во всех других случаях выражения этой валентности никаких конкретных (в частности, дейктических) требований на субъекта ожидания и А3 не накладывается. Интерпретация любой ситуации, в которой при *подождать* реализуется данная валентность (скажем, предположение о том, что будет делать субъект ожидания после того, как его бездействие отменено)

[6] Может показаться, что в конструкциях типа (49-50) мы имеем дело не с заполнением активной валентности глагола *подождать,* а с сочинением предикатов: нужное значение (49), например, получается тогда из совокупности смыслов *подожди* ‘не выполняй некоторое действие’ и *не уходи* ‘не выполняй действие «уход»’. Обратим, однако, внимание, что в таких конструкциях невозможно использовать сочинительный союз: фразы типа **подожди и не уходи* невозможны. Это обстоятельство убеждает нас в том, что присоединяемая бессоюзно клауза действительно заполняет валентность *подождать,* аналогично тому, как это происходит при заполнении валентности глагола *изловчиться* сочиненной клаузой (ср. анализ *изловчился и прыгнул* в Богуславский 1996:31 и сл). Разница состоит лишь в том, что в последнем случае заполняющая валентность клауза требует союза, а в случае с *подождать* исключает его.

выходит за рамки лексической семантики, поскольку требует обращения к внеязыковой действительности и может производиться, например, с помощью аксиоматики здравого смысла;

- предложная группа, вводимая предлогом *до* + род.пад. Существительное в такой группе прототипически обозначает временную точку или событие, как в примерах (55-56), но может быть и предметным, выражая событие метафорически, как в (57):

(55) *Подождите до первого числа, когда жалованье получу* (А. П. Чехов);

(56) *Привалов являлся как раз в то время, когда хозяину нужно было уходить из дому, и он каждый раз упрашивал гостя подождать до его возвращения, чтобы пообедать вместе* (Д. Н. Мамин-Сибиряк);

(57) *Разумнее подождать до ближайшего порта* (т.е. до прихода в порт);

- придаточное предложение, вводимое несколькими типами союзов, в частности,

(i) союзом *пока* или его синонимом *покамест* или *покуда;* ср.

(58) *В зале были свободные столики, но они подождали, пока освободятся места у стойки* (В. Аксенов);

(59) *Префект молча подождал, пока весь полк не свернул с дороги, огибая город* (Б. Васильев)[7];

(60) *Нефедов молча с ним чокнулся и подождал, покуда генерал пригубит первым* (Г. Владимов);

(61) *Михаил все это сейчас вспомнил и подождал, покамест Раечка не прошла мимо* (Ф. Абрамов);

(ii) союзом *когда*, ср.

(62) **—** *Подождем, когда стемнеет, а тогда полезем через забор* (Н. Носов);

(iii) союзом *чтобы,* ср.

(63) *Желающий получить пшеничный пирог должен подождать, чтобы смололи муку* (Г. В. Плеханов);

(iv) союзом *что*[8]*,* ср.

(64) **—** *Подождал, что под сердцем шевельнется нежность и окатит горячим, но горячим почему-то не окатило* (В. Шукшин);

(65) *Я постоял рядом, подождал, что он заговорит со мною…* (А.Рекемчук).

(v) Добавим, что в разговорной речи союз, вводящий придаточное, которое выражает А3, может вообще опускаться, ср.

(66) *Подожди, я найду ключи (= подожди, пока я найду ключи).*

- бессоюзное придаточное, обычно содержащее модальный глагол или частицу *ли*, ср.

(67) **—** *Товарищи, вы подождите, может быть, будет еще машина сегодня* (Л. Гинзбург);

(68) *К чести Аннушки надо сказать, что она была любознательна и решила ещё подождать, не будет ли каких новых чудес* (М. Булгаков).

- придаточное, вводимое союзным словом; ср.

(69) *Не правда ли: вы уже свободно двигаете рукой? Вот подождем, что завтра скажет доктор…* (Е.Замятин);

(70) *Мне захотелось подождать, чем кончится вся эта история* (Л. Кассиль).

---

[7] Различие в значении и употреблении *пока* и *пока не* не составляет специфики глагола *подождать*, и мы не будем здесь его рассматривать.

[8] Это весьма редкий способ выражения данной валентности. Среди всех примеров контактного расположения слов *подождать* и *что*, целиком просмотренных авторами в основном и газетном подкорпусах НКРЯ, нашлись только три фразы, в которых этот способ реализован.

d. A4 – валентность длительности ожидания.

Эта валентность выражается именной группой в винительном падеже, формируемой словом со значением временного интервала, как в (71), или синтаксического эквивалента такой группы (количественной или аппроксимативной конструкции), как в (72-73). Если при глаголе *подождать* присутствует отрицание, винительный падеж может меняться на родительный (74).

(71) *— Я извиняюсь, — сказал он, и лицо его потемнело, — вы не можете подождать минутку?* (М. Булгаков);

(72) *— Я могу подождать читателя еще сто лет, — так примерно сказал Кеплер, — если сам господь ждал зрителя шесть тысяч лет* (Ю.Олеша);

(73) *Подождём ещё с недельку — вреда от этого не будет, а оснований прибавится…* (Ю.Домбровский);

(74) *Неужели там и в самом деле кто-нибудь обидится тем, что я не хочу подождать двух недель?* (Ф. М. Достоевский).

4.3 Слабая валентность источника

В целом валентная структура *подождать* повторяет валентную структуру доминанты гнезда – глагола *ждать*. Однако между ними есть и расхождения. Одно из них заключается в том, что при *подождать* практически не выражается валентность источника ожидания, вводимого предлогом *от* и достаточно обычного для *ожидать,* как в

(75) *Она ждала от отца сочувствия и понимания, но ничего подобного в нём не находила* (Л. Улицкая).

Примеры заполнения такой валентности у *подождать* единичны:

(76) *Но если он еще и арт-директор* […] *ему все равно придется взвешивать качество выбранного спектакля и решать, что ему сегодня важнее: включить в афишу «плохой» спектакль, условно говоря, Стуруа или Някрошюса или подождать от них следующего* (НКРЯ, дискуссия в журнале «Театральная жизнь»).

4.4 Совместная выразимость валентностей подождать

В принципе все основные несубъектные валентности этого глагола (A2, A3 и A4) факультативны, и чаще всего при нем выражается только одна из них. Тем не менее совместная встречаемость этих валентностей вполне обычна: ср.

(77) *Подождем Вальгана* [A3] *с этим вопросом* [A2] (Г. Николаева);

(78) *Она же отвечала, чтобы он подождал умирать* [A2]*, пока* [A3] *она не докончит свое вышивание* (С.М.Голицын);

(79) *Мамон, вели подождать казнить* [A2] *до приезда немчина-лекаря* [A3]*; да смотри, чтобы на злодеях железа не спали!*.. (И. И. Лажечников).

(80) *И ещё сутки* [A4] *надо будет подождать, когда* [A3] *пропитаю с внутренней стороны* (А.Иванов).

Вместо заключения: нерешенные проблемы

Разумеется, предложенное здесь разделение валентностей этого семантически сложного глагола нельзя считать окончательным. Интерпретация ряда конструкций с участием *подождать* требует дополнительного исследования. В частности, нам не до конца ясно, как следует трактовать выражения типа

(81) *Все же они решили подождать отца до полудня* (А.И.Мусатов):

если полностью принять наше распределение валентностей, то получается, что сильная валентность A3 реализуется здесь дважды, чего быть не должно. Возможно, с одной стороны, следует

отличать валентность, прекращающей ожидание, от валентности временной точки; с другой стороны, возможно, что валентность временной точки стоит объединить с валентностью длительности (правда, при этом придется объяснить поведение конструкций типа *Подождал два часа до вечера*).

Еще одна тема, которой мы не уделили должного внимания, – это инвентарь диатез рассмотренного глагола. В частности, необходимо определить, как соотносятся друг с другом конструкции типа *Подождем с решением до завтра* и *Решение подождет до завтра.*

Наконец, необходимо было бы внимательнее рассмотреть семантические особенности *подождать* на фоне других глаголов данного словообразовательного гнезда, а также глаголов, входящих с *подождать* в один синонимический ряд (*погодить, повременить,* плюс несколько глаголов в императиве – *постой, не спеши, не торопись*; ср. *Письмецо в конверте погоди, не рви* – Б.Окуджава; *Ты постой обижаться* – А.Платонов; *Не торопись с результатами, подожди с оценками – О.Арестова*).

Все это – предмет будущего исследования.

Благодарности

Авторы выражают признательность за поддержку данной работы Российскому фонду фундаментальных исследований (грант № 19-07-00842). Авторы благодарны также анонимным рецензентам «Диалога», сделавшим ценные замечания, которые позволили устранить ряд неточностей и уточнить некоторые выводы.

Литература

[1] Богуславский И.М. Сфера действия лексических единиц. М., Школа «Языки русской культуры», 1996. 464 с.
[2] Зализняк А.А. Русское именное словоизменение. М., Наука, 1967. 369 с.
[3] Зализняк Анна А., Левонтина И.Б., Шмелев А.Д. Константы и переменные русской языковой картины мира. М., Языки славянских культур. 2012. 696 с.
[4] Падучева Е.В. Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. М., Школа «Языки русской культуры», 1996. 464 с.
[5] Падучева Е.В. Делимитативный способ действия. // Известия РАН. Серия литературы и языка. 2020. Т. 79, № 3, с. 5-12.
[6] Тестелец Я.Г. Падеж как фактор идентичности при эллипсисе в русском языке// Кибрик А.Е. (ред.), Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2011). М.: РГГУ. 2011. С. 656–667.

Building Dataset and Morpheme Segmentation Model for Russian Word Forms

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

Sapin A.S.
Lomonosov Moscow State University
Moscow, Russia
alesapin@gmail.com

Abstract

The paper describes a way to generate a dataset of Russian word forms, which is needed to build an appropriate neural model for morpheme segmentation of word forms. The developed generation procedure produces word forms segmented into morphs that are classified by morpheme types, based on existing dataset of segmented lemmas and additional dictionary data, as well as fine-grained classification of Russian inflectional paradigms, which makes it possible to correctly process word forms with alternating consonants and fluent vowels in endings. The built representative dataset (more than 1,6 million word forms) was used to develop a neural model for morpheme segmentation of word forms with classification of segmented morphs. The experiments have shown that in detecting morphs boundaries the model has comparable quality with the best segmentation models for lemmas (98% of F-measure), slightly outperforming them in word-level classification accuracy (with score 91%).

Keywords: morphological segmentation; morpheme analysis of Russian word forms; neural models for morphology; morpheme segmentation with classification

DOI: 10.28995/2075-7182-2021-20-154-161

Построение датасета и модели морфемной сегментации для словоформ русского языка

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

Сапин А.С.
МГУ имени М.В.Ломоносова
Москва, Россия
alesapin@gmail.com

Аннотация

В статье описан способ генерации датасета с русскими словоформами, который необходим для построения соответствующей нейронной модели морфемной сегментации словоформ. Разработанная процедура генерации формирует словоформы, сегментированные на морфы, которые классифицируются по типам морфем, опираясь на существующий датасет сегментированных лемм и дополнительные словарные данные, а также дробную классификацию русских флективных парадигм, что позволяет правильно обрабатывать словоформы с чередованием согласных и беглыми гласными в окончаниях. Построенный представительный датасет (более 1,6 млн. словоформ) был применен для разработки нейронной модели морфемной сегментации словоформ с классификацией сегментированных морфов. Эксперименты показали, что при обнаружении границ морфов модель имеет сопоставимое качество с наилучшими моделями сегментации для лемм (98% F-меры), немного превосходя их по аккуратности классификации на уровне слов (91%).

Ключевые слова: морфологическая сегментация; морфемный анализ словоформ русского языка; нейросетевые модели морфологии; морфемная сегментация с классификацией

1 Introduction

Morpheme segmentation is a kind of morphological analysis, which implies breaking words into constituent morphs, the surface forms of morphemes (roots and affixes), for example, *taste-less*, Rus. *без-вкус-н-ый*. Morphemes are the smallest meaningful units of texts, so information about morphemic

structure of words is helpful for various NLP problems, in particular, recognition of semantically related words: cognates with the same root, paronyms (words that have similar morphs but differ in meaning) and so on. In lexical semantics, morphemic structure of words may be exploited to overcome data sparseness inherent to natural languages. The work [2] shows that even simple subword information can improve distributional word vectors representations, therefore more accurate linguistic information about word structures is useful for deriving meaning of rare and out-of-vocabulary words.

The data sparseness problem is more complicated for languages with rich morphologies, such as Russian, which is a highly inflective language with many affixes (prefixes, suffixes, postfixes) of various types and meanings. Significantly varying word forms are present in Russian texts, among them unknown words are often encountered, and their lemmas are unknown. For morphology rich languages the task of morpheme segmentation of words is especially complicated task, as it requires not only splitting into morphs but also classification of resulted morphs by labeling their main types (Prefix, Root, Suffix, Ending), for example:

без:PREF/*вкус*:ROOT/*н*:SUFF/*ый*:END, *taste*:ROOT/*less*:SUFF.

The first works on automatic morpheme segmentation were pure statistical, either dictionary-based [8] or corpus-based [7]. For a long time, only unsupervised and semi-supervised machine learning techniques were applied for the task, because of absence of representative datasets with labeled segmented morphemes for training. The most known solutions were implemented in Morfessor system [7, 10], which performs only morpheme segmentation and exploits unsupervised machine learning methods to be trained on a large text collection, showing about 70-80% of F-measure for detected morpheme boundaries, for English, Finnish, and Turkish words.

Recently proposed work [9] presents a dictionary-based morpheme segmentation method supplemented by application of word vectors representations (word embeddings), but like the previous methods does not involve classification of segmented morphs and achieves no more than 85% F-measure for English words.

The problem of morpheme segmentation with classification of segmented morphs remained almost unexplored until recent works [4, 5, 11] undertaken for Russian, in which powerful supervised machine learning techniques were applied. The implemented methods consider the task of morpheme segmentation with classification as sequence labeling [12] and classify letters of words being segmented to main types of morphs. Relevant labeled data were exploited for training the segmentation models, and among them, the most volume dataset obtained from derivation dictionary [13], it contains about 96 thou. segmented words (lemmas, i.e. normalized forms of words). The trained high quality segmentation models rely on various approaches.

- Convolutional neural network model[1] (CNN) [11];
- Gradient boosted decision trees (GBDT) model[2] [4];
- Long short-term memory neural network model[3] (Bi-LSTM) [5].

Besides approaches, the implemented models differ in classification schemes: the CNN model was trained based on BMES labeling scheme with 22 classes of letters, accounting for beginning (B), middle (M), ending (E) positions of a letter in the corresponding affix (prefix, root, suffix, postfix), as well as single (S) letter variants of affixes, and also hyphen and linking letter in multi-root and hyphenated words. Unlike the CNN model, in the works [4, 5] the number of letter classes was reduced to 10, since the set of BMES labels is redundant even for recognizing successive affixes and roots.

Evaluation of these CNN, GBDT and Bi-LSTM models trained on the same Russian datasets has shown their comparable quality [4, 5]: up to 98-99% of F-measure for morpheme boundaries (depending on training datasets and model hyperparameters), and about 96-98% of classification accuracy for letters and 87-88% for whole words. For now, these models present state-of-the-art methods for the considered task, outperforming the previously developed ones, both for morpheme segmentation and for segmentation with classification. However, they were developed only for segmenting lemmas, not for words in various grammatical forms encountered in texts. Meanwhile, for significantly varying Russian word forms (for verbs, up to 20 forms exist, differing in several affixes) the models for segmenting lemmas cannot work with similar quality. Thus, an appropriate morpheme segmentation mod-

[1] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation
[2] https://github.com/alesapin/GBDTMorphParsing
[3] https://github.com/alesapin/RussianMorphParsing

el applicable for word forms is to be built, but there were no relevant datasets with segmented word forms.

In our work we have built a representative dataset of Russian word forms, which is suitable to train a supervised machine learning model for morpheme segmentation with classification. For this purpose, a generation procedure was developed, which produces word forms segmented into classified morphs, based on the available dataset with Russian lemmas from [13] and additional data from Russian morphological dictionaries. The latter includes, first and foremost, the fine-grained classification of Russian inflectional paradigms taken from the system [3], which makes it possible to correctly process many word forms with alternating consonants and fluent vowels in endings.

We have exploited the built dataset for training a neural model intended for morpheme segmentation of word forms, along with classification of segmented morphs. CNN architecture was chosen as a core of the model. Experiments have shown that the developed model achieves 98% of F-measure for detected morphs boundaries and also gives up to 91% of accuracy for classification of morphs in whole words, while the analogous CNN model trained on lemmas works poorly for word forms, giving only about 40% for word-level classification accuracy. Thereby, the quality of the developed model for word forms is comparable with the best supervised machine learning models for segmenting lemmas, even with slightly outperforming them in classification accuracy.

The paper first explains the generation procedure we have developed, as well as the generated dataset with segmented word forms and labeled morphs. Then the CNN model trained on this dataset is described, and the results of its experimental evaluation are reported and discussed. Finally, some conclusions are presented.

2 Generating Dataset with Segmented Word Forms

The developed procedure generates and segments word forms for given lemmas that are split into classified morphs and taken from the dataset[4] obtained from dictionary [13] (hereafter, Tikhonov's dataset), thus extending it. This dataset encompasses 96,046 words (lemmas) of main parts of speech (POS): nouns, adjectives, verbs, and adverbs. Segmented morphs of words are classified according to main morpheme types (Prefix, Root, Suffix, Ending, Postfix), hyphen (*чей-либо*) and also linking letter for multi-root words (e.g., *вод-**о**-наливной*); successive prefixes and suffixes (if any) are labeled, for example: the verb *полюбоваться* (*to admire*) is segmented and labeled as *по*:PREF/*люб*:ROOT/*ова*:SUFF/*ть*:SUFF/*ся*:POSTFIX.

The generation procedure depends on part of speech (POS) of input lemma and performs segmentation of its possible word forms, based on known segmentation of the lemma and grammatical information about Russian flexions and word formation suffixes. The significant problem was related to processing alternating consonants and fluent vowels in ending parts of word forms (in roots and end suffixes), mainly for nouns and verbs. Below we present examples.

Lemma *звери**нец*** − *звер:*ROOT/*ин:*SUFF/***ец***:SUFF
Word forms *звери**нца*** − *звер:*ROOT/*ин:*SUFF/***ц***:SUFF/*а*:END
*звери**нцу*** − *звер:*ROOT/*ин:*SUFF/***ц***:SUFF/*у*:END

Lemma *лечь* − *ле*:ROOT/*чь*:SUFF
Word forms *легла* − *лег*:ROOT/л:SUFF/a:END
но *ля**жет*** − *ля**ж***:ROOT/*ет*:END

For verbs, fluent vowels are encountered not only in endings but also in prefixes, e.g.:

Lemma *отмереть* − ***от***:PREF/***мер***:ROOT/*е*:SUFF/*ть*:SUFF
Word forms *отмер* − ***от***:PREF/***мер***:ROOT
но *отомрет* − ***ото***:PREF/***мр***:ROOT/*ет*:END

To overcome the problem and to correctly recognize morphs (prefixes, roots, suffixes) while segmenting, we have exploited not the known canonical inflection-class system for Russian [14], but the system of numerous inflectional classes from system CrossLexica [3]. CrossLexica's system includes 313 classes for nouns, 25 classes of adjectives (they are also encompass passive participles), and 289 classes for verbs. Approximately 35% of the classes describe words with alternating consonants or

[4] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation

fluent vowels. Specification of each inflexion class includes endings (more precise, pseudo-flexions) of all word forms, and also an example of a word belonging to this class. Here is an example of noun inflexion class.

/ 96*/ {"ец","ца","цу","ец","цем","це","цу","цу"} , /*ранец*/*

The above mentioned words *зверинец* and *полюбоваться* have classes 96 and 34, respectively.

For our purposes, we have manually labeled boundaries between affixes in all endings of the classes, for example:

/ 96*/{ "ец-","ц-а","ц-у","ец-","ц-ем","ц-е","ц-у","ц-у"}, /*ранец*/*

Moreover, we have supplemented the specifications of the verbs classes with labeled gerund suffixes (*а/я, в,вши/ши)*, e.g.: *разлегшись* − *раз*:PREF/*лег*:ROOT/*ши*:SUFF/*сь*:SUFF, since they were not originally listed in CrossLexica's inflectional classes. For participles, the verbs classes specify only endings for active forms, whereas endings for passive participles are described in classes for adjectives. Here is an example of verb class.

/ 34*/ {"-ова-ть-ся","-ова-л-ся","-ова-л-а-сь","-ова-л-о-сь","-ова-л-и-сь",*
"у-ю-сь","у-ешь-ся","у-ет-ся","у-ем-ся","у-ете-сь","у-ют-ся",
"у-й-ся","у-йте-сь","-ова-вш-ий-ся","у-ющ-ий-ся", "у-я-сь","ова-вши-сь" },
*/*жаловаться*/*

One can notice that labeled endings (pseudo-flexions) include word formation suffixes of verbs (*ова*/*ева*, *ыва*/*ива*, *вш*/*ш*, *уш*/*ющ*, *л*, etc.) and postfix (*ся*, *сь*). To enhance consistency in verbs forms, we have additionally replaced in the original dataset the last label SUFF in infinitives of verbs by label END (since there is no full agreement between linguists about classification of infinitive morphs *ть, ти, чь*), e.g., *от*:PREF/*мер*:ROOT/*е*:SUFF/*ть*:END.

While applying our generation procedure, all words (lemmas) from Tikhonov's dataset were considered. For each particular lemma, our procedure finds its inflectional class indicated in the CrossLexica's dictionary, generates all its word forms according to the endings of the found class and then segments each word form, based both on known labels of the lemma being processed and on data taken from the class specification. More precisely, the beginning part of the word form copies segmentation and labels of the lemma, while the rest part is segmented according to the ending from the class. The following pair of lemma and its word form illustrates the process:

Lemma *пожаловаться*: *по*:PREF/*жал*:ROOT/*ова*:SUFF/*ть*:END:/*ся*:POSTFIX

Word form *пожаловалась*: *по*:PREF/*жал*:ROOT/*ова*:SUFF/*л*:SUFF/*а*:END/*сь*:POSTFIX

If the given lemma to be segmented is absent in CrossLexica's dictionary, all necessary word forms are taken from Open Corpora dictionary[5] [1], and inflexion class is automatically restored by the following rule: the set of all endings for an assigned class should coincide with the endings set of OpenCorpora's word forms.

Lemmas absent both in Open Corpora and in CrossLexica's dictionary accounted for about 10% of Tikhonov's dataset. Nouns and adjectives were mostly also processed, but in semi-automatic manner: word forms were predicted by morphological processor CrossMorphy[6], their class was restored by the above-described rule, with the following manual validation and necessary correction. Some difficult cases of nouns and adjectives were discarded, as well as all verbs lemmas absent in both dictionaries (the most of them are very rare or even out of use, such as *каландрироваться, окулироваться*). Overall, only 1,950 lemmas of Tikhonov's dataset (less than 2 %) were omitted.

As a result, about 98% of lemmas from the source dataset were processed and a dataset with segmented and classified word forms was built, its total size is 1,613,047 elements: 28% nouns, 45% adjectives and participles, 27% verb forms, and 0.05% adverbs. The built dataset consists of groups, each group encompasses word forms for a particular processed lemma, hereafter we call such groups inflectional. Groups for nouns are relatively small (6 singular forms, 6 plural), and groups are larger for adjectives (24 elements) and for verbs (15-18 elements). A verb group includes all personal forms of present, future, past tenses, and imperative forms, as well as two forms of active participle and 1-3 forms of gerund. Below we present fragments of the group for verb *связать* (*to tie*):

[5] http://opencorpora.org
[6] https://github.com/alesapin/XMorphy

с:PREF/*вяз*:ROOT/*а*:SUFF/*ть*:END
с:PREF/*вяз*:ROOT/*а*:SUFF/*л*:SUFF
с:PREF/*вяз*:ROOT/*а*:SUFF/*л*:SUFF/*а*:END
с:PREF/*вяз*:ROOT/*а*:SUFF/*л*:SUFF/*о*:END
с:PREF/*вяз*:ROOT/*а*:SUFF/*л*:SUFF/*и*:END
с:PREF/*вяж*:ROOT/у:END
с:PREF/*вяж*:ROOT/*ешь*:END
... ...
с:PREF/*вяз*:ROOT/*а*:SUFF/*вши*:SUFF/*ий*:END
с:PREF/*вяз*:ROOT/*а*:SUFF/*в*:SUFF
с:PREF/*вяз*:ROOT/*а*:SUFF/*вши*:SUFF

While testing our generation procedure, we have manually verified some fragments of the resulting dataset, to make sure that it is correct. It has been observed quite many errors in labeling segmented words (lemmas) in the source Tikhonov's dataset, mainly in classification of root morphs. We have corrected more than 1,5 thou. errors, and the corrected version of the dataset with segmented lemmas is now freely available[7], as well as the created dataset[8] with word forms.

3 Neural Morpheme Segmentation Model for Word Forms

To build and evaluate a morpheme segmentation model based on the generated dataset with word forms, among the best approaches for morpheme segmentation, namely CNN, GBDT, and Bi-LSTM, we have chosen convolutional neural network (CNN), because CNN is much faster to train, without lose in quality. At the same time, we did not use the auxiliary correction procedure and ensembles of several models proposed for original CNN model [11], since in our work such techniques do not significantly improve quality of segmentation.

Our CNN model for segmenting word forms was implemented with Keras library [6] (based on Tensorflow). Any input word (word form) is represented as a vector: one-hot encoded letters concatenated with information about is a particular letter vowel or not, and also concatenated with POS tag of the word, which is taken from the morphological dictionaries. POS labels include nouns, adjectives, verbs (personal forms), participles (active forms), gerunds, and adverbs. Similar to works [5, 6] we apply simplified labeling scheme of letters, with 10 classes.

The resulted CNN model has three layers with 512 units in each one, dropout of 40%, and ReLU activation function. The last layer is fully connected and completed with a softmax activation function, which outputs a probability distribution over all possible letter classes. Preliminary experiments with various hyperparameters of the model have shown that additional layers do not significantly improve its quality (the model with three layers gives sufficient results, losing to four-layers network less than 1%). Among the gradient descent algorithms (Adam, RMSprop, SGD), the better results were shown by Adam with a fixed learning rate of 0.001.

For our experiments, the generated dataset was randomly divided in proportion 70:10:20 for training, validation, and testing, respectively. Two variants of random dividing the dataset and corresponding trained models were studied:

- Random mixing of all labeled word forms, with subsequent splitting them to training and testing subsets ─ Model with Simple Mixing;
- Random mixing of inflectional groups (each group consists of all word forms corresponding to the same lemma), with subsequent splitting to training and testing subsets (thus, splitting does not divide the groups) ─ Model with Group Mixing.

Besides these two implemented models, we have also trained CNN model (of the same architecture) only on lemmas taken from the generated dataset (more precise, from its training subset) − Model on Lemmas. Programming code of these implemented morpheme segmentation models is available at GitHub[9] (our training, testing and validation sets are fixed for reproducibility).

[7] https://github.com/cmc-msu-ai/NLPDatasets
[8] https://drive.google.com/file/d/1_-0zKmmr2MS8NhQee16dZcRWz7cAwkt2/view?usp=sharing
[9] https://github.com/alesapin/XMorphy/tree/master/scripts/rule

We have evaluated both the quality of segmentation and classification accuracy of our segmentation models, the results are given in Table 1 and Table 2, respectively. The last rows of the Tables correspond to Model only on Lemmas. All scores were computed twice: for all word forms and only for lemmas.

The quality of segmentation (cf. Table 1) was measured in precision (P) and recall (R) of morph boundaries and F-measure (computed as mean harmonic of the recall and precision, F1). One can see that Simple Mixing Model slightly outperforms its counterpart (Group Mixing) in all the scores for morphs boundaries, but both models for word forms are much better than Model on Lemmas in scores for word forms (99-98% compared with 86-90% of F1-measure). As for scores for lemmas, they are almost similar for Group Mixing Model and Model only on Lemmas.

Model:	**Word forms**			**Lemmas**		
Training Set	P	R	F1	P	R	F1
Simple Mixing	99.56	99.71	99.63	99.41	99.55	99.48
Group Mixing	98.22	99.05	98.63	98.16	98.95	98.55
Only Lemmas	86.56	90.67	88.57	98.06	98.53	98.30

Table 1: Evaluation of morpheme segmentation for word forms and lemmas (%)

Table 2 corresponds to classification accuracy of the segmented morphs, for letters and for whole words. The former is the ratio of correctly recognized classes of letters to the number of all letters, the latter estimates the ratio of completely correctly segmented words with true classes of all their letters. Simple Mixing Model again outperforms its counterpart in all the scores, slightly for letters and significantly for words (97.34% and 91.06%). We can explain this as follows: since for the Simple Mixing Model inflectional groups may be divided while splitting to training and testing subsets, the testing subset may contain some word forms of the groups, whose elements are present in the training subset, and this improves evaluation results.

Thus, the group mixing is the more proper way of training and evaluating models on word forms, and scores of our Group Mixing Model are more adequate. This is additionally confirmed by comparing Group Mixing Model and Model only on Lemmas: their quality with respect to lemmas are almost similar, both in morpheme boundaries (Table 1, 98.55% and 98.30% of F1-measure) and accuracy in letters (Table 2, 97.54% and 97.13%); and at the same time these scores are highly close to those of the best neural morpheme segmentation models [5, 11].

Model:	**Word Forms**		**Lemmas**	
Training Set	Letters	Words	Letters	Words
Simple Mixing	99.42	97.34	99.15	96.40
Group Mixing	97.66	91.06	97.54	91.03
Only Lemmas	81.67	41.02	97.13	89.32

Table 2: Classification accuracy for word forms and lemmas (%)

It is surprising that our Group Mixing Model shows better results in classification accuracy, both for all word forms and for lemmas, than state-of-the-art results for lemmas: 91% for word-level accuracy compared with 87-88% [4, 5, 11]. In our opinion, the main reason is related to the learned knowledge: trained patterns for word forms help to more correctly parse lemmas. Other factors may also have influenced: exploiting corrected Tikhonov's dataset with lemmas and accounting for various POS during training. Apparently, these factors also improved quality of our Model only on Lemmas, for word-level classification accuracy: 89.32 % instead of 87-88% of state-of-the-art results [5, 11].

The last rows of Tables 1, 2 show that the Model only on Lemmas significantly loses when applied to word forms: much worse F1-measure on morpheme boundaries (88.57%) and even worse classification accuracy (81.67% for letters and 41.02% for words). It means, that for highly inflective languages, segmentation of word forms should be performed by models trained on relevant datasets.

4 Conclusions and Future Work

We have built the representative and volume dataset with Russian word forms split into morphs classified by main morpheme types. The rule-based generation procedure developed for this purpose relies on the analogous dataset containing only segmented lemmas, as well as on several known and proven morphological resources.

In our work, the built dataset was intended specifically to implement a neural morpheme segmentation model for word forms, which is important for morphologically-rich and highly inflective Russian. Experimental evaluation of implemented CNN models for segmenting word forms have shown their comparable quality with the state-of-the-art models for lemmas, and the properly trained model for word forms (Group Mixing Model) even outperforms the state-of-the-art results obtained for lemmas, giving about 91% in word-level classification accuracy.

The built dataset, programming code of the generation procedure, and the implemented neural models are of free access, as well as the corrected version of Tikhonov's dataset with lemmas. We hope they can be useful for other NLP tasks and experiments with Russian texts.

In our opinion, further progress in automatic morpheme segmentation for languages such as Russian may be achieved only after accounting some phonological features of words in training datasets, including introduction of iota sign [j] and two variants of consonants (hard and soft). Another interesting task for research involves creating a combined machine learning model performing both traditional inflectional morphological analysis and morpheme segmentation of a given word form.

Acknowledgements

We thank the author of CrossLexica for the provided dictionary resources.

References

[1] Bocharov V., et al. (2011), Quality assurance tools in OpenCorpora project [Instrumenty kontrolya kachestva dannyh v proekte Otkrytyj Korpus], Computational Linguistics and Intelligent Technologies: Papers from the Annual Int. Conference "Dialogue 2011", Bekasovo [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011"], Moscow, pp.101–109.

[2] Bojanowski P., Grave, E., Joulin, A., Mikolov, T. (2017), Enriching word vectors with subword information, Transactions of the Association for Comp. Linguistics, 5, pp. 135–146.

[3] Bolshakov I.A. (2013), CrossLexica – Universum of links between Russian words [CrossLexica – universum svyazey mezshdu russkimi slovami], Business Informatics [Biznes Informatica], No 3 (25), pp.12–19.

[4] Bolshakova E., Sapin A. (2019), Comparing models of morpheme analysis for Russian words based on machine learning, Computational Linguistics and Intellectual Technologies: Proceedings of the Int. Conference "Dialogue 2019" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2019"], Moscow, pp. 104–113.

[5] Bolshakova E., Sapin A. (2019), Bi-LSTM Model for Morpheme Segmentation of Russian Words // Artificial Intelligence and Natural Language: Proceedings of the conference AINL 2019, CCIS, vol. 1119. Springer, Cham, pp. 151–160.

[6] Chollet F. (2015), Keras: Deep learning library for theano and tensorflow. Access mode: https://keras.io/

[7] Creutz M., Lagus K. (2007), Unsupervised models for morpheme segmentation and morphology learning // ACM Transactions on Speech and Language Processing, 4 (1), Article 3.

[8] Harris S. Zellig (1967), Morpheme boundaries within words: Report on a computer test // Transformations and Discourse Analysis Papers, 73, pp. 68–77.

[9] Sakakini T., Bhat S., Viswanath P. (2017), MORSE: Semantic-ally Drive-n MORpheme SEgment-er // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp. 552–561.

[10] Smit P., Virpioja S., Gronroos S., Kurimo M. (2014), Morfessor 2.0: Toolkit for statistical morphological segmentation // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the ACL, Gothenburg, pp. 21–24.

[11] Sorokin A., Kravtsova A. (2018) Deep Convolution Networks for Supervised Morpheme Segmentation of Russian Language // Proceedings of the Conference on Artificial Intelligence and Natural Language, AINL 2018, St-Petersburg, Springer, Cham, pp. 3–10.

[12] Sutskever I., Vinyals O., Le Q. V. (2014), Sequence to sequence learning with neural networks // Advances in neural information processing systems, pp. 3104–3112.

[13] Tikhonov A.N. (1990), Word Formation Dictionary of Russian language [Slovoobrazovatel'nyj slovar' russkogo yazyka], Moscow, Russkij yazyk Publ.
[14] Zaliznjak A.A. (1977), Grammatical dictionary of Russian: Inflection. [Grammaticheskij slovar' russkogo yazyka], Moscow, Russkij yazyk Publ.

On (non-)compatibility of genitive partitive and imperfective in Russian: a corpus study

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

Abstract

The paper provides the results of the study of the use of the genitive case with partitive semantics as the means of direct object marking within imperfective verbs in Russian. The genitive partitive is traditionally claimed to be compatible with perfective verbs and as an exception with imperfective verbs used as the substitution for perfective verbs in neutralization contexts. The analysis of the data from the Russian National Corpus and the Russian-language Internet shows that the use of the genitive partitive within imperfective verbs is neither rare nor marginal. The compatibility level of the genitive and imperfective aspectual correlates of prefixed perfective verbs is dependent on the imperfectivability level and frequency. The use of the genitive partitive is sensitive to the semantics of the imperfective, however, it means the coverage of a broader range of phenomena than it is traditionally assumed. Although the use of the genitive partitive is mostly restricted to neutralization contexts such as iterativity and historical present, a number of gradual achievement imperfective verbs with progressive semantics as well as verbs that refer to constant situations are compatible with the genitive partitive.

Ключевые слова: Russian language, verbal aspect, Aktionsarten, direct object, partitive genitive case, perfective verbs, imperfective verbs, corpus study

DOI: 10.28995/2075-7182-2021-20-162-178

К вопросу о (не)сочетаемости родительного партитивного и несовершенного вида в русском языке: корпусное исследование

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

Аннотация

В статье приводятся результаты исследования сочетаемости родительного падежа с партитивной семантикой как средства оформления прямого дополнения и несовершенного вида глагола в русском языке. В литературе распространена точка зрения, согласно которой употребление родительного партитивного возможно только при перфективных глаголах и как исключение при имперфективных глаголах в контекстах нейтрализации видового противопоставления. Анализ материала Национального корпуса русского языка показывает, что употребление родительного партитивного при имперфективных глаголах не представляет собой редкое или маргинальное явление. Уровень сочетаемости родительного партитивного и имперфективных видовых коррелятов приставочных глаголов определяется уровнем имперфективируемости и собственной частотностью глагольных лексем. Употребление родительного партитивного падежа чувствительно к частным значениям несовершенного вида, однако круг явлений оказывается шире, чем традиционно обсуждается в литературе. Несмотря на преимущественное употребление в контекстах нейтрализации, таких как итеративность и настоящее историческое, употребление родительного партитивного наблюдается также при градационных глаголах в актуально-длительном значении и глаголах, реферирующих к постоянным ситуациям.

Ключевые слова: русский язык, глагольный вид, способы действия, прямое дополнение, родительный партитивный падеж, перфективные глаголы, имперфективные глаголы, корпусное исследование

1 Вводные замечания

В литературе распространена точка зрения, согласно которой употребление родительного (далеех— род.) падежа с партитивным значением возможно при глаголах совершенного вида (далеех— СВ) и невозможно при глаголах несовершенного вида (далее — НСВ) [20], [12: 182–190], [13], [6: 249], [18: 39], ср. (1)–(2):

(1) *Дурасиков помолчал, прокашлялся,* ***выпил воды*** *из стакана и сказал: «Тогда ладно…»* [Дина Рубина. Медная шкатулка (2011-2015)]

(2) *Хальдор поднялся к себе и долго, шумно* ***пил воду****.* [Елена Хаецкая. Хальдор из светлого города (1997)] (****воды***)

В ряде работ [20: 2236], [12: 182] делается оговорка, согласно которой запрет на род. падеж в позиции прямого дополнения действует лишь в отношении случаев использования НСВ в актуально-длительном значении и не распространяется на случаи употреблении НСВ в итеративном значении или в настоящем историческом, то есть в контекстах нейтрализации видового противопоставления, где НСВ является функциональной заменой СВ. При этом в случае вхождения глагола СВ в «видовую тройку» в контекстах нейтрализации используется вторичный имперфектив [21: 50], см. (3).

(3) *После репетиций мы шли напротив в «Артистик»,* ***выпивали коньячку*** *и продолжали свои разговоры, беседы*. [Михаил Рощин, Татьяна Бутрова. Драматургия и проза жизни // «Октябрь», 2003]

Несмотря на то, что проблематика сочетаемости НСВ и род. падежа неоднократно затрагивалась в литературе, вопрос о степени распространенности таких сочетаний (с точки зрения как лексического разнообразия глаголов НСВ, так и соотношения род. и вин. падежей) до сих пор подробно не рассматривался. В работе последовательно анализируется возможность употребления род. партитивного при глаголах НСВ, являющихся видовыми коррелятами глаголов СВ, способных к генитивному управлению. Цель исследования — проверка на корпусных данных и уточнение представлений о наличии зависимости между грамматическим значением вида и допустимостью оформления прямого дополнения род. падежом.

2 Материал и методика исследования

Для анализа сочетаемости глаголов с род. падежом прямого дополнения и его соотношения с формой вин. падежа была реализована следующая исследовательская процедура.

По Малому академическому словарю (далее — МАС) [3] был получен список глагольных лексем СВ, для которых в рамках указанного словаря зафиксирована возможность управления род. падежом. К анализу привлекались более или менее объемные (более 10 лексем) морфологические группы глаголов СВ, для которых характерно употребление с род. партитивным (всего 531 лексема): глаголы с префиксами *по-* — 51 лексема, *на-* — 372 лексемы, *под-* — 48 лексем, *при-* — 27 лексем, *до-* — 19 лексем, *от-* — 14 лексем. Как представляется, рассмотрение групп глаголов, объединенных определенными морфемными показателями, позволяет, с одной стороны, получить разнообразный языковой материал для анализа, с другой — выявить возможные сходства и различия в употреблении глаголов, входящих в разные группы.

В рамках исследования принята восходящая к С. И. Карцевскому [10] точка зрения, согласно которой тождество лексического значения возможно лишь в суффиксальных парах, где имперфективный коррелят образован от перфективного глагола при помощи суффикса *-(ы/и)ва-* (реже *-а-*). Для каждой перфективной глагольной лексемы, с опорой на результаты исследования имперфективируемости русских префиксальных глаголов (см. [9], а также базу данных по имперфективируемости: http://www.rusimpdb.ru) определялся имперфективный видовой коррелят (при его наличии), напр., *добавить — добавлять*, *надергать — надергивать*. При анализе учитывались как представленные в МАС конвенциональные имперфективные корреляты, так и глаголы, отсутствующие в словаре, но зафиксированные по данным Национального

корпуса русского языка (далее — НКРЯ) [17] и русскоязычного сегмента сети Интернет (рунет), напр., *накуковать — накуковывать*. Кроме того, вслед за рядом авторов [14: 364], [4], [7], принято решение считать глаголы прерывисто-смягчительного способа действия (далее — СД) результатом имперфективации делимитативов, т. е. объединять в видовые пары такие глаголы, как *попить — попивать*.

По данным основного подкорпуса НКРЯ методом ручной выборки составлен перечень и подсчитано количество всех зафиксированных в корпусе употреблений глаголов СВ и соответствующих им глаголов НСВ с дополнением в форме род. и вин. падежей. Поскольку употребление род. партитивного в позиции прямого дополнения возможно только в тех случаях, когда объект выражен кумулятивной именной группой (существительным с вещественной, реже отвлеченной, семантикой либо существительным в множественном числе, напр., *попить воды*, *прибавить скорости*, *нарубить дров*, см. [2: 83]), при подсчете случаев употребления вин. падежа также учитывались только примеры с кумулятивными именами. В базу данных не включались случаи употребления падежных форм в отрицательных контекстах и примеры, где падежная форма не определяется однозначно: употребления с несклоняемыми существительными (напр., *кофе*) в отсутствие определяющих слов либо иных средств, снимающих неоднозначность, а также примеры с одушевленными объектами, где наблюдается формальное совпадение род. и вин. падежей (учитывались сочетания типа *пострелять дичи*, но не *пострелять уток*).

При отсутствии примеров в НКРЯ, осуществлялся дополнительный поиск примеров в рунете.

3 Анализ результатов

В настоящей работе рассматриваются следующие аспекты, связанные с употреблением род. партитивного при глаголах НСВ: способность к употреблению с род. партитивным для глаголов СВ и НСВ в сопоставлении с данными об имперфективируемости префиксальных перфективов (раздел 3.1), количественные данные о соотношении род. и вин. падежей при глаголах СВ и НСВ (3.2), семантика НСВ и способность к употреблению с род. партитивным (3.3).

3.1 Способность к употреблению с родительным партитивным и имперфективируемость

В Табл. 1 ниже приведены данные (в абсолютных числах) о соотношении общего количества глагольных лексем с отмеченной в МАС способностью к управлению род. падежом (531 лексема) и количества глаголов СВ и НСВ, для которых по данным НКРЯ и рунета зафиксированы примеры употребления с формой род. партитивного в позиции прямого дополнения. Употребление с род. партитивным считается зафиксированным в случае обнаружения в анализируемом источнике (НКРЯ либо рунете) по крайней мере одного примера соответствующего функционирования.

Оценивая общее по всем рассматриваемым приставочным группам количество глагольных лексем НСВ, демонстрирующих случаи управления род. падежом, можно сделать вывод что такое употребление для глаголов НСВ не является редким или маргинальным: по суммарным данным НКРЯ и рунета сочетание с род. падежом наблюдается для 252 глаголов НСВ, что составляет 53,85% от общего количества лексем СВ, реализующих способность к употреблению с род. падежом.

	Количество лексем СВ (по МАС)	**СВ с род. партитивным по (НКРЯ)**	**СВ с род. партитивным (НКРЯ+рунет)**	**НСВ с род. партитивным (НКРЯ)**	**НСВ с род. партитивным (НКРЯ+рунет)**
по-	51	46	51	2	10
на-	372	257	326	57	168
под-	48	27	36	15	30
при-	27	21	24	16	18
от-	19	11	17	8	15
до-	14	11	13	7	11
Всего	**531**	**373**	**468**	**105**	**252**

Таблица 1: Количество лексем, употребляемых в сочетании с род. партитивным (данные МАС, НКРЯ и рунета)

Несмотря на то, что имперфективные глаголы показывают далекий от нулевого уровень сочетаемости с род. партитивным, сниженные показатели относительно данных для перфективных глаголов требуют объяснения.

Для наглядности представим приведенные выше данные в виде диаграммы, демонстрирующей относительные показатели реализованной способности к употреблению с род. партитивным глагольных лексем СВ, для которых такая возможность зафиксирована в МАС, и их коррелятов НСВ.

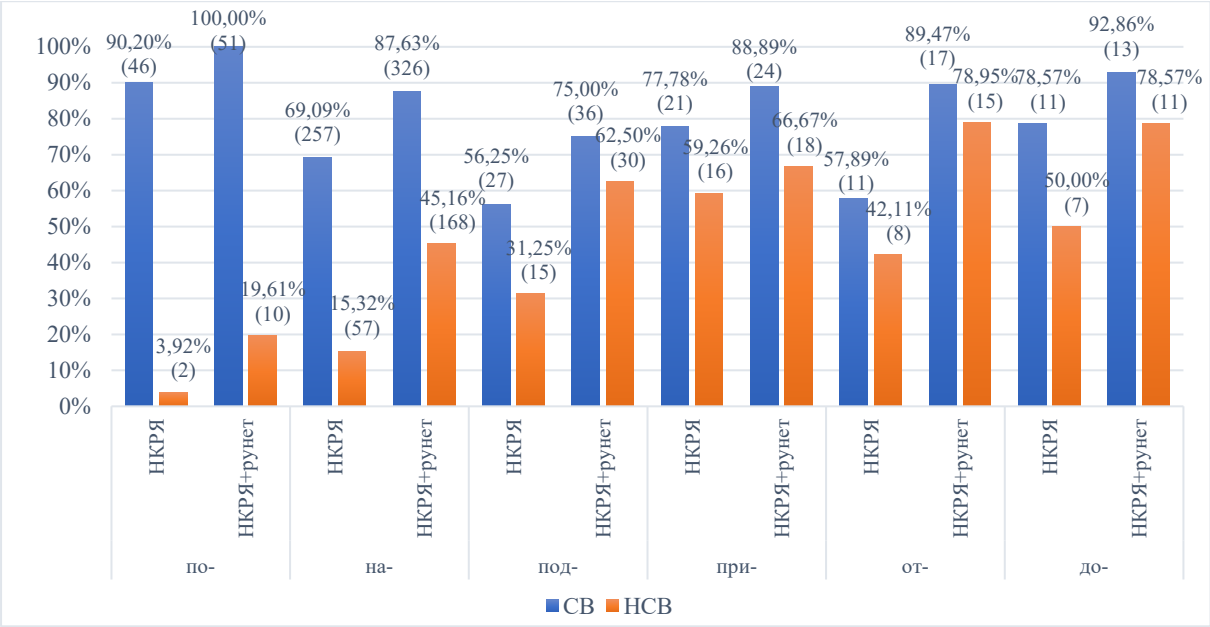


Рисунок 1: Употребление *по-*, *на-*, *под-*, *при-*, *от-*, *до*-глаголов СВ и НСВ с род. партитивным (относительные данные)

Рис. 1 показывает, что способность к употреблению с род. партитивным неодинакова у различных с точки зрения морфемного состава глаголов НСВ.

По-глаголы. Наибольшая разница между количеством глаголов СВ, для которых зафиксированы примеры употребления с род. партитивным, и аналогичными данными для глаголов НСВ, наблюдается в группе *по*-глаголов. Суммарные данные НКРЯ и рунета показывают, что для всех перфективных *по*-глаголов с отмеченной в МАС способностью к генитивному управлению обнаруживаются соответствующие примеры употребления. При этом использование род. падежа наблюдается только при 10 глаголах НСВ, что составляет около 19,6 % от общего количества глаголов СВ в выборке (см. Рис. 1); по данным НКРЯ — *посыпáть*, *попивать* (см. (4)–(5)), по данным рунета — *пожевывать*, *покапывать*, *покушивать*, *понюхивать*, *похлебывать*, *понавешивать*, *понаделывать*, *понастраивать* (см., напр., (6)).

(4) *Лазунка часто встает, шевелит угли костра да лопаткой* ***посыпает сырого песку****, чтоб хозяин не сжег сапоги…* [А. П. Чапыгин. Разин Степан (1927)]

(5) *Мама Маринина снова попивает портвешку, хотя меру все-таки знает, может, по Марининым молитвам*. [Майя Кучерская. Современный патерик: чтение для впавших в уныние (2004)]

(6) *В Балашихе если что , в 90-е таких домов не было. Это в конце нулевых и в 10-е годы стали* ***понастраивать муравейников***. (https://brstate.com/v/l2V8yIeCkKClq4s/my-popali.html)

Различие в количестве *по*-глаголов СВ и НСВ, реализующих способность к управлению род. партитивным, может объясняться общим низким уровнем имперфективируемости данной префиксальной группы, см. [8], [1]. Следует отметить, что большинство перфективных *по*-глаголов,

для которых наблюдается сочетаемость с род. партитивным, составляют лексемы, относящиеся к СД: делимитативному (25 лексем), аттенуативному (3 лексемы), кумулятивно-дистрибутивному (разновидности дистрибутивного СД) (14 лексем), поэтому при сопоставлении данных более корректно рассматривать показатели имперфективируемости СД с префиксом *по-*. В [1] показано, что уровни имперфективируемости аттенуативного и дистрибутивного СД значительно снижены по сравнению с данными по префиксальной группе в целом, в то время как делимитативный СД демонстрирует сопоставимый с общим по префиксальной группе (и даже несколько повышенный) уровень имперфективируемости (по суммарным данным МАС+рунет: дистрибутивный СД — 36,9%, аттенуативный СД — 48,39%, делимитативный СД — 73,03%, общий уровень имперфективируемости *по*-глаголов — 65,47% и 68,92% по максимальной и минимальной выборкам соответственно). Из 25 делимитативных лексем в пределах рассматриваемой выборки глаголов, способных к употреблению с родительным падежом, 18 (72%) имеют имперфективные корреляты, только для 6 из которых обнаруживаются примеры генитивного управления. Можно предположить, что невысокая доля имперфективных глаголов, демонстрирующих сочетаемость с род. падежом, на фоне данных о высоком уровне имперфективируемости делимитативного СД в целом, объясняется тем, что несмотря на наблюдаемую композициональность семантики глаголов прерывисто-смягчительного СД (префикс *по-* указывает на делимитативную семантику, а суффикс *-ыва-*/*-ива-* — на итеративность, см. [14: 364], [1: 162]), пары типа *попить – попивать* не в полной мере удовлетворяют критерию видовой парности, а именно, глагол НСВ не способен служить заменой СВ, например, в таком контексте нейтрализации видового противопоставления, как настоящее историческое.

На-глаголы. Наиболее многочисленную группу глаголов, для которых в МАС зафиксирована возможность управления род. падежом, составляют глагольные лексемы с префиксом *на-* — 372 лексемы. В литературе отмечается практически обязательное оформление род. партитивным дополнения при глаголах с префиксом *на-*, в частности, для входящего в состав данной префиксальной группы кумулятивного СД [5], [15], [16].

Глаголы с префиксом *на-* и семантикой накопления можно разделить на два типа: 1. «прототипическое» значение кумулятивного СД, в соответствии с определением М. А. Шелякина, «действия, направленные на достижение значительного количества одних и тех же результатов путем многократного осуществления действия исходного глагола» [19: 144], см. (7); 2. значение простого накопления объекта, соответствующее определению кумулятивного СД в [21: 114]: «Глаголы этого класса обозначают “накопление результата” действия», см. (8).

(7) ***Шишек накидал, веток сухих наломал —*** *и вот тебе топливо», — рассказывает Дмитрий.* [коллективный. Неделя. Герои // «Огонек», 2014]

(8) *Когда он приподнимал голову, чтобы* ***набрать воздуху****, в его глазах плясали два маленьких костра.* [Фазиль Искандер. Первое дело (1956)]

Рис. 1 показывает, что примеры употребления с род. падежом фиксируются в анализируемых источниках для 168 имперфективных лексем с префиксом *на-*, что составляет немногим меньше половины (45,16%) от общего количества перфективных лексем, для которых в МАС дается указание на возможность генитивного управления, при этом наблюдается существенное (практически трехкратное: в абсолютных числах — с 57 до 168) возрастание доли лексем НСВ, демонстрирующих примеры сочетания с род. падежом по данным рунета по сравнению с показателями НКРЯ. Как представляется, относительно низкая доля глаголов НСВ с префиксом *на-*, для которых фиксируются примеры употребления с род. падежом, может объясняться не ограничением на сочетаемость значений род. партитивного и НСВ, а особенностями функционирования, в частности, фактической неупотребительностью конкретных глаголов кумулятивного СД. Приведем несколько наблюдений.

Из Табл. 1 и Рис. 1 видно, что из 372 глагольных лексем СВ, для которых возможность употребления с род. падежом отмечена в МАС, 326 (87,63%) демонстрируют соответствующие случаи употребления в исследуемых источниках. При этом следует отметить, что глаголы СВ (46 лексем, 12,37%), для которых не зафиксированы примеры употребления с родительным партитивным, либо не демонстрируют случаев употребления ни с одной из падежных форм, либо

показывают низкую частотность и представлены единичными случаями употребления с винительным падежом. Корреляты НСВ таких глаголов также не употребляются в НКРЯ и рунете.

М. А. Шелякин отмечает, что имперфективные варианты глаголов кумулятивного СД встречаются изредка и только со значением многократности [19: 144]. Однако данные [1: 168] демонстрируют относительно высокий уровень имперфективируемости глаголов кумулятивного СД, по суммарным данным МАС и рунета достигающий 81,63%. В рамках выборки, включающей 326 глаголов, для которых в НКРЯ и рунете фиксируются случаи употребления с родительным партитивным, имперфективируемость устанавливается на уровне 91,72%, или 299 лексем по суммарным данным МАС+рунет (по МАС — 71,47%, 233 лексемы), при этом для 91 лексемы не обнаруживается однозначно определяемых примеров употребления как род., так вин. падежа. Из оставшихся 208 имперфективных лексем 168, или 80,77%, демонстрируют примеры употребления с род. партитивным. Отметим, что не зафиксированные в МАС, но образованные по регулярной модели неконвенциональные корреляты глаголов кумулятивного СД демонстрируют сочетаемость с род. падежом, см. (9)–(10).

(9) *А мы,помню школярами коxды были — салфетки сворачивали в несколько раз треуголкой и яки зверско-голодная моль,ножницами* ***дырок*** *в ней* ***накрамсывали*** *сикась-накось, а потома — разворачиваешь и дивные «снежинки» получались с узорами усякими.* (http://www.ozersk74.ru/news/usernews/219211)

(10) *Но да, уговаривать и* ***наобещивать гор*** *— это очень не полезное дело.* (https://www.b17.ru/forum/topic.php?id=114081&p=22)

Таким образом, практически реализуемая способность видовых коррелятов лексем кумулятивного СД с префиксом *на-* зависит как от зафиксированного уровня имперфективируемости, так и от фактической употребительности перфективных глаголов и их имперфективных коррелятов. В случае, если у глагола НСВ фиксируются случаи употребления в НКРЯ либо рунете, в большинстве случаев наблюдается также оформление прямого дополнения родительным падежом.

Под-, при-, до- и от-глаголы. Основную часть глаголов с префиксами *под-*, *при-*, *до-* и *от-* составляют выделяемые в классификации СД М. А. Шелякина [19] (и отсутствующие, например, в классификации, изложенной в [21]) комплетивно-партитивный СД с префиксами *под-* (46 из 48, 95,83%): *подкинуть*, *подкопить*, *при-* (19 из 27, 70,37%): *придать*, *приработать*, *до-* (9 из 14, 64,3%): *добавить*, *добрать*, отделительно-партитивный СД с префиксом *от-* (19 из 19, 100%): *отлить*, *отсыпать*.

Как видно из Табл. 1, в рамках приставочных групп с префиксами *под-*, *при-*, *до-* и *от-* доли глаголов НСВ, демонстрирующих примеры употребления с родительным падежом в НКРЯ и рунете, по отношению к данным для глаголов СВ, выше, чем в рассмотренных ранее группах с префиксами *по-* и *на-*. В первую очередь, это может объясняться более высоким уровнем имперфективируемости глаголов с префиксами *под-*, *при-*, *до-* и *от-*. Глаголы перечисленных приставочных групп характеризуются высоким уровнем имперфективируемости. Согласно МАС, в рамках рассматриваемых выборок парными являются 43 из 48 (89,6%) глаголов с префиксом *под-*, 26 из 27 (96,3%) — с префиксом *при-*, 14 из 14 (100%) — с префиксом *до-*, 16 из 19 (84,2%) — с префиксом *от-*. При привлечении данных НКРЯ и рунета имперфективируемость во всех группах достигает тотального уровня.

Приведенные данные позволяют сделать вывод, что доля НСВ, способных к употреблению с род. партитивным в рамках различных префиксальных групп, в основном определяется, во-первых, количеством глаголов СВ, демонстрирующих примеры использования с род. партитивным в НКРЯ и рунете, во-вторых, уровнем имперфективируемости, наблюдаемым для всей приставочной группы в целом и для входящих в нее СД в частности. Отклонение от данной тенденции, наблюдаемое для группы делимитативных *по-*глаголов можно связать с особенностями видового противопоставления, устанавливаемого в парах «делимитативный СД — прерывисто-смягчительный СД».

3.2 Родительный падеж при глаголах СВ и НСВ: количественные данные

Анализ количественного соотношения форм род. и вин. падежей при 105 глагольных лексемах НСВ, для которых в НКРЯ зафиксированы случаи употребления с род. партитивным, и сопоставление полученных данных с показателями для соответствующих глаголов СВ показывает, что выбор род. падежа для оформления прямого дополнения при глаголах СВ осуществляется чаще, чем при глаголах НСВ. В Приложениях А–Д приводятся данные о количестве обнаруженных в основном подкорпусе НКРЯ примеров употребления род. и вин. падежей для каждого из 105 имперфективных глаголов и их перфективных коррелятов. Тенденция к преобладанию винительного падежа при глаголах НСВ наблюдается как для выборки в целом, так и для приставочных групп в отдельности (по показателям средних значений, медианам, стандартному отклонению). Данные о соотношении частот по всем 105 глаголам приводятся на Рис. 2[1], в то время как Табл. 2. показывает наличие статистически значимой зависимости между значением вида и падежным оформлением прямого дополнения. Аналогичные данные для каждой префиксальной группы приводятся в Приложениях.

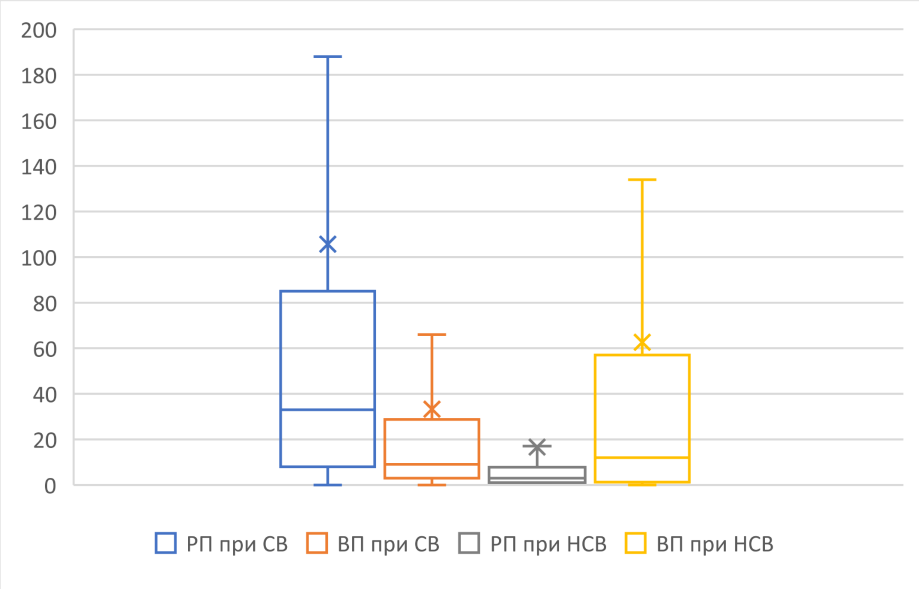


Рисунок 2: Частоты конструкций с родительным и винительным падежами при СВ и НСВ

	РП	**ВП**	**Всего**
СВ	10989	3458	14447
НСВ	1729	6515	8244
Всего	12718	9973	22691

Различие статистически значимо: χ^2= 6464,608, p<<0,01.

Таблица 2: Распределение конструкций с родительным и винительным падежами при глаголах СВ и НСВ

Различия в соотношении род. и вин. падежей при глаголах СВ и НСВ может объясняться упоминаемыми в литературе ограничениями на сочетаемость род. партитивного с определенными частными значениями НСВ, в частности, с актуально-длительным. Таким образом, контексты возможной вариативности составляют лишь определенную долю употреблений НСВ. В пользу такого объяснения свидетельствуют случаи отклонения от тенденции к преобладанию вин. п. при НСВ, наблюдаемые у имперфективных глаголов, для которых значение многократности является основным (*нарывать* <цветов>, *накупать* <подарков>).

[1] На диаграмме типа «ящик с усами» x обозначает среднее арифметическое, а средняя линия ящика — медиану для каждого представленного ряда данных.

Отклонение от тенденции к оформлению объекта вин. падежом наблюдается также для нескольких глаголов с абстрактной семантикой, референтирующих преимущественно к постоянным ситуациям: *подбавлять*, *прибавлять (2)*, *придавать*, см. (11).

(11) *Западные инвесторы очень консервативны, проверенные юрисдикции* ***придают*** *им* ***уверенности***. [Неоклеус Андреас. Кипру можно верить // «Эксперт», 2013]

В случаях типа (11) наблюдается употребление род. падежа за пределами контекстов нейтрализации видового противопоставления. Такие примеры не противоречат утверждению о несочетаемости род. партитивного и актуально-длительного значения НСВ: структура ситуации не предполагает развития действия, направленного на достижение некоторого предела (результата), выражаемого СВ, обозначаемая глаголом НСВ ситуация результативна в любой момент своего существования.

3.3 Актуально-длительное значение НСВ и употребление родительного партитивного

Как было отмечено выше, в литературе постулируется невозможность употребления род. партитивного при глаголах НСВ в актуально-длительном значении. Однако данные НКРЯ и рунета показывают, что имперфективные корреляты некоторых глаголов комплетивно-партитивного и отделительно-партитивного СД, а также кумулятивного СД с семантикой постепенного накопления результата («тип 2», см. выше) демонстрируют примеры сочетания с род. партитивным в том числе при употреблении в актуально-длительном значении, см. (12)–(14).

(12) *И пока он* ***набирает*** *из колонки* ***воды*** *в мутно-желтую большую банку, а потом протирает мокрой тряпкой памятник, я смотрю на небо, вижу его прохладную голубизну, вижу его слабое, подтаявшее по краям облачко…* [Ирина Муравьева. Ляля, Наташа, Тома (1991)]

(13) *— Необыкновенные? — воскликнул я с видом любопытства,* ***подливая*** *ему* ***чая***.
[М. Ю. Лермонтов. Герой нашего времени (1839-1841)]

(14) *— лениво отозвался Петров, щедро* ***отсыпая*** *на газетный лоскуток* ***злющего самосаду***.
[Анатолий Ткаченко. В заливе измены (1975) // «Огонек», 1961]

Видовые пары глаголов комплетивно-партитивного и отделительно-партитивного способов действия, а также кумулятивные глаголы с префиксом *на-*, демонстрируют сходства в семантике видового противопоставления. Глагол СВ в таких парах глаголов не указывает на достижение естественного предела. Ю. С. Маслов выделял в отдельную группу пары глаголов, обозначающих ситуации, для которых «нет возможности выделить “критическую точку”, знаменующую переход к новому состоянию, границу, отделяющую новое состояние от старого» [11: 86] (для обозначения видовых пар с указанным типом семантического соотношения используется термин «градативы» [12], или «градационные пары» [21]). Как представляется, обнаруживается сходство между градационными глаголами НСВ и упомянутыми выше глаголами, обозначающими постоянные ситуации (см. (11)): в обоих случаях рассматриваемые контексты НСВ не являются контекстами нейтрализации видового противопоставления, при этом обозначаемая длительная ситуация может рассматриваться как результативная в любой произвольный момент времени.

4 Выводы

Приведенные выше наблюдения показывают, что использование формы род. падежа с партитивной семантикой как средства оформления прямого дополнения при имперфективных глаголах оказывается сложнее, чем принято считать в литературе. Среди основных итогов исследования можно отметить следующие.

Употребление род. падежа при имперфективных глаголах не представляет собой редкое или маргинальное явление: соответствующее употребление наблюдается для видовых коррелятов

более чем половины глаголов СВ, демонстрирующих примеры сочетания с род. падежом в НКРЯ и рунете.

Доли имперфективных лексем, демонстрирующих примеры генитивного управления, неодинаковы в различных префиксальных группах и коррелируют с уровнем имперфективируемости, а также принципиальной с употребительностью (и частотностью) конкретных префиксальных лексем.

Данные о количественном соотношении род. и вин. падежей при глаголах СВ и НСВ свидетельствуют о преобладании вин. падежа при имперфективных глаголах, что, как представляется, является следствием ограниченной сферы употребления род. партитивного, сочетающегося только с определенными значениями из семантического спектра НСВ, в частности, со значениями, реализуемыми в контекстах нейтрализации видового противопоставления.

Употребление род. падежа при имперфективных глаголах чувствительно к семантическим особенностям НСВ и к семантике видового противопоставления, при этом рассматриваемая проблематика оказывается шире, чем, как правило, обсуждается в литературе. Так, например, обнаружено, что употребление род. партитивного не ограничивается контекстами нейтрализации видового противопоставления: данные НКРЯ и рунета позволяют сделать вывод о возможности употребления род. падежа при имперфективных глаголах, обозначающих постоянные ситуации. Также при ряде имперфективных глаголов кумулятивного, комплетивно-партитивного и отделительно-партитивного СД, входящих в градационные видовые пары, возможно употребление род. падежа в случаях, если НСВ имеет актуально-длительное значение, что постулируется в литературе как невозможное. Наконец, нестандартной семантикой видового противопоставления может объясняться сниженная способность глаголов прерывисто-смягчительного СД, рассматриваемых как результат имперфективации глаголов делимитативного СД, к употреблению с род. падежом.

Благодарности

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

Acknowledgements

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

References

[1] Chuikova Oksana Iu. (2020) On the secondary imperfectivation of *po*-perfectives in Russian [Ob osobennostyakh vtorichnoj imperfektivatsii glagolov s prefiksom *po*- v russkom yazyke], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”, vol. 19 (26), pp. 160–176.

[2] Chuikova Oksana Iu. (2020) Towards the issue on the interaction of verbal aspect and the case of direct object in Russian [K voprosu o vzaimodejstvii glagol'nogo vida i padezha pryamogo dopolneniya v russkom yazyke], Proceedings of the 7th International Aspectological Conference “Interrelation between Aspect and Adjacent Categories”, May 5–8, 2020 [Vzaimodejstvie aspekta so smezhnymi kategoriyami. Materialy VII Mezhdunarodnoj konferentsii Komissii po aspektologii Mezhdunarodnogo komiteta slavistov (Sankt-Peterburg, 5–8 maya 2020 goda).], Izd-vo RGPU im. A. I. Gertsena, St. Petersburg, pp. 81–93.

[3] Evgenieva Anastasija P. (ed.) (1981–1984), Dictionary of Russian Language in 4 volumes [Slovar’ russkogo jazyka v 4 tomah], 2nd ed. Academy of Science of USSR, Institute of Russian Language, Moscow, available at: http://feb-web.ru/feb/mas/mas-abc/14/ma239217.htm

[4] Fedotov Maksim L., Chuikova Oksana Ju. (2013), On the definition of limitative aspectual meaning and on the features of “delimitative” verbs in Russian [K opredeleniiu aspektual’nogo znacheniia delimitativa i voprosu ob osobennostiakh “delimitativnoi” derivatsii 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 statei i vospominanii k 100-letiiu Yu. S. Maslova], SPbU Publ., St. Petersburg, pp. 153-203.

[5] Filip Hana (2005), Measures and indefinites // Reference and Quantification: The Partee Effect, CSLI Press, Stanford, California, pp. 229–289.

[6] Glovinskaya Marina Ya. (2001) Polysemy and Synonymy in the Tense-Aspect System of the Russian Verb [Mnogoznachnost' i sinonimiya v vido-vremennoj sisteme russkogo glagola], Azbukovnik Publ., Moscow.
[7] Gorbova Elena V. (2019), To the Restriction on Imperfectivation: 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 Elena V., Chuikova Oksana Iu. (2020) Aktionsarten and the secondary imperfectivation (the case of *po-*, *pro-*, *u*-verbs) [Sposoby dejstviya russkogo glagola i vtorichnaya imperfektivatsiya (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 [Vzaimodejstvie aspekta so smezhnymi kategoriyami. Materialy VII Mezhdunarodnoj konferentsii Komissii po aspektologii Mezhdunarodnogo komiteta slavistov (Sankt-Peterburg, 5–8 maya 2020 goda).], Izd-vo RGPU im. A. I. Gertsena, St. Petersburg, pp. 136–148.
[9] Gorbova Elena V., Chuikova Oksana Iu., Sharygina Sofya S. (2021), Imperfectivability of Russian prefixal perfectives: regularity and peculiarities [Imperfektiviruemost' russkikh pristavochnykh perfektivov: regulyarnost' i specifika], Topics in the study of language [Voprosy Jazykoznanija], no. 4, pp. 91–130.
[10] Karcevskij Sergey J. (1962), Aspect [Vid], Issues in the Russian Aspect [Voprosy glagol'nogo vida], Izd-vo Inostrannoi Literatury, Moscow, pp. 218–230.
[11] Maslov Iury S. (2004), Selected works. Aspectology. General Linguistics [Izbrannye trudy: Aspektologiia. Obshchee iazykoznanie], Yazyki slavyanskikh kul'tur, Moscow.
[12] Paducheva Elena V. (1996), Semantic studies: Semantics of tense and aspect in Russian; Semantics of the narrative [Semanticheskie issledovaniya: Semantika vremeni i vida v russkom yazyke. Semantika narrative], Yazyki russkoj kul'tury Publ., Moscow.
[13] Paducheva Elena V. (1998), On non-compatibility of partitive and imperfective in Russian, Theoretical linguistics, vol. 24 (1), pp. 73–82.
[14] Pazel'skaya Anna G., Tatevosov Sergey G. (2008), Verbal Noun and the Structure of Russian verb [Otglagol'noe imya i struktura russkogo glagola], Verbal Derivation Research [*Issledovaniya po glagol'noj derivatsii*], Yazyki slavyanskikh kul'tur, Moscow, pp. 348–379.
[15] Pereltsvaig Asya (2006), Small nominlas, Natural Language and Linguistic Theory, vol. 24, pp. 433–500.
[16] Romanova Eugenia (2006), Constructing Perfectivity in Russian: Ph.D. Dissertation, University of Tromsø, Tromsø.
[17] Russian National Corpus [Nacional'nyi korpus russkogo jazyka] (2003–2019), available at: http://www.ruscorpora.ru
[18] Shatunovskij Ilya B. (2009), Problems of the Russian Aspect [*Problemy russkogo vida],* Yazyki slavyanskikh kul'tur Publ., Moscow.
[19] Shelyakin Mikhail (2008), Category of aspectuality of the Russian verb [Kategoriya aspektual'nosti russkogo glagola], URSS Publ., Moscow.
[20] Wierzbicka Anna (1967), On the semantics of the verbal aspect in Polish // To honor Roman Jakobson, Mouton, The Hague–Paris, pp. 2231–2249.
[21] Zalizniak Anna A., Shmelev Aleksey D. (2000), Introduction to the study of Russian aspect [Vvedenie v russkuyu aspektologiyu], Yazyki russkoj kul'tury Publ., Moscow.

Приложение А. Соотношение родительного и винительного падежей при СВ и НСВ с префиксом *на-*

	СВ				НСВ			
	РП по НКРЯ	**ВП по НКРЯ**	**Отношение шансов**	**сумма СВ**	**РП по НКРЯ**	**ВП по НКРЯ**	**Отношение шансов**	**сумма НСВ**
налить (1)	1822	470	3,88	2292	362	1075	0,34	1437
набрать (1)	1148	100	11,48	1248	160	424	0,38	584
накупить	514	28	18,36	542	**21**	9	2,33	30
насы́пать (2)	346	122	2,84	468	31	150	0,21	181
нарезать (1)	85	201	0,42	286	1	65	0,02	66
нарвать (1)	201	12	16,75	213	**4**	0	——	4
натаскать (1)	187	13	14,38	200	6	7	0,86	13
напустить	167	26	6,42	193	25	30	0,83	55
напечь	164	11	14,91	175	1	0	——	1
наделать (1)	162	9	18,00	171	1	1	1,00	2
набросать	106	43	2,47	149	5	15	0,33	20
нагнать (4)	85	54	1,57	139	15	98	0,15	113
наломать	111	9	12,33	120	2	2	1,00	4
нажить	26	84	0,31	110	15	95	0,16	110
нарезать (2)	75	34	2,21	109	1	13	0,08	14
нагнать (3)	64	41	1,56	105	4	31	0,13	35
наложить	75	10	7,50	85	31	143	0,22	174
навалить (2)	60	9	6,67	69	8	21	0,38	29
натащить	55	8	6,88	63	6	7	0,86	13
нацедить	56	3	18,67	59	4	9	0,44	13
наскрести (2)	43	7	6,14	50	2	6	0,33	8
навешать	36	11	3,27	47	3	19	0,16	22
натереть	15	29	0,52	44	1	14	0,07	15
набить (1)	39	4	9,75	43	5	32	0,16	37
надергать (1)	41	1	41,00	42	1	3	0,33	4
накопать (2)	36	5	7,20	41	1	1	1	2
намести (2)	37	3	12,33	40	4	4	1	8
насовать (1)	34	6	5,67	40	1	0	——	1
навалить (1)	16	23	0,70	39	4	16	0,25	20
нашить	35	4	8,75	39	2	2	1	4
нанести (1)	34	4	8,50	38	1	2	0,50	3
накачать	15	12	1,25	27	1	51	0,02	52
напилить	21	4	5,25	25	1	0	——	1

настлать (1)	22	3	7,33	25	4	7	0,57	11
навалить (3)	24	0	——	24	3	0	——	3
нагрести	21	3	7,00	24	3	14	0,21	17
нагулять	13	9	1,44	22	4	26	0,15	30
навесить	13	7	1,86	20	3	19	0,16	22
намазать	4	14	0,29	18	1	42	0,02	43
набрать (4)	7	6	1,17	13	2	6	0,33	8
настелить	12	0	——	12	4	7	0,57	11
набить (2)	7	4	1,75	11	1	2	0,50	3
надуть	7	4	1,75	11	1	3	0,33	4
накласть	10	1	10,00	11	31	143	0,22	174
намять	9	2	4,50	11	1	3	0,33	4
наметать2 (1)	6	4	1,50	10	1	1	1	2
настричь	8	2	4,00	10	1	1	1	2
надергать (2)	8	0	——	8	1	0	——	1
наменять (2)	8	0	——	8	2	0	——	2
наскрести (1)	6	0	——	6	1	1	1	2
нажечь	3	1	3,00	4	1	0	——	1
накатать (1)	2	1	2,00	3	2	1	2	3
напасти	3	0	——	3	1	0	——	1
настрелять	3	0	——	3	1	0	——	1
намыть (1)	1	1	1,00	2	2	1	2	3
накатить1	1	0	——	1	2	1	2	3
Среднее	**109,09**	**26,11**	**6,80**	**135,2**	**14,34**	**46,84**	**0,59**	**61,18**
Медиана	**30**	**6**		**39,5**	**2**	**6,5**		**9,5**
Стандартное отклонение	**288,93**	**69,71**		**349,02**	**52,21**	**154,30**		**206,06**
Сумма	6109	1462		7571	803	2623		3426

хи-квадрат			
	РП	**ВП**	**Total**
СВ	6109	1462	7571
НСВ	803	2623	3426
Total	6912	4085	10997
Различие статистически значимо: χ^2=3308,754, p<<0,01.			

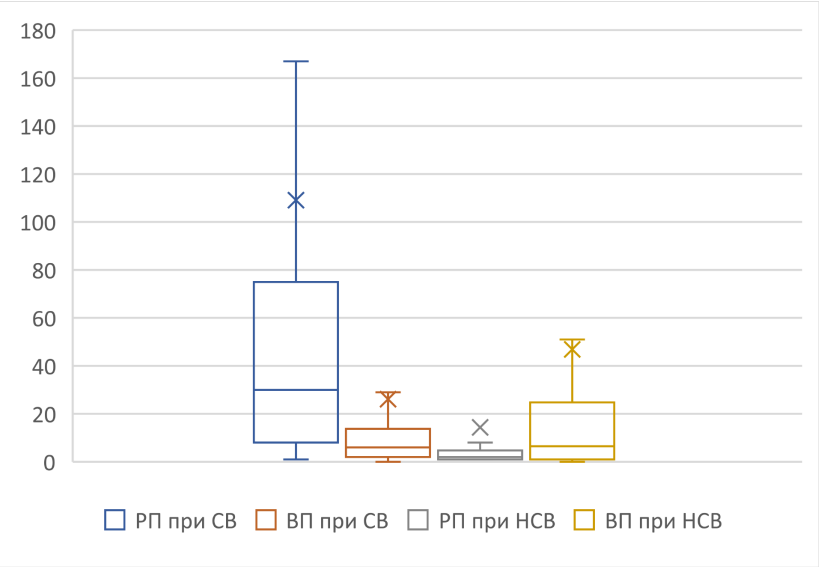


Рисунок А: Частоты конструкций с родительным и винительным падежами при СВ и НСВ с префиксом *на-*

Приложение Б. Соотношение родительного и винительного падежей при СВ и НСВ с префиксом *под-*

	СВ				НСВ			
	РП по НКРЯ	**ВП по НКРЯ**	**Отношение шансов**	**Сумма СВ**	**РП по НКРЯ**	**ВП по НКРЯ**	**Отношение шансов**	**Сумма НСВ**
подбросить	175	26	6,73	201	41	134	0,31	175
подсы́пать	127	43	2,95	170	30	59	0,51	89
подлить	120	16	7,50	136	154	167	0,92	321
поддать (1)	123	2	61,50	125	11	75	0,15	86
поднести	88	21	4,19	109	27	69	0,39	96
подкинуть	86	18	4,78	104	14	54	0,26	68
подбавить	73	2	36,50	75	**46**	15	3,07	61
подложить	59	10	5,90	69	37	165	0,22	202
подкопить	36	5	7,20	41	2	5	0,40	7
подмешать	18	21	0,86	39	9	58	0,16	67
подпустить (2)	26	8	3,25	34	2	1	2,00	3
подкупить	12	5	2,40	17	1	4	0,25	5
подпустить (3)	13	3	4,33	16	5	1	5,00	6
подкачать	1	3	0,33	4	1	9	0,11	10
подкосить	1	1	1,00	2	1	1	1,00	2
Среднее	**63,87**	**12,27**	**9,96**	**76,13**	**25,40**	**54,47**	**0,98**	**79,87**
Медиана	**59**	**8**		**69**	**11**	**54**		**67**
Стандартное отклонение	**54,55**	**11,87**		**62,38**	**38,98**	**59,20**		**91,17**
Сумма	958	184		1142	381	817		1198

хи-квадрат			
	РП	**ВП**	**Total**
СВ	958	184	1142
НСВ	381	817	1198
Total	1339	1001	2340
Различие статистически значимо: χ^2= 645,834, p<<0,01.			

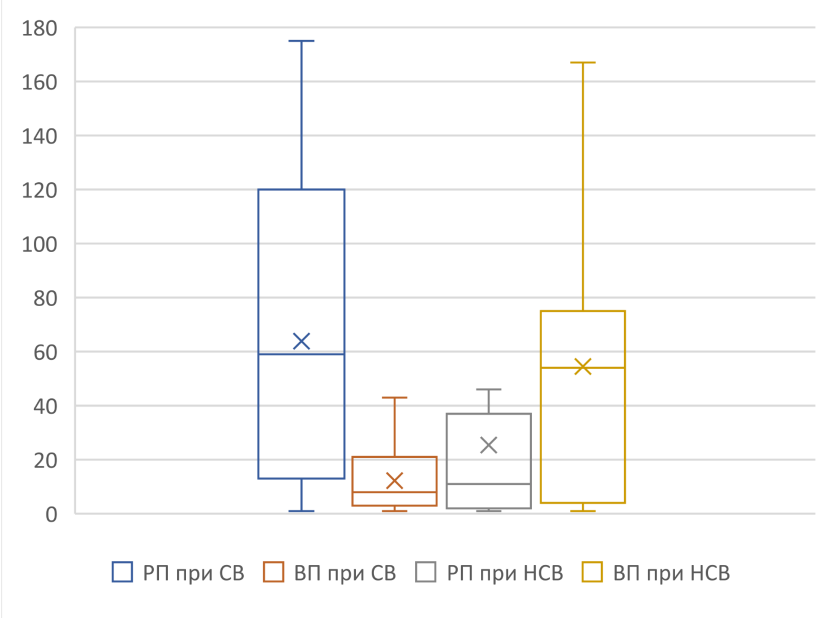


Рисунок Б: Частоты конструкций с родительным и винительным падежами при СВ и НСВ с префиксом *под-*

Приложение В. Соотношение родительного и винительного падежей при СВ и НСВ с префиксом *при-*

	СВ				НСВ			
	РП по НКРЯ	**ВП по НКРЯ**	**Отношение шансов**	**Сумма СВ**	**РП по НКРЯ**	**ВП по НКРЯ**	**Отношение шансов**	**Сумма НСВ**
прибавить (2)	835	166	5,03	1001	**255**	85	3,00	340
прислать	31	177	0,18	208	17	598	0,03	615
придать	139	66	2,11	205	**73**	61	1,20	134
прихватить	47	139	0,34	186	5	21	0,24	26
принять	28	150	0,19	178	1	33	0,03	34
прикупить	63	61	1,03	124	6	33	0,18	39
пригубить	28	65	0,43	93	1	14	0,07	15
прибавить (1)	44	28	1,57	72	22	40	0,55	62
приготовить	15	19	0,79	34	1	13	0,08	14
прихлебнуть	20	6	3,33	26	5	406	0,01	411
примешать	8	11	0,73	19	3	69	0,04	72

прирезать	9	6	1,5	15	1	2	0,50	3
прикопить	7	2	3,5	9	1	8	0,13	9
припустить	6	1	6	7	3	1	3	4
прикушать	6	0	—	6	1	0	—	1
приработать	1	0	—	1	1	1	1	2
Среднее	**80,44**	**56,06**	**1,91**	**136,5**	**24,75**	**86,56**	**0,67**	**111,31**
Медиана	**24**	**23,5**		**53**	**3**	**27**		**30**
Стандартное отклонение	**204,05**	**65,35**		**243,21**	**64,00**	**167,91**		**181,73**
Сумма	1287	897		2184	396	1385		1781

хи-квадрат			
	РП	**ВП**	**Total**
при- СВ	1287	897	2184
при- НСВ	396	1385	1781
Total	1683	2282	3965
Различие статистически значимо: χ^2= 539,187, p<<0,01.			

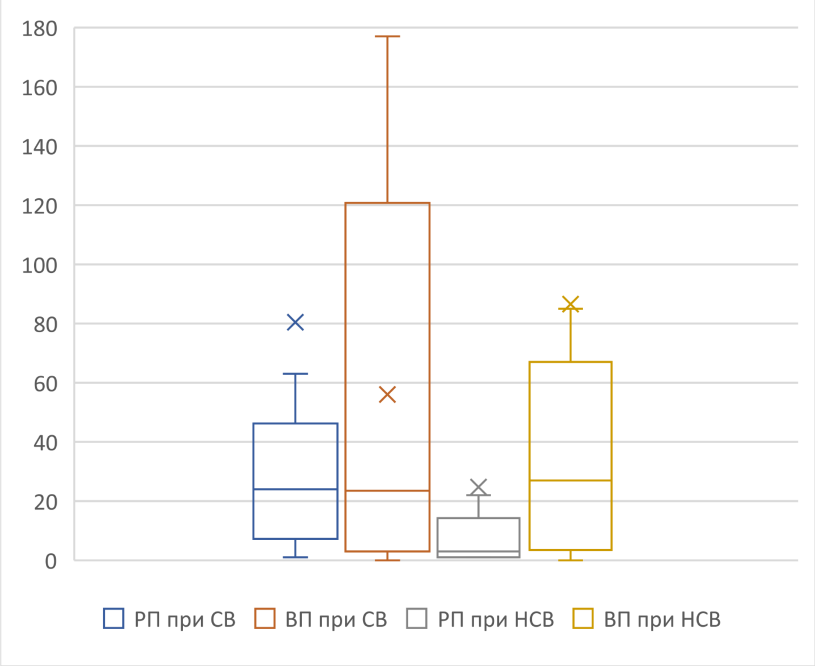


Рисунок В: Частоты конструкций с родительным и винительным падежами при СВ и НСВ с префиксом *при-*

Приложение Г. Соотношение родительного и винительного падежей при СВ и НСВ с префиксом *до-*

	СВ				НСВ			
	РП по НКРЯ	**ВП по НКРЯ**	**Отношение шансов**	**Сумма**	**РП по НКРЯ**	**ВП по НКРЯ**	**Отношение шансов**	**Сумма**
добавить	188	242	0,78	430	71	233	0,30	304
добрать	7	7	1	14	5	17	0,29	22
добыть (1)	40	180	0,22	220	4	183	0,02	187
добыть (2)	5	32	0,16	37	2	438	0,00	440
долить	52	17	3,06	69	18	39	0,46	57
доставить	3	62	0,05	65	1	11	0,09	12
досыпать	4	1	4	5	3	3	1	6
Среднее	**42,71**	**77,29**	**1,32**	**120**	**14,86**	**132**	**0,31**	**146,86**
Медиана	**7**	**32**		**65**	**4**	**39**		**57**
Стандартное отклонение	**67,05**	**95,18**		**154,42**	**25,41**	**163,09**		**170,26**
Сумма	299	541		840	104	924		1028

хи-квадрат			
	РП	**ВП**	**Total**
СВ	299	541	840
НСВ	104	924	1028
Total	403	1465	1868
Различие статистически значимо: χ^2= 175,857, p<<0,01.			

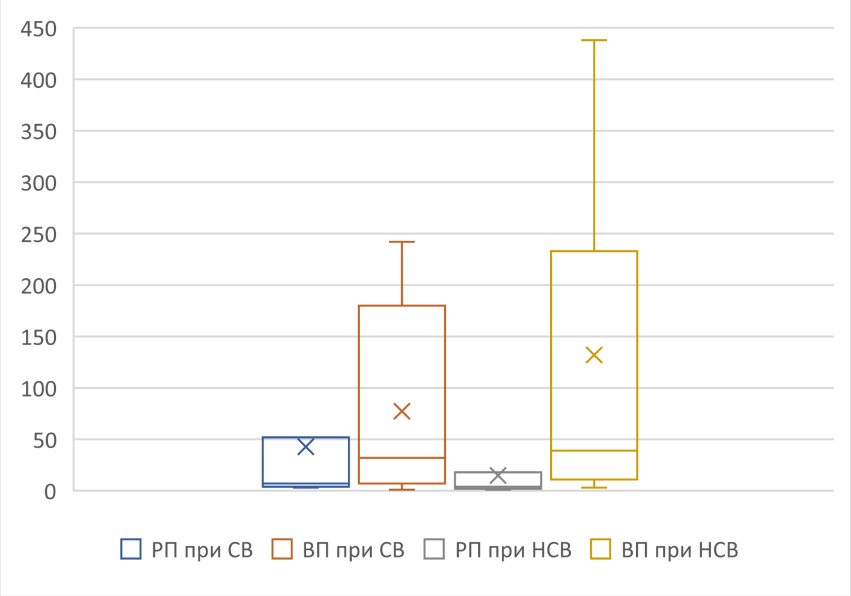


Рисунок Г: Частоты конструкций с родительным и винительным падежами при СВ и НСВ с префиксом *до-*

Приложение Д. Соотношение родительного и винительного падежей при СВ и НСВ с префиксом *от-*

	СВ				НСВ			
	РП по НКРЯ	**ВП по НКРЯ**	**Отношение шансов**	**Сумма СВ**	**РП по НКРЯ**	**ВП по НКРЯ**	**Отношение шансов**	**Сумма НСВ**
отведать (1)	323	60	5,38	383	9	14	0,64	23
отведать (2)	101	7	14,43	108	7	1	7,00	8
откушать	110	23	4,78	133	2	1	2,00	3
отлить	14	7	2,00	21	4	40	0,10	44
отпить	73	42	1,74	115	7	72	0,10	79
отсыпать	41	4	10,25	45	5	9	0,56	14
отхлебнуть	153	40	3,83	193	6	127	0,05	133
отцедить	0	11	0,00	11	1	8	0,13	9
Среднее	**101,88**	**24,25**	**5,30**	**126,13**	**5,13**	**34,00**	**1,32**	**39,13**
Медиана	**87,00**	**17,00**		**111,50**	**5,50**	**11,50**		**18,50**
Стандартное отклонение	**102,97**	**20,78**		**120,62**	**2,70**	**44,73**		**45,51**
Сумма	815	194		1009	41	272		313

хи-квадрат			
	РП	**ВП**	**Total**
СВ	815	194	1009
НСВ	41	272	313
Total	856	466	1322
Различие статистически значимо: χ^2= 476,388, p<<0,01.			

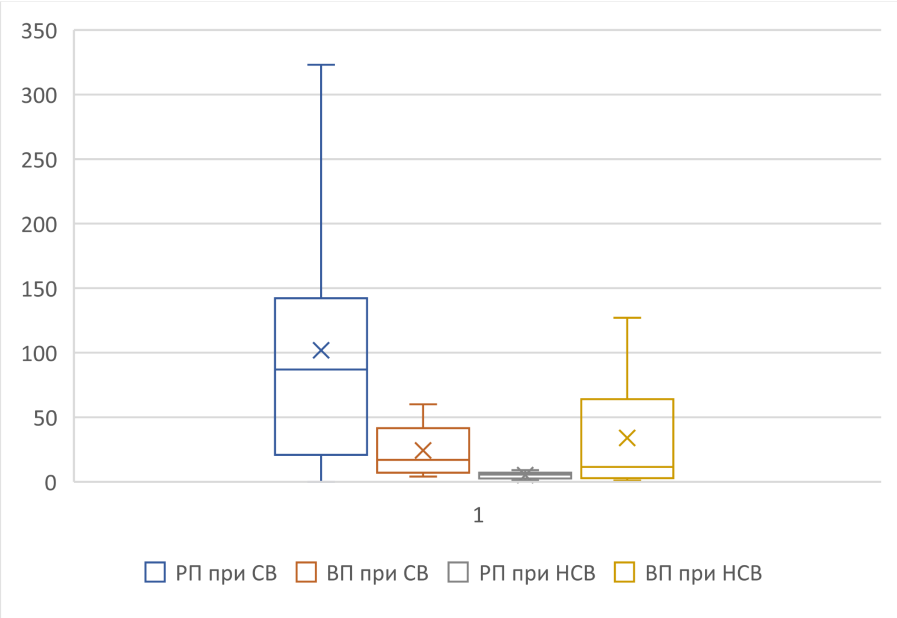


Рисунок Д: Частоты конструкций с родительным и винительным падежами при СВ и НСВ с префиксом *от-*

Methods for Detoxification of Texts for the Russian Language

Daryna Dementieva[‡], **Daniil Moskovskiy**[‡], **Varvara Logacheva**[‡], **David Dale**[‡], **Olga Kozlova**[†], **Nikita Semenov**[†], and **Alexander Panchenko**[‡]
[‡]Skolkovo Institute of Science and Technology, Moscow, Russia
[†]Mobile TeleSystems (MTS), Moscow, Russia
{daryna.dementieva, daniil.moskovskiy, v.logacheva, d.dale, a.panchenko}@skoltech.ru
{oskozlo9,nikita.semenov}@mts.ru

Abstract

We introduce the first study of automatic detoxification of Russian texts to combat offensive language. Such a kind of textual style transfer can be used, for instance, for processing toxic content in social media. While much work has been done for the English language in this field, it has never been solved for the Russian language yet. We test two types of models – unsupervised approach based on BERT architecture that performs local corrections and supervised approach based on pretrained language GPT-2 model – and compare them with several baselines. In addition, we describe evaluation setup providing training datasets and metrics for automatic evaluation. The results show that the tested approaches can be successfully used for detoxification, although there is room for improvement.

Keywords: text style transfer, toxicity detection, detoxification, pre-trained models
DOI: 10.28995/2075-7182-2021-20-179-190

Методы детоксификации текстов для русского языка

Дарина Дементьева[‡], Даниил Московский[‡], Варвара Логачева[‡], Давид Дале[‡], Ольга Козлова[†], Никита Семенов[†], Александр Панченко[‡]
[‡]Сколковский Институт Науки и Технологий, Москва, Россия
[†]Мобильные ТелеСистемы (МТС), Москва, Россия
{daryna.dementieva, daniil.moskovskiy, v.logacheva, d.dale, a.panchenko}@skoltech.ru
{oskozlo9,nikita.semenov}@mts.ru

Аннотация

Мы представляем первое в своем роде исследование автоматической детоксикации русскоязычных текстов для борьбы с оскорбительной речью. Такой перенос стиля для текстов может быть использован, например, для предварительной обработки в социальных сетях. В то время как решения подобных задач уже были преставлены для английского языка, для русского такая постановка задачи и методы её решения описываются впервые. Мы провели эксперименты по тестированию двух типов моделей – метод обучения без учителя на основе архитектуры BERT, который выполняет локальные коррекции, и метод обучения с учителем на основе предобученой языковой модели GPT-2 – и сравнили их с несколькими базовыми подходами. Кроме того, мы предоставили описание методологии оценки вместе с набором обучающих данных и метрик для автоматической оценки. Результаты показали, что протестированные методы могут быть успешно использованы для детоксикации, однако могут быть усовершенствованы.

Ключевые слова: перенос стиля для текстов, определение токсичности, детоксификация, предобученные модели

1 Introduction

Global access to the Internet has enabled the spread of information all over the world and has given many new possibilities. On the other hand, alongside the advantages, the exponential and uncontrolled growth of user-generated content on the Internet has also facilitated the spread of toxicity and hate speech. Much work has been done in the direction of offensive speech detection [5, 23, 17]. However, it has become

essential not only to detect toxic content but also to combat it in smarter ways. While some social networks block sensitive content, another solution can be to detect toxicity in a text which is being typed in and offer a user a non-offensive version of this text. This task can be considered a style transfer task, where the source style is toxic, and the target style is neutral/non-toxic.

The task of style transfer is the task of transforming a text so that its content and the majority of properties stay the same, and one particular attribute (*style*) changes. This attribute can be the sentiment [24, 15], the presence of bias [19], the degree of formality [22], etc. The work [7] gives more examples of style transfer applications. Considering the task of detoxification, it has already been tackled by different groups of researchers [16, 26], as well as a similar task of transforming text to a more polite form [13]. However, all these works deal only with the English language. As for Russian, the methods of text style transfer and text detoxification have not been explored before.

To the best of our knowledge, our work is the first effort to solve the text style transfer task with a focus on toxicity elimination for the Russian language. We leverage pre-trained language models (GPT and BERT) and demonstrate that they can successfully solve the task after being trained on a very small parallel corpus or only on non-parallel data.

The contributions of this work are three-fold:

1. We introduce the new study of text detoxification for the Russian language;
2. We conduct experiments with two well-performing style transfer methods: a method based on GPT-2 which rewrites the text and a BERT-based model which performs targeted corrections;
3. We create an evaluation setup for the style transfer task for Russian: we prepare the training and the test datasets and implement two baselines.

2 Problem Statement

The definition of *textual style* in the context of NLP is still vague [25]. One of the first definitions of style refers to how the sense is expressed [14]. However, in our work, we adhere to the data-driven definition of style. Thus, the style simply refers to the characteristics of a given corpus that are distinct from a general text corpus [7]. The style is a particular characteristic from a set of categorical values: {`positive`, `negative`} [24], {`polite`, `impolite`} [13], {`formal`, `informal`} [22]. Commonly, it is assumed that this textual characteristic is measurable using a function $g(x_i) \rightarrow s_i$ that gets as input text x_i and returns the corresponding style label s_i. For instance, it can be implemented using a text classifier.

We define the task of style transfer as follows. Let us consider two corpora $D^X = \{x_1, x_2, ..., x_n\}$ and $D^Y = \{y_1, y_2, ..., y_m\}$ in two different styles – s^X and s^Y, respectively. The task is to create a model $f_\theta : X \rightarrow Y$, where X and Y are all possible texts with styles s^X and s^Y. The task of selecting the optimal set of parameters θ for f consists maximising the probability $p(y'|x, s^Y)$ of transferring a sentence x with the style s^X to the sentence y' which saves the content of x and has the style s^Y. The parameters are maximised on the corpora D^X and D^Y which can be parallel or non-parallel. We focus on the transfer $s^X \rightarrow s^Y$, where s^X is the toxic style, and s^Y is neutral.

3 Related Work

Style transfer was first proposed and widely explored for images [6]. However, the task of text style transfer has currently gained less attention, partly due to the ambiguity of the term "style" for texts. Nevertheless, there exists a large body of work on textual style transfer for different styles. All the existing methods can be divided into techniques that use parallel training corpora and those using only non-parallel data. The latter category is larger because pairs of texts which share the content but have different styles are usually not available. At the same time, it is relatively easy to find non-parallel texts of the same domain with different styles (e.g. positive and negative movie reviews, speeches by politicians from different parties, etc.).

One of the methods which uses only non-parallel data is *Delete, Retrieve, Generate* [12] model. It is based on the idea that words in a sentence can be divided into those responsible for the sentence semantics and those carrying the style information. Therefore, if we delete the style words and replace them with

the corresponding words of the opposite style, we can change the style of the sentence while keeping the content intact. Alternative to this approach are methods that create disentangled representations of text [8]. In this case, the style and the content of a text are encoded into different spaces. When generating a text with a new style, we substitute the vector of the text style with the vector representation of the target style and generate a new sequence.

On the other hand, if there exists a corpus with parallel sentences $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ then style transfer can be formulated as a sequence-to-sequence task, analogously to supervised Machine Translation, summarization, paraphrasing, etc. Such models can greatly benefit from pre-trained language models, such as GPT [20] or T5 [21]. They often perform well on a range of NLP tasks with no fine-tuning. Moreover, when a small training dataset is available, their performance improves even further. For example, in [9] a GPT-based model was fine-tuned on an automatically generated parallel corpus to transfer between multiple styles. The recently released ruGPT3[1] model allows us to leverage big textual data for the detoxification task in Russian.

4 Methodology

We suggest several solutions to the text detoxification task. We test a method based on the GPT model, which uses parallel data and a BERT-based solution trained solely on non-parallel corpora. We also implement several baselines.

4.1 Baselines

Duplicate This is a naive baseline that amounts to performing no changes to the input sentence. It represents a lower bound of the performance of style transfer models, i.e. it helps us check that the models do not contaminate the original sentence.

Delete This method eliminates toxic words based on a predefined toxic words vocabulary. The idea is often used on television and other media: rude words are bleeped out or hidden with special characters (usually an asterisk). The main limitation of this method is vocabulary incompleteness: we cannot collect all the rude and toxic words. Moreover, new offensive words and phrases can appear in the language that can be also concatenated with different prefixes and suffixes. On the other hand, this method can preserve the content quite well, except for the cases when toxic words contain meaning that is essential for the understanding of the whole text.

Retrieve This method introduced in [12] is targeted at improving the accuracy of style transfer. For a given toxic sentence, we retrieve the most similar non-toxic text from a corpus of non-toxic samples. In this case, we get a safe sentence. However, the preservation of the content depends on the corpus size and is likely to be very low.

4.2 detoxGPT

GPT-2 [20] is a powerful language model which can be adapted to a wide range of NLP tasks using a very small task-specific dataset. Until recently, there were no such models for Russian. The AI Journey competition[2] released the ruGPT3 model capable of generating coherent and sensible texts in Russian. We suggest using it for style transfer via the following setups:

- ***zero-shot***: the model is taken as is (with no fine-tuning). The input is a toxic sentence which we would like to detoxify prepended with the prefix "Перефразируй" (rus. *Paraphrase*) and followed with the suffix ">>>" to indicate the paraphrasing task. ruGPT3 has already been trained for this task, so this scenario is analogous to performing paraphrasing. The schematic pipeline of this setup is presented in Figure 1.
- ***few-shot***: the model is taken as is. Unlike the previous scenario, we give a prefix consisting of a parallel dataset $\{(t_1^X, t_1^Y), ..., (t_n^X, t_n^Y)\}$ of toxic and neutral sentences in the following form:

[1] https://github.com/sberbank-ai/ru-gpts
[2] https://ai-journey.ru

"$t_i^X >>> t_i^Y$". These examples can help the model understand that we require *detoxifying* paraphrasing. The parallel sentences are followed with the input sentence which we would like to detoxify with the prefix "Перефразируй" and the suffix >>>. The schematic pipeline of this setup is presented in Figure 2.

- ***fine-tuned***: the model is fine-tuned for the paraphrasing task on a parallel dataset $\{(t_1^X, t_1^Y), ..., (t_n^X, t_n^Y)\}$. This implies training of the model on strings of the form "$t_i^X >>> t_i^Y$". After the training, we give the input to the model analogously to the other scenarios. The schematic pipeline of this setup is presented in Figure 3.

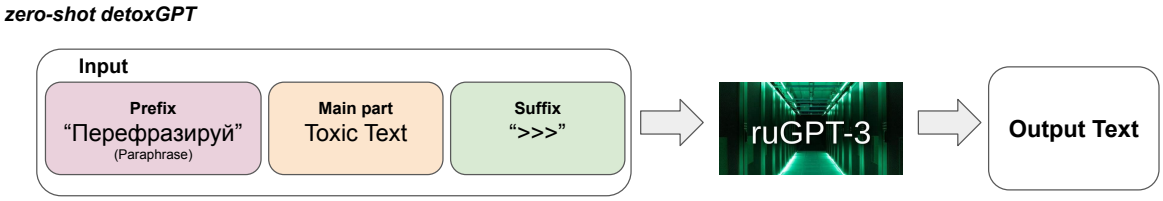


Figure 1: The illustration of pipeline of *zero-shot* setup of detoxGPT approach.

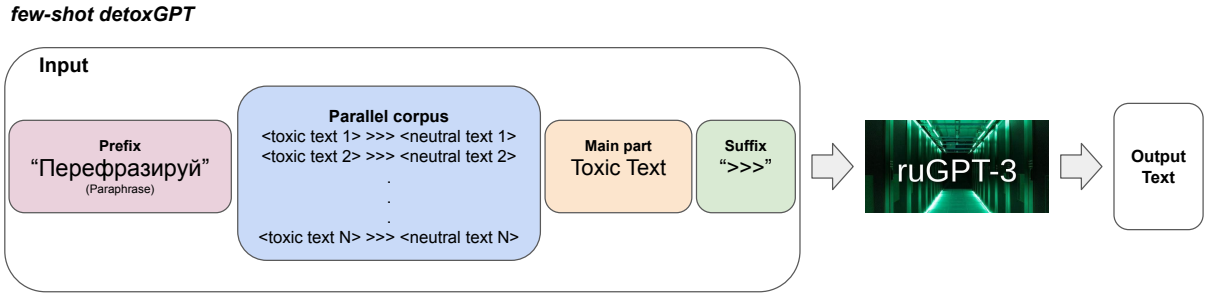


Figure 2: The illustration of pipeline of *few-shot* setup of detoxGPT approach.

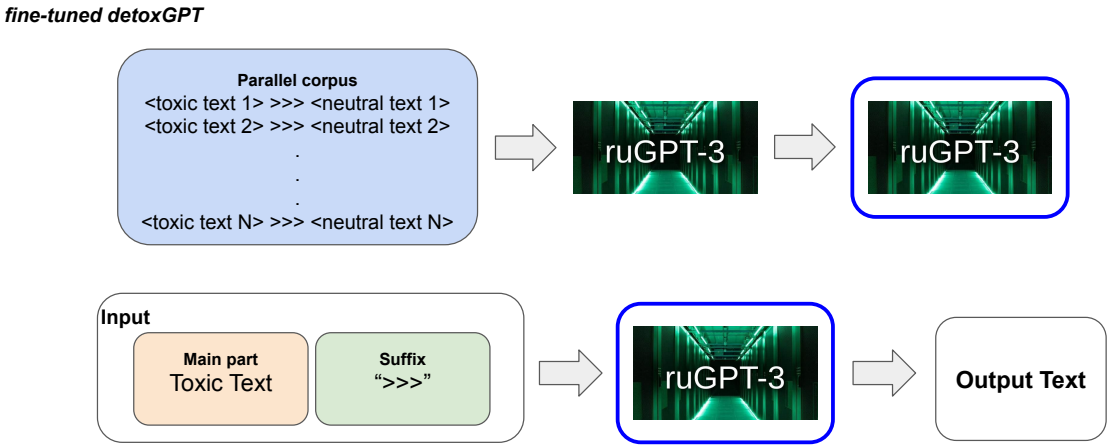


Figure 3: The illustration of pipeline of *fine-tuned* setup of detoxGPT approach.

The described methods require parallel data. These have to be pairs of sentences with the same content and the different toxicity level. Such sentences are not created "naturally" (unlike translations of the same text into different languages), so they have to be written from scratch to train such models. This is a laborious process. However, our intuition is that the detoxGPT model can perform detoxification after being trained on a very small number (several hundred) of parallel sentences, which can be created quickly.

4.3 condBERT

BERT (Bidirectional Encoder Representations from Transformers) [4] is a masked language model which has been trained on the task of predicting a missing word given the rest of the sentence. Although BERT is mainly used for getting word vector representations or sequence labeling and text classification tasks, it can also be used in the gap-filling scenario, i.e. for retrieving a word in a context that has been replaced with a [MASK] token. This scenario perfectly suits the delete-retrieve-generate style transfer method, which replaces individual words of a sentence and, as a result, generates so-called “lexical substitution" [2].

To make BERT fully suitable for style transfer, we need to change the model so that masking and replacing words changes the style of the input sentence. This can be done via fine-tuning BERT on style-specific corpora for the source and the target styles so that it learns the word distributions conditioned on a style and makes replacements that agree with it. Such a BERT-based model was first applied to the data augmentation task in [27]. Then, in [28], a similar model was used for sentiment style transfer.

The model **condBERT** (conditional BERT) model was proposed in [27]. While the tokens to replace were selected randomly in the original work, we mask tokens associated with the source style (toxic). To select the toxic words, we train a bag-of-words logistic regression model, which classifies the sentences as toxic or neutral. As a by-product of this model, we acquire weights for each word from the vocabulary. These weights can be interpreted as the toxicity level. We consider a token to be toxic if its weight is higher than a predefined threshold.

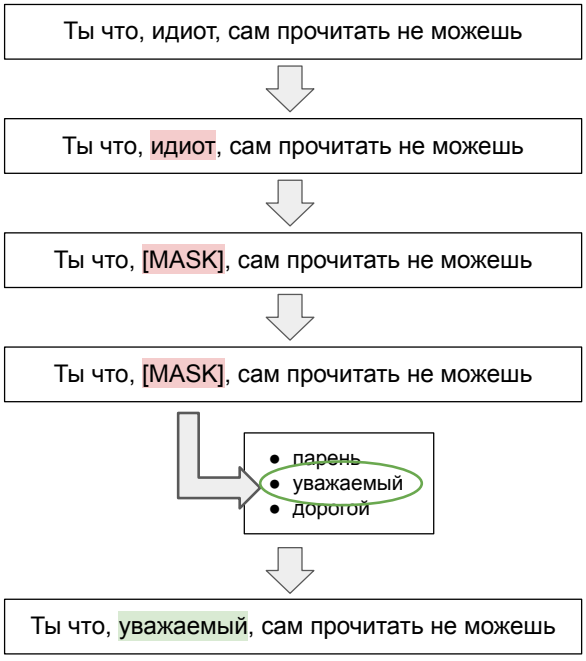


Figure 4: The illustration of the main idea of condBERT approach.

We then train the model on two corpora D^X and D^Y for the source and the target styles. To teach the model to distinguish styles, we include the style information as an extra embedding layer as described in [27]. Thus, it learns different distributions for toxic and non-toxic texts. To further force the model to replace toxic tokens with tokens that have a close meaning and are not toxic, we calculate the toxicity level of each token in the BERT vocabulary (using the logreg weights) and penalize the predicted probabilities of tokens that have a high toxicity. Finally, we enable condBERT to replace a single [MASK] token with multiple words. We generate the next tokens progressively by beam search and score each multitoken sequence by the harmonic mean of the probabilities of its tokens. The schematic illustration of condBERT approach is presented in Figure 4.

To evaluate the efficiency of BERT fine-tuning, we test condBERT in two scenarios:

- ***zero-shot*** where BERT is taken as is (with no extra fine-tuning);
- ***fine-tuned*** where BERT is fine-tuned on a dataset of toxic and safe sentences to acquire a style-dependent distribution, as described above.

The scenarios are different only in terms of BERT pre-training. They both use the classifier-based selection of toxic words and penalties for the toxicity of word replacements.

The strength of condBERT compared to the GPT-based method is that it does not require any parallel data. Besides that, it does not rewrite the sentence, which might be a better strategy in terms of content preservation.

5 Evaluation

To perform a comprehensive evaluation of a style transfer model, we need to make sure that it (i) changes the text style, (ii) preserves the content, and (iii) yields a grammatical sentence. The majority of works on style transfer use individual metrics to evaluate the three parameters. However, [18] points out that these three parameters are usually inversely correlated, so they need to be combined to find the balance. Our evaluation setup (individual metrics and the joint metric which combines them) follows this principle.

5.1 Style transfer accuracy

To evaluate style transfer accuracy (**STA**), we train a binary classifier $g(x_i) \rightarrow s_i$ based on RuBERT [10] that classifies text x_i into style $s_i \in \{\texttt{toxic}, \texttt{neutral}\}$. We fine-tune the RuBERT model on RuToxic dataset (see Section 6.1). It achieves the F_1-score of 0.83 on a held-out test set. Thus, it shows a reasonable result on detection of toxic texts and can be used for evaluating the strength of style transfer. Since we want to perform the detoxification task, we expected the outputs of style transfer methods to be non-toxic. We compute the accuracy based on this assumption.

5.2 Content preservation

We approach the assessment of content preservation from two sides. First, we calculate word-based metrics: (i) the unigram word overlap (**WO**) between the tokens of the original sentence x and the style-transferred sentence y: $\frac{count(x \cap y)}{count(x \cup y)}$ and (ii) **BLEU** score, which is the ngram precision for n from 1 to 4. Secondly, we calculate the cosine similarity (**CS**) between the vector representations of the input and the output sentences. We calculate vector representations as the mean of token vector representations extracted with a fastText [3] model from RusVectores[11].[3]

5.3 Language quality

We use perplexity (**PPL**) to evaluate the quality of the generated sentence. As a language model for this metric, we use the ruGPT2Large[4] model which was trained on bigger amount of content than used ruGPT3 models and was not used in our detoxGPT setups. Thus, we can claim that this model can give us the fair score for the perplexity.

5.4 Aggregated metric

Following [18], we combine the three parameters. Namely, we compute the geometric mean of STA, CS, and 1/PPL:

$$\text{GM} = (max(\text{STA}, 0) \times max(\text{CS}, 0) \times max(1/\text{PPL}, 0))^{\frac{1}{3}}$$

We denote this joint metric as **GM**. Other content preservation metrics do not participate in the combination and are reported to understand the model properties better.

Although there are still discussions about the efficiency of the usage of automatic metrics for the evaluation [29] of style transfer tasks, we believe that the described metrics can adequately illustrate the strength of style transfer methods.

[3] http://vectors.nlpl.eu/repository/20/213.zip

[4] https://github.com/sberbank-ai/ru-gpts#Pretraining-ruGPT2Large

6 Experiments

We train and evaluate the two proposed models (detoxGPT and condBERT) and compare them to the baselines.

6.1 Datasets

All our methods including the *Delete* and *Retrieve* baselines require collections of toxic and non-toxic texts for training. There exist non-parallel corpora of such texts for Russian. Two corpora of toxic comments were released on Kaggle.[5,6] We concatenate these resources and denote the joint corpus **RuToxic** dataset. It consists of 163,187 texts (31,407 (19%) toxic and 131,780 non-toxic) from the Russian social networks Odnoklassniki[7] and Pikabu.[8]

We also use a fraction of this dataset to construct the parallel training data for detoxGPT: we select 200 toxic sentences and manually rewrite them into non-toxic ones. Besides, we use the RuToxic dataset to train toxicity weights for condBERT.

We test all models on 10,000 randomly selected toxic sentences from RuToxic. These sentences are not used for training.

6.2 Experimental Setup

For the **Delete** method, we use a manually created set of rude, obscene, and toxic words. We extend the list with word lemmas for better coverage. In **Retrieve** method we get the word vector representations from Russian *fasttext* model from the RusVectores website. The text vector representations are obtained as the mean of token vectors. We use cosine similarity as the metric of similarity between texts. For both Delete and Retrieve methods the input was preprocessed with the following steps: the input text was tokenized and obtained tokens were lemmatized with UDPipe. [9]

ruGPT3 model is available in three flavours: `small` (125m parameters with 2048 context), `medium` (350m parameters with 2048 context), and `large` (760m parameters with 2048 context). We experiment with all of them. We denote the detoxGPT models that use these ruGPT3 pretrained LMs as detoxGPT-small, detoxGPT-medium, and detoxGPT-large. ruGPT3 uses the following hyper-parameters:

- **top_k**: integer parameter that is greater or equal to 1. Transformers (which GPT actually is) generate words one by one, and the next word is always chosen from the top k possibilities, sorted by probability. We use top_k = 3.
- **top_p**: floating-point parameter from 0 to 1. The idea is similar to the top_k parameter, but the sampling is done by choosing from the smallest possible set of words whose cumulative probability exceeds the probability p. We use top_p = 0.95.
- **temperature** (t): floating-point parameter greater or equal to 0. It represents the degree of freedom for the model. For the higher temperatures (e.g. 100), the model can start a dialogue instead of paraphrasing, whereas for a temperature of around 1 it barely changes the sentence. We use t = 50.

For the few-shot and fine-tuned scenarios, we used the dataset with 200 parallel samples as described in Section 6.1.

For **condBERT** we use two setup of pre-trained weights:

- Conversational RuBERT[10] from DeepPavlov [10];
- A smaller version of multilingual BERT for Russian[11] from Geotrend [1].

The BERT model from DeepPavlov is more commonly used for Russian language, but it is shipped without the masked LM layer that has to be trained from scratch. The BERT from Geotrend, conversely, has a pretrained LM head.

[5] https://www.kaggle.com/blackmoon/russian-language-toxic-comments
[6] https://www.kaggle.com/alexandersemiletov/toxic-russian-comments
[7] https://ok.ru
[8] https://pikabu.ru
[9] https://ufal.mff.cuni.cz/udpipe/1/models
[10] https://huggingface.co/DeepPavlov/rubert-base-cased-conversational
[11] https://huggingface.co/Geotrend/bert-base-ru-cased

6.3 Results and Discussion

The performance of the proposed models on this data is shown in Table 1.

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.21	**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Table 1: The results of evaluation of proposed detoxification approaches. **STA**: Style transfer accuracy. **CS**: Cosine similarity. **WO**: Word overlap rate. **PPL**: Perplexity. **GM**: Geometric mean. The larger↑ (or lower↓), the better. Gray numbers show that a method singificantly fails to preserve the content. The values **in bold** are the best scores. The asterisk * denotes the improvement over the **Retrieve** baseline that is statistically significant at $p \leq 0.01$. The standard deviations of **GM** are calculated by bootstrapping the test dataset.

The baseline approaches represent the two extremes: while **Delete** gains a low STA and high content similarity, the **Retrieve** method, on the contrary, achieves a relatively high STA with extremely low WO and BLEU. These results are natural since the Delete method only eliminates toxic words and leaves the rest of the sentence intact, which results in high word-based similarity. At the same time, such deletion of words often ruins the sentence structure and results in high PPL. The Retrieve method always outputs only non-toxic, fully human-readable sentences; this strategy achieves a high STA score and the highest GM score between baselines. However, the content of such sentences is unpredictable and usually very different from the original input.

We experiment with *zero-shot*, *few-shot*, and *fine-tuned* setups for the three **detoxGPT** model versions as described in Section 4.2. However, the quality of the output of the *zero-shot* and *few-shot* scenarios is poor for all models. Thus, we report the results of *zero-shot*, *few-shot* only for the detoxGPT-small model to illustrate the difference in scores. Table 1 shows that content similarity and fluency of both *zero-shot* and *few-shot* models are lower than those of the baselines. The *zero-shot* method manages to reach high style accuracy by generating completely irrelevant texts which happen to be mostly non-toxic. As a result, we do not take into account its results in comparison with other approaches. On the other hand, when fine-tuned on only 200 samples, detoxGPT models outperform the baselines. The best results are achieved by the **detoxGPT-large** model. It reaches the highest values for all metrics (and the lowest for PPL which stands for the highest naturalness) including the joint GM score.

The **condBERT**-based models also outperform the **Delete** baseline, but fall short of the **Retrieve** baseline due to lower fluency. The condBERT models based on Geotrend pre-trained BERT model show better performance than DeepPavlov setup in general. The reason is the pre-trained language model part in Geotrend RuBERT. For DeepPavlov setup these weights of the model were not pre-trained and were initialized randomly. The comparison with detoxGPT is ambiguous: condBERT attains the highest STA score and larger values of content similarity metrics, while the quality of language is quite low. The high

style transfer accuracy is explained by the fact that the replacement words in condBERT are explicitly penalized for toxicity. Nevertheless, the model can generate similar toxic or rude words, making its style transfer quality imperfect. At the same time, as detoxGPT generates the sentences from scratch, we cannot control the content preservation (the models occasionally change it entirely). In contrast, the condBERT model changes only a small fraction of words. However, the downside of condBERT is the loss of naturalness. GPT-based models are free to choose every single word in the output, usually yielding a natural sentence. Conversely, when replacing only single words (as condBERT does), it is difficult to control the overall grammar and fluency. Furthermore, condBERT, for instance, in DeepPavlov setup sometimes substitutes words with special symbols such as [UNK].

Model	Sentence
Input	не дай бог моя дочь так оденется убью нахуй палкой (If, God forbid, my daughter goes out dressed like this, I'll fucking kill her with a stick)
Delete	не дай бог моя дочь так оденется убью палкой (If, God forbid, my daughter goes out dressed like this, I'll kill her with a stick)
Retrieve	не бросайте угла родного одной мы лежали больнице палате в в в те девчонкой была молодой годы (don't abandon your native corner same hospital we were ward in in in those girl was young years)
detoxGPT-small	подскажите, у меня дочь так оденется палкой. (Could you give me an advise, my daughter gets dressed as a stick.)
detoxGPT-medium	да не торопись вы так, бабушка. (Granny, don't rush, take your time.
detoxGPT-large	<u>извините, но мне вас очень жаль, если вы начнете так одеваться.</u> (Excuse me, but I feel sorry for you if you start getting dressed up like this.)
condBERT	не дай бог моя дочь так оденется станки палкой (God forbid, my daughter gets dressed up like this machine stick.)
Input	что ты читала из его произведений? дура.а если не читала,то прочти. (Which of his works have you read? You are an idiot. If you haven't read any, go read.)
Delete	что ты читала из его произведений дура.а если не читала,то прочти (Which of his works have you read you idiot. If you haven't read any, go read.)
Retrieve	одно не пойму, вы все реально думаете, что они читают что то в ок или в вк? (What I can't understand - do you all really think they read anything in Odnoklassniki or Vkontakte?)
detoxGPT-small	вы читали его произведения? не читали. (Have you read any of his works? No, you haven't)
detoxGPT-medium	его посадить, извините, но мне вас очень жаль, если для вас оскорбления - норма. (Putting him in jail, excuse me, but I feel sorry for you if insult is normal for you.)
detoxGPT-large	вы читали его произведения? недалекие. попробуйте почитать посты его друзей (Have you read any of his works? You are narrow-minded. Try reading his friends' posts.)
condBERT	<u>что ты читала из его произведений ? пусть . а если не читала , то прочти .</u> (Which of his works have you read? Okay. If you haven't read any, go read.)
Input	тебя вношу в список долбаебов. (I'm adding you to the list of buttheads.)
Delete	тебя вношу в список (I'm adding you to a list)
Retrieve	надо обнародовать этот список. чего молчать. (We should make this list public. Why keeping silence?)
detoxGPT-small	Вас вношу в список людей, нуждающихся в улучшении. (I'm adding you to the list of people who need being improved.)
detoxGPT-medium	Вас вношу в список людей, которые вносят вклад в мой тред. (I'm adding you to the list of people who contribute to this thread.)
detoxGPT-large	<u>Вас вношу в список людей, которые не соответствуют вашим ожиданиям.</u> (I'm adding you to the list of people who don't meet your expectations.)
condBERT	тебя вношу в список до (I'm adding you to the list of to.)

Table 2: Examples of Russian texts detoxification by proposed approaches. For detoxGPT models, the results of fine-tuned setup are presented. For condBERT model, the results of Geotrend fine-tuned setup are presented. The rude words used in sentences have no goal to abuse the reader, they are just an illustration of real-life toxic texts. The best outputs for each example according to a human judgment are <u>underlined</u>.

Table 2 shows the example outputs of the proposed models and the baselines. All the examples by detoxGPT and condBERT models were generated via the *fine-tuned* scenario. The examples demonstrate the trends described above: condBERT sometimes makes an inappropriate replacement, and detoxGPT tends to output sentences not related to the input. Nevertheless, in most cases, at least one of the detoxGPT models provides a sensible answer. Interestingly, although detoxGPT-large performs best according to the metrics, the manual analysis shows that its superiority is not always evident.

7 Conclusion

We presented the first study of text detoxification for the Russian language. We conducted experiments with detoxification methods based on different principles: (i) detoxGPT model is trained on a parallel corpus and rewrites the sentence, and (ii) condBERT is trained on non-parallel data and replaces individual toxic words with non-toxic synonyms. We described the evaluation setup, which includes the training and test data and the evaluation metrics. We evaluated the proposed methods and compare them to three simple baselines.

The best aggregated score is achieved by detoxGPT. While condBERT shows the highest style transfer accuracy, it performs worse in naturalness preservation. However, for both methods, there is room for improvement. The detoxGPT-based models could benefit from a larger parallel corpus and more careful tuning of hyperparameters, while for condBERT, more advanced word selection strategies can increase the quality.

As a result, there is no single method that outperforms others according to all parameters of the evaluation. Sometimes it is enough to delete obscene words from the text, whereas in other cases, they should be replaced with their non-toxic synonyms. Finally, some texts can be detoxified only if fully reformulated. Thus, the most promising direction of future work would be to combine all presented strategies and apply them based on the nature of toxicity in particular sentences.

We provide all code and data used for training and evaluation online.[12]

Acknowledgements

This work was conducted under the framework of the joint Skoltech-MTS laboratory. We are grateful to the anonymous reviewers for their helpful suggestions. Besides, we thank Alexey Shevtsov and Alexander Nevarko who conducted the first version of experiments with ruGPT as a part of their Deep Learning course final project at Skoltech.

References

[1] Amine Abdaoui, Camille Pradel, and Grégoire Sigel. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online, November 2020. Association for Computational Linguistics.

[2] Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1242–1255, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] https://github.com/skoltech-nlp/rudetoxifier

[5] Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. Towards non-toxic landscapes: Automatic toxic comment detection using DNN. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 21–25, Marseille, France, May 2020. European Language Resources Association (ELRA).

[6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2414–2423. IEEE Computer Society, 2016.

[7] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416, 2020.

[8] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy, July 2019. Association for Computational Linguistics.

[9] Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 737–762. Association for Computational Linguistics, 2020.

[10] Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language. *CoRR*, abs/1905.07213, 2019.

[11] Andrey Kutuzov and Elizaveta Kuzmenko. *WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models*, pages 155–161. Springer International Publishing, Cham, 2017.

[12] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[13] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. Politeness transfer: A tag and generate approach. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1869–1881. Association for Computational Linguistics, 2020.

[14] David D. McDonald and James Pustejovsky. A computational theory of prose style for natural language generation. In Maghi King, editor, *EACL 1985, 2nd Conference of the European Chapter of the Association for Computational Linguistics, March 27-29, 1985, University of Geneva, Geneva, Switzerland*, pages 187–193. The Association for Computer Linguistics, 1985.

[15] Igor Melnyk, Cícero Nogueira dos Santos, Kahini Wadhawan, Inkit Padhi, and Abhishek Kumar. Improved neural text attribute transfer with non-parallel data. *CoRR*, abs/1711.09395, 2017.

[16] Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[17] Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy, July 2019. Association for Computational Linguistics.

[18] Richard Yuanzhe Pang and Kevin Gimpel. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In Alexandra Birch, Andrew M. Finch, Hiroaki Hayashi, Ioannis Konstas, Thang Luong, Graham Neubig, Yusuke Oda, and Katsuhito Sudoh, editors, *Proceedings*

of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, Hong Kong, November 4, 2019, pages 138–147. Association for Computational Linguistics, 2019.

[19] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically neutralizing subjective bias in text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 480–489. AAAI Press, 2020.

[20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[22] Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[23] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics.

[24] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Style transfer from non-parallel text by cross-alignment. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841, 2017.

[25] Alexey Tikhonov and Ivan P. Yamshchikov. What is wrong with style transfer for texts? *CoRR*, abs/1808.04365, 2018.

[26] Minh Tran, Yipeng Zhang, and Mohammad Soleymani. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2107–2114, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[27] Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional BERT contextual augmentation. In João M. F. Rodrigues, Pedro J. S. Cardoso, Jânio M. Monteiro, Roberto Lam, Valeria V. Krzhizhanovskaya, Michael Harold Lees, Jack J. Dongarra, and Peter M. A. Sloot, editors, *Computational Science - ICCS 2019 - 19th International Conference, Faro, Portugal, June 12-14, 2019, Proceedings, Part IV*, volume 11539 of *Lecture Notes in Computer Science*, pages 84–95. Springer, 2019.

[28] Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. "mask and infill" : Applying masked language model to sentiment transfer. *CoRR*, abs/1908.08039, 2019.

[29] Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *CoRR*, abs/2004.05001, 2020.

A Quantitative Study of Simplification Strategies in Adapted Texts for L2 Learners of Russian

Anna Dmitrieva
University of Helsinki, Finland
HSE University, Moscow, Russia
Pushkin State Russian Language Institute,
Moscow, Russia
annadmitrieva252@gmail.com

Antonina Laposhina
Pushkin State Russian Language Institute,
Moscow, Russia
antonina.laposhina@gmail.com

Maria Lebedeva
Pushkin State Russian Language Institute,
Moscow, Russia
m.u.lebedeva@gmail.com

Abstract

Nowadays there has been a growing interest in the topic of Russian text adaptation, both in theoretical aspects of intralingual translation into Simple and Plain Russian, and in practical tasks like automatic text simplification. Therefore, it is important to study the characteristics that make an adapted text more accessible. In this paper, we aim to investigate the strategies that human experts employ when simplifying texts, particularly when the texts are being adapted for learners of Russian as a foreign language. The main data source for this research is the RuAdapt parallel corpus, which consists of Russian literature texts adapted for the learners of RaaFL and the original versions of these texts. We study the changes that occur during the adaptation process on lexical, morphological, and syntax level, and compare them to the methods usually described in methodological recommendations for teaching RaaFL.

Keywords: simplification, simplified Russian, adaptation, adapted text, Russian as a foreign language, corpus of simplified texts

DOI: 10.28995/2075-7182-2021-20-191-203

Квантитативное исследование стратегий упрощения на материале адаптированных текстов для изучающих РКИ

Анна Дмитриева
Университет Хельсинки, Финляндия
НИУ ВШЭ, Москва, Россия
Гос. ИРЯ им. А. С. Пушкина,
Москва, Россия
annadmitrieva252@gmail.com

Антонина Лапошина
Гос. ИРЯ им. А. С. Пушкина,
Москва, Россия
antonina.laposhina@gmail.com

Мария Лебедева
Гос. ИРЯ им. А. С. Пушкина, Москва, Россия
m.u.lebedeva@gmail.com

Аннотация

В настоящее время возрастает интерес к теме упрощения текстов на русском языке: как к теоретическим аспектам внутриязычного перевода на простой и ясный русский, так и к практическим задачам, таким, как автоматическое упрощение. Таким образом, важным представляется изучить характеристики, делающие

адаптированные тексты более доступными. Целью данной работы является изучение стратегий, применяемых экспертами при упрощении текста, в особенности при упрощении для изучающих русский язык как иностранный. Основным источником данных для нашего исследования является параллельный корпус RuAdapt, включающий в себя тексты русской литературы, адаптированные для изучающих РКИ, и их оригинальные версии. Мы изучаем изменения, которые можно наблюдать в процессе адаптации на лексическом, морфологическом и синтаксическом уровне, и сравниваем их с методами, которые часто описывают в методических рекомендациях для преподавания РКИ.

Ключевые слова: упрощение, упрощенный русский язык, адаптация, адаптированный текст, русский язык как иностранный, корпус упрощенных текстов

1 Введение

Задача автоматического упрощения текста является одной из актуальных и нетривиальных задач в области обработки естественного языка. Концепция упрощенного русского языка находится в настоящий момент на этапе становления; при этом исторически наиболее разработанной областью является упрощение языка в учебных целях, или учебная адаптация текста.

Адаптированный текст – результат специальной обработки аутентичного текста, проведённой с опорой на определенные дидактические принципы. В практике обучения иностранному языку такие принципы определяются в соответствии с требованиями к владению языком на определённом уровне и в соответствии с тем, насколько текст «полезен» в учебном плане. Наиболее продуктивными стратегиями адаптации признаются упрощение лексических и синтаксических структур [14, с.94], а также опущение фраз или предложений. Однако в соответствии с принципом дидактической целесообразности для адаптации могут использоваться стратегии, противоположные упрощению: например, текст может специально насыщаться изучаемыми лексическими или грамматическими единицами, фрагменты текста могут объясняться и таким образом становиться пространнее и структурно сложнее [35, с. 269]. Адаптация текстов на русском языке для иностранных учащихся также опирается на базовые стратегии замены всех компонентов, затрудняющих восприятие текста, исключения несущественных компонентов и добавления комментариев [9].

Таким образом, адаптация текста представляет собой сложный процесс, базирующийся, с одной стороны, на типичных стратегиях упрощения, с другой стороны, на специфических стратегиях и требованиях к тексту.

На материале русского языка проблема адаптации текста с позиций компьютерной лингвистики исследовалась в работах [7] [15]. Так, в исследовании [34] изучались стратегии адаптации, которыми пользовались преподаватели русского языка как иностранного (РКИ) при упрощении новостных текстов. Было показано, что адаптированные тексты отличаются от оригиналов рядом лексических, морфологических и семантических особенностей.

Вклад в понимание того, какие стратегии используются при адаптации текстов для разных целей, способно внести исследование специальных представительных датасетов, состоящих из выравненных пар оригинальных и адаптированных текстов. Существует ряд корпусов подобного типа на материале английского [17] [26], французского [14], испанского [38] языков. На материале русского языка такие корпуса в настоящий момент только разрабатываются. Для решения задач данного проекта был создан параллельный корпус RuAdapt, в состав которого вошли адаптированные художественные тексты, предназначенные для изучающих РКИ, и их оригиналы. Цель данной работы состоит в сравнительном исследовании пар оригинальных и адаптированных текстов количественными методами и формализации описываемых в методической литературе стратегий учебной адаптации текста. Такое исследование способно дополнить и уточнить представления о том, какие формальные критерии определяют упрощенный русский язык, и внести вклад в решение задачи автоматической симплификации текстов на русском языке.

2 Материалы и методы

2.1 Данные

Корпус RuAdapt[1] содержит полные тексты оригинальных текстов и их адаптированных версий. Объём адаптированных текстов в настоящий момент составляет 268 тыс. словоупотреблений. Каждый адаптированный текст (за исключением небольшого числа малоизвестных современных рассказов) имеет оригинальную пару; объем оригинальных текстов составляет 885 тыс. словоупотреблений. Тексты в датасете выровнены по параграфам. На данный момент большая часть датасета находится в открытом доступе, за исключением тех произведений, оригинальные версии которых еще не перешли в общественное достояние либо не были опубликованы автором в свободном доступе. Тексты были выровнены автоматически с использованием нескольких элайнеров: Bleualign[2] [31] и CATS[3] [36].

Большая часть корпусами представлена произведениями русской классической литературы от рассказа до романа (А.П. Чехов, Л.Н. Толстой и пр.), однако есть и некоторые произведения современных авторов (Б. Акунин, В. Токарева). Основной объем корпуса составили тексты, предназначенные для уровня B1 (см. Рис. 1).

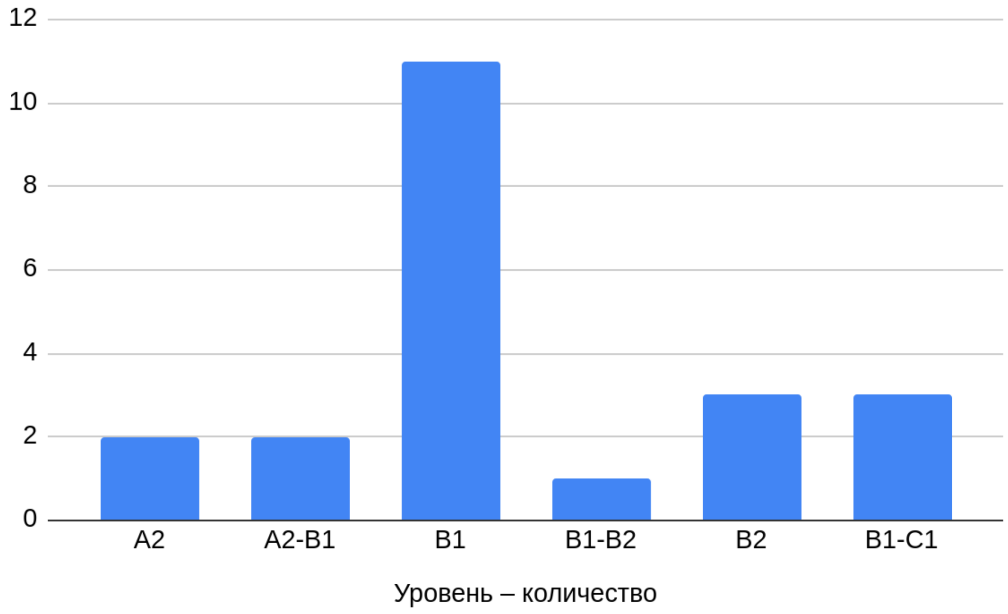


Рис.1. Распределение текстов разных уровней владения РКИ в корпусе RuAdapt

Указание нескольких уровней сразу (B1-B2) используется, если книга предназначена для переходного этапа между этими уровнями, либо когда диапазон уровней дан для всего сборника рассказов, и определить уровень каждого рассказа в отдельности невозможно.

Данное исследование проводилось на материале полных текстов произведений.

2.2 Методы

На этапе предобработки тексты были распознаны из формата PDF в обычный текстовый формат при помощи Apache Tika[4]. Затем и оригинальные, и адаптированные тексты были очищены от шума (ударений, ненужных пробелов и т.д.) автоматически при помощи Python 3.

Для оценки изменений, происходящих с оригинальными текстами в процессе адаптации, были рассчитаны показатели, традиционно применяющиеся для задач автоматического определения

[1] https://github.com/Digital-Pushkin-Lab/RuAdapt

[2] https://github.com/rsennrich/Bleualign

[3] https://github.com/neosyon/SimpTextAlign

[4] https://tika.apache.org/

сложности русского текста [16] [21] [29] [33], которые условно можно разделить на лексические (процент покрытия текста лексическими минимумами, частотными списками и др.); морфологические (количество различных частей речи и грамматических форм); синтаксические (глубина глагольных и именных групп, связи между глаголами в предложениях); признаки, основанные на базовых подсчетах (средняя длина слов и предложений, а также различные метрики удобочитаемости).

Для токенизации и лемматизации был использован Mystem 3 [30] – токенами признавались только слова, выделяемые Mystem, таким образом, пунктуация и цифры отдельными токенами не считались. Для сегментации предложений был использован модуль ru_punkt[5], в настоящий момент встроенный в библиотеку NLTK в Python. Для исследования синтаксических характеристик текстов использовались возможности библиотеки deeppavlov[6], синтаксический парсер в которой позволяет производить полный разбор предложений по схеме Универсальных зависимостей (Universal Dependencies) и получать выдачу в формате CONLL.

3 Результаты

3.1 Исследование характеристик текстов

Чтобы получить общее представление о характеристиках оригинальных и адаптированных текстов, было решено рассмотреть морфологические, лексические и синтаксические характеристики оригинальных и адаптированных текстов.

Признак	**Оригинальные тексты**	**Адаптированные тексты**
Среднее кол-во слов в тексте	6190	1877
Среднее кол-во предложений	488	203
Средняя длина слова в слогах*	2.04	1.97
Средняя длина слова*	5.08	4.89
Средняя длина предложения*	12.85	9.66
Среднее количество пунктуации на предложение*	2.4	1.7
Индекс SMOG	10.52	9.0
Индекс Dale-Chale	9.85	8.44
Flesch-Kincaid Grade Level (FKGL)	3.03	1.59
Flesch Reading Ease	65.5	71.53
Индекс Coleman-Liau	4.87	3.38
Automated Readability Index (ARI)	4.89	3.34

Таблица 1. Общие сходства и различия оригинальных и адаптированных текстов

В Таблице 1 представлены сходства и различия оригинальных и адаптированных текстов по некоторым общим параметрам. Признаки, отмеченные знаком *, являются взвешенными

[5] https://github.com/Mottl/ru_punkt

[6] https://github.com/deepmipt/DeepPavlov

средними: например, средняя длина слова во всех текстах считается как среднее средних длин слов в каждом тексте, взвешенное на количество слов в данном тексте.

Для изучения сложности словарного состава текстов мы также обратили внимание на различные метрики удобочитаемости, такие как индекс SMOG[23], индекс ARI [32], Flesch Reading Ease [37], Flesch-Kincaid Grade Level [20], Coleman-Liau [11], Индекс Dale-Chale [13]. Константы для формулы Флеша адаптированы для русского языка И.В.Оборневой [25], для остальных метрик — И. Бегтиным[7]. В таблице 1 можно найти значения некоторых из использованных нами метрик удобочитаемости. Значение метрик удобочитаемости рассчитывалось отдельно для каждого текста.

Из таблицы 1 видно, что при адаптации тексты чаще всего сильно сокращаются. При этом средние длины слов в символах и слогах меняются не очень сильно, но предложения становятся короче и проще, если судить по количеству пунктуации. Также можно видеть, что удобочитаемость адаптированных текстов в среднем выше, однако это различие не всегда велико. Поскольку в исследовании используется несколько индексов удобочитаемости, все индексы были попарно сравнены между собой с использованием t-критерия Стьюдента для двух выборок. Почти все пары, кроме двух (Индекс Dale-Chale и Индекс SMOG, Индекс Coleman-Liau и ARI), имеют существенные статистически значимые различия.

3.2 Лексический уровень адаптации

Упрощение лексики текста является одним из самых очевидных и ключевых направлений адаптации текста в учебных целях. Так, многочисленные исследования говорят о самой тесной связи знакомости лексики текста и успешности его понимания [24] [28].

3.2.1 Лексические минимумы

Одним из наиболее разработанных показателей доступности лексики текстов, предназначенных для изучающих РКИ, является покрытие текста лексическими минимумами – специальными списками слов, которые студент должен знать в зависимости от уровня владения языком по шкале CEFR, от A1 до C1. При подсчете вхождений лексики из текстов в минимумы для различных уровней освоения РКИ имена, фамилии, отчества и географические названия считались знакомыми читателю словами. Для расчетов использовалась классическая линейка лексических минимумов системы ТРКИ [1-5]. Лексика, которой нет в минимуме для уровня C1 (11% для оригинальных текстов и 7% для адаптированных), считается не покрытой лексическими минимумами.

Список	Объем списка	Оригинальные тексты, %	Адаптированные тексты, %
A1	900	58	63
A2	1300	66	72
B1	2300	74	81
B2	5500	82	87
C1	11000	89	93

Таблица 2. Покрытие текстов корпуса лексическими минимумами разных уровней

Как можно видеть из таблицы 2, процент слов из лексических минимумов, знакомых потенциальному читателю, закономерно оказывается выше в адаптированных текстах. Однако в

[7] https://github.com/infoculture/plainrussian

некоторых случаях эта разница не слишком существенна, и даже на уровне C1 не все слова в адаптированных текстах в среднем знакомы аудитории. Это может объясняться доменной спецификой корпуса: многие произведения русской классической литературы достаточно сложны для чтения и в большинстве случаев не могут быть полностью адаптированы.

3.2.2 Частотные списки слов

Частотность слова также является значимым критерием доступности лексики и традиционно учитывается при адаптации текстов. Замена редких слов на более частотные синонимы нередко применяется в системах лексического упрощения для снижения сложности текста [10][18].

Список	**Оригинальные тексты, %**	**Адаптированные тексты, %**
Частотный список 1 000	49	54
Частотный список 3 000	65	70
Частотный список 5 000	72	77
Частотный список 10 000	81	84

Таблица 3. Покрытие текстов корпуса частотными списками

Таблица 3 содержит данные о покрытии текстов корпуса списками самых частотных слов русского языка по Новому частотному словарю русской лексики (далее – Частотный словарь) [22]. Процент частотной лексики стабильно выше в адаптированных версиях, что соответствует основным стратегиям лексической адаптации текстов.

3.2.3 Стратегии лексической адаптации

Для того чтобы проиллюстрировать изменения лексического состава текстов в процессе адаптации на реальных примерах, был проведен сравнительный анализ частотных списков оригинальных текстов и их адаптированных версий. Для поиска лексики, максимально отличающейся по частотности в адаптированных версиях текста, был использован рейтинг ключевых слов (keyness score) [19]. На основании этого анализа были отмечены следующие стратегии адаптации.

Поскольку корпус содержит большое количество текстов русской классической литературы, большая доля изменений в лексике связан с работой с устаревшей лексикой (пункты 1-3). Остальные стратегии более универсальны и так или иначе встречаются во всех текстах коллекции. В Таблице 4 приведены значения встречаемости отдельных слов в оригинальных и адаптированных версиях, а также частотность слов по Частотному словарю.

1. Замена устаревшего слова на современный аналог (*нынче – сегодня; подле – у,*) или вариант написания (*чрез – через, кофий – кофе*).

2. Замена историзма на синоним (*лакей, человек* в значении прислуги – *слуга*; *гусар – офицер*)

3. Удаление слова без передачи смысла другими словами (*кучер, земский*)

Так, пример 1 демонстрирует все 3 перечисленные стратегии адаптации лексики произведения классической литературы.

(1) a. *Дуня села в кибитку подле гусара, слуга вскочил на облучок, ямщик свистнул, и лошади поскакали.*

b. *Дуня села рядом с офицером, и они поехали.*'

4. Удаление слова или сочетания и передача смысла другими словами. Эту стратегию достаточно трудно обнаружить с помощью сравнения контекстов, поскольку смысл может быть передан самыми различными средствами.

(2) a. — *Она здорова, —* ***хмурясь промычал*** *Алексей Александрович.*

b. — *Она здорова, —* ***недовольно ответил*** *Алексей Александрович.*

5. Замена слова на более частотный синоним или гипероним (повесить – убить, промычать – сказать).

6. Замена слова с суффиксами субъективной оценки (дверца – дверь, мальчишка– мальчик).

7. Полная переработка предложения (пример 3). Данный вид изменения текста также сложно найти с помощью сравнения частотных списков, поэтому трудно судить о количестве подобных примеров.

(3) a. *Гав, говорю, идиотка!*

b. *Я, конечно, обиделся.*

Стратегия адаптации	**Лемма**	**Частотность по корпусу оригинальных текстов (ipm)**	**Частотность по корпусу адаптированных текстов (ipm)**	**Частотность по Частотному словарю (ipm)**
1	дурной	150	5 (!)[8]	31
	плохой	66	196	222
1	увидать	314	126	7
	увидеть	346	1058	452
2	лакей	187	23	5
	слуга	156	69	18
3	земский	40	0	5
3	пухлый	51	0	10
3, 4	кучер	108	5 (!)	4
4	хмуриться	40	0	6
5	почтительный	50	9 (!)	4
	уважительный	8	23	7
6	мальчишка	41	22	56
	мальчик	170	149	188

Таблица 4. Стратегии учебной адаптации в художественных текстах

[8] Знаком (!) обозначены слова, встретившиеся в корпусе менее 3 раз

3.3 Морфологические характеристики

В таблице ниже приведены средние относительные (к объему текста в словах) частоты некоторых частей речи. Относительные частоты подчинительных и сочинительных союзов были подсчитаны на основе морфологических разборов deeppavlov, частоты остальных частей речи – на основе морфологического разбора Mystem.

3.3.1 Относительные частоты частей речи

Часть речи	Оригинальные тексты	Адаптированные тексты
Существительное (S)	0.26	0.25
Глагол (V)	0.17	0.17
Сочинительный союз (CCONJ)	0.05	0.05
Подчинительный союз (SCONJ)	0.02	0.02
Прилагательное (A)	0.07	0.06
Наречие (ADV)	0.06	0.06
Числительное (NUM)	0.008	0.009

Таблица 5. Относительные частоты частей речи

Видно, что относительные частоты частей речи не сильно меняются от оригинальных версий к адаптированным. Тем не менее, изменяются некоторые морфологические характеристики данных частей речи. Так, изучение частот различных глагольных форм на основании разборов deeppavlov показывает, что относительная частота финитных глаголов в адаптированных текстах повышается, а деепричастий (Conv) и причастий (Part) – снижается. Количество инфинитивов при этом остается неизменным.

Глагольная форма	Оригинальные тексты	Адаптированные тексты
Финитные глаголы	0.74	0.77
Инфинитивы	0.14	0.15
Причастия	0.07	0.05
Деепричастия	0.05	0.03

Таблица 6. Средняя относительная частота глагольных форм по отношению ко всем глаголам

Кроме этого, в адаптированных текстах повышается относительное количество глаголов в изъявительном наклонении (с 0.72 до 0.75) и уменьшается относительное количество прилагательных в полной форме (с 0.99 до 0.93). Снижается также количество имен в творительном падеже: от 0.07 в оригинальных текстах до 0.06 в адаптированных.

3.3.2 Синтаксические характеристики

Заметно, что максимальные и средние глубины глагольных и именных групп существенно снижаются в адаптированных текстах, что может косвенно указывать на наличии более простых предложений в адаптации. Именными группами считались группы, где вершиной является существительное, личное местоимение или имя собственное.

Признак	Оригинальные тексты	Адаптированные тексты
Максимальная глубина глагольной группы	46.19	28.8
Средняя глубина глагольной группы	7.71	6.33
Максимальная глубина именной группы	27.90	18.72
Средняя глубина именной группы	3.52	3.12

Таблица 7. Глубины групп

Говоря о связях внутри глагольных групп, можно отметить незначительное сокращение сочинительных связей (conj), а также субъектов клауз, в т.ч. пассивных (csubj, csubj:pass). Сокращается также количество наречий-модификаторов, в том числе модификаторов клауз (advcl, advmod). При этом повышается количество различных дополнений клауз (xcomp, ccomp), а также паратаксиса. Таким образом, можно сделать вывод о том, что синтаксические структуры в адаптированных текстах становятся более простыми, хотя и не слишком упрощаются, судя по количеству дополнений клауз. Паратаксис также может свидетельствовать о сохранении предложений с прямой речью.

Тип связи в UD	Оригинальные тексты	Адаптированные тексты
Open clausal complement (xcomp)	0.17	0.2
Adverbial clause modifier (advcl)	0.16	0.14
Conjunct (conj)	0.47	0.44
Parataxis	0.12	0.13
Clausal complement (ccomp)	0.06	0.08
Clausal subject (csubj)	0.013	0.012
Clausal subject – passive (csubj:pass)	0.0016	0.0019
Adverbial modifier (advmod)	0.0006	0.0003

Таблица 8. Относительные частоты связей внутри глагольных групп

Можно проследить изменения в синтаксических структурах на следующем примере:

(4) a. *Анвар даже пробовал выговорить доблестному злодею помилование, но озлобившиеся министры были непреклонны, и наутро убийцу повесили на дереве. Дамы из гарема, так горячо любившие своего Черкеса, пришли посмотреть на его казнь, горько плакали и посылали ему воздушные поцелуи.*

b. *Когда эфенди узнал о том, что случилось, он просил министров не быть слишком жестокими к его другу. Но министры его даже слушать не стали. Утром Гасана убили. Женщины во дворце плакали.*

Видно, что была упразднена сочинительная связь в первом предложении в пользу нескольких простых предложений. Также исключаются причастия и адвербиальные модификаторы (так горячо любившие), при этом замена не осуществляется. С разделением сложных предложений снижается также глубина глагольных и именных групп.

3.4 Статистическое тестирование и моделирование зависимости между классом текста и его характеристиками

Для изучений возможных зависимостей между классом текста (оригинал или адаптированный) был применен коэффициент ранговой корреляции Кендалла. В результате исследования было выяснено, что наибольшую отрицательную корреляцию с классом текста имеют такие метрики, как процент слов, входящих в лексические минимумы для ТРКИ, некоторые лексические списки, а также значение формулы Оборневой. Наибольшую же положительную корреляцию имеют такие признаки, как некоторые формулы удобочитаемости, относительная частота причастий и деепричастий, максимальная и средняя глубина глагольной группы, количество устаревших слов, слов в творительном падеже и некоторые типы связей внутри глагольной группы (advmod, obj).

Для дальнейшего исследования зависимости между признаками и классом текста была построена логистическая регрессия. Для построения данной модели использовалась библиотека sklearn [27]. Перед подачей в модель тексты были перемешаны, размер тестового множества составил 0.2 от всей выборки. Также перед подачей в модель значения признаков были масштабированы от 0 до 1. Регрессия использовалась с параметрами по умолчанию, для оптимизации был выбран метод покоординатного спуска (liblinear solver). F1-мера классификации составила 71 для оригинальных текстов и 70 для класса адаптированных текстов. Это позволяет говорить о способности признаков объяснять класс, к которому принадлежит текст.

При изучении признаков, имеющих наибольшую значимость в модели, было обнаружено, что в решении определения текста к классу адаптированных текстов большую роль снова играют такие признаки, как количество слов из лексических минимумов и некоторых списков лексики. Кроме того, влияние оказывает количество фамилий в тексте, количество существительных. На принадлежность к классу оригинальных текстов указывают такие признаки, как процент длинных слов (т.е. слов длиннее 3 слогов), некоторые формулы удобочитаемости (формула Флеша в адаптации Оборневой, Индекс Dale-Chale и SMOG), количество сочинительных союзов, а также максимальная глубина именных и глагольных групп.

4 Выводы

Изучив результаты анализа характеристик оригинальных и адаптированных текстов, можно прийти к нескольким выводам. Во-первых, общее сокращение объемов текстов, а также сокращение длин предложений и их сложности (количества пунктуации, глубины глагольных и именных групп) свидетельствует о том, что одной из основных стратегий упрощения является саммаризация. При этом она происходит не только на уровне удаления целых отрывков произведения, но и на уровне предложений. Кроме этого, количественное исследование позволило подтвердить применение ряда описанных в литературе стратегий адаптации на морфологическом и синтаксическом уровне, в частности, замены деепричастий и причастий, замены сложных предложений несколькими простыми и пр.

Наиболее заметным на проанализированном материале оказывается редукция лексической сложности в процессе адаптации, о которой можно судить исходя из снижения процента редких слов и слов, выходящих за пределы лексических минимумов. При этом можно заметить, что некоторые стратегии упрощения, обнаруженные в других исследованиях, например, замена аббревиатур, очень мало представлены в исследуемом корпусе, вероятно, из-за специфики домена.

Другим интересным наблюдением является то, что при адаптации на лексическом уровне в упрощенных текстах заметно повышается количество слов из частотных списков. Это свидетельствует о том, что подобные списки (так же, как и лексические минимумы) можно использовать для автоматического лексического упрощения текстов. Использование частотных списков в этой задаче является распространенной практикой [18], которая, однако, на материале русского языка еще не применялось.

Для дальнейшего изучения стратегий упрощения необходимо будет сопоставить предложения из оригинальных текстов с предложениями из адаптированных, чтобы изучить, как именно происходят замены и/или удаления отдельных слов и фрагментов текста. Кроме этого, для создания дополнительных материалов для изучения стратегий упрощения, применяемых экспертами, будут привлечены эксперты-преподаватели РКИ.

Важным результатом данного исследования стал параллельный датасет RuAdapt, который может быть использован не только для изучения стратегий упрощения русского языка, подобного проведенному в настоящем исследовании, но и для создания либо дообучения систем автоматического упрощения текста. Поскольку в настоящее время задача автоматического упрощения часто рассматривается как монолингвальный машинный перевод [38], параллельные датасеты, где “обычным” сегментам соответствуют их упрощенные варианты, оказываются необходимы для обучения нейронных сетей для перевода. Кроме того, данные о стратегиях упрощения, полученные на основе подобных параллельных датасетов, можно применять для улучшения методов оценки результатов автоматического упрощения.

Понадобятся дополнительные исследования, чтобы сказать, насколько RuAdapt может улучшить качество нейронных моделей для упрощения, будучи использован в паре с другими датасетами, не относящимися к домену художественной литературы. Большую часть датасета составляют классические литературные произведения, актуальные для русского читателя по сей день. К тому же тексты, применяемые в учебных целях и упрощенные для читателей, осваивающих язык, нередко становятся частью параллельных датасетов для упрощения [6][8]. Это позволяет предположить возможность успешного использования RuAdapt с датасетами, содержащими общую лексику (например, такими, как датасет соревнования RuSimpleSentEval).

Благодарности

Работа выполнена с использованием средств государственного бюджета по госзаданию на 2020–2024 годы (проект FZNM-2020-0005).

References

[1] Andryshina N.P., Kozlova T.V. Lexical minimum of Russian as a foreign language. Level A1. Common language (4th ed.). – St. Petersburg, Zlatoust. 80 p. – 2012.

[2] Andryshina N.P., Kozlova T.V. Lexical minimum of Russian as a foreign language. Level A2.Common language (5th ed.). – St. Petersburg, Zlatoust. 116 p. – 2015.

[3] Andryshina N.P. (ed.) Lexical minimum of Russian as a foreign language. Level B1. Common language (9th ed.). – St. Petersburg, Zlatoust. 200 p. – 2017 (a).

[4] Andryshina N.P. (ed.), Lexical minimum of Russian as a foreign language. Level B2. Common language (7th ed.) – St. Petersburg, Zlatoust. 164 p. – 2017 (b).

[5] Andryshina N.P. (ed.). Lexical minimum of Russian as a foreign language. Level C1. Common language. – St. Petersburg, Zlatoust. 201 p. – 2018.

[6] Arfé B., Oakhill J., Pianta E. The text simplification in TERENCE //Methodologies and Intelligent Systems for Technology Enhanced Learning. – Springer, Cham, 2014. – P. 165-172.

[7] Baranova Yu. N., Elipasheva T. S. Creating an informational resource for Russian learner text analysis. [Sozdanie vspomogatel'nogo informacionnogo resursa dlya analiza uchebnyh tekstov na russkom yazyke.] // Chelovek v informacionnom prostranstve, Yaroslavl'. – 2014. – P. 232-246.
[8] Brouwers L. et al. Syntactic sentence simplification for French //Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR). – 2014. – P. 47-56.
[9] Brygina A.V. Linguistic principles of fiction text adaptation [Lingvisticheskie principy adaptirovaniya hudozhestvennogo teksta] // Ph.D. dissertation synopsis, RUDN university, Russia. – 2005. Available at: https://search.rsl.ru/ru/record/01003298898
[10] Chen X., Meurers D. Characterizing text difficulty with word frequencies //Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. – 2016. – P. 84-94.
[11] Coleman M., Liau T. L. A computer readability formula designed for machine scoring //Journal of Applied Psychology. – 1975. – Vol. 60. – №. 2. – P. 283.
[12] Crossley S. A., Yang H. S., McNamara D. S. What's so Simple about Simplified Texts? A Computational and Psycholinguistic Investigation of Text Comprehension and Text Processing //Reading in a Foreign Language. – 2014. – Vol. 26. – №. 1. – P. 92-113.
[13] Dale E., Chall J. S. A formula for predicting readability: Instructions //Educational research bulletin. – 1948. – P. 37-54.
[14] Gala N., Tack A., Javourey-Drevet L., Francois T., Ziegler J.C. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers //Language Resources and Evaluation for Language Technologies (LREC). – 2020. – P. 1353-1361.
[15] Karpov N., Sibirtseva V., Bogdanov D., Dmitrieva A., Elian E., Kleshnin E., Markiva E., Teplukhina T., Violentova L. Development of modern electronic textbook of Russian as a foreign language: content and technology //Higher School of Economics Research Paper No. WP BRP. – 2012. – T. 6. Available at: https://ideas.repec.org/p/hig/wpaper/06hum2012.html.
[16] Karpov N., Baranova J., Vitugin F. Single-sentence readability prediction in Russian //International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2014. – P. 91-100.
[17] Kauchak D. Improving text simplification language modeling using unsimplified text data //Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers). – 2013. – P. 1537-1546.
[18] Keskisärkkä R. Automatic text simplification via synonym replacement. – 2012. Ph.D. dissertation. Linköping University, Sweden, available at: https://www.diva-portal.org/smash/get/diva2:560901/FULLTEXT01.pdf
[19] Kilgarriff A. Simple maths for keywords //Proceedings of the Corpus Linguistics Conference. Liverpool, UK. – 2009. Available at: http://ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc
[20] Kincaid J. P. et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. – Naval Technical Training Command Millington TN Research Branch, 1975.
[21] Laposhina A. N., Veselovskaya T. S., Lebedeva M. U., Kupreshchenko O. F. Automated Text Readability Assessment For Russian Second Language Learners // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". – Issue 17 (24). – 2018. Available at: http://www.dialog-21.ru/media/4312/laposhina%D0%B0n.pdf
[22] Lyashevskaya O.N., Sharov S.A. Modern Russian frequency dictionary (based on the data from the Russian National Corpus) [Chastotnyj slovar' sovremennogo russkogo yazyka (na materialah Nacionalnogo korpusa russkogo yazyka)] // Azbukovnik, Moscow. – 2009.
[23] Mc Laughlin G. H. SMOG grading-a new readability formula //Journal of reading. – 1969. – Vol. 12. – №. 8. – P. 639-646.
[24] Nation I. How large a vocabulary is needed for reading and listening? //Canadian modern language review. – 2006. – Vol. 63. – №. 1. – P. 59-82.
[25] Oborneva I. V. Automatic evaluation of text perception quality. [Avtomatizaciya ocenki kachestva vospriyatiya teksta] // Vestnik Moskovskogo gorodskogo pedagogicheskogo universiteta. Seriya: Informatika i informatizaciya obrazovaniya, (5). – 2005. – P. 86-91.
[26] Pavlick E., Callison-Burch C. Simple PPDB: A paraphrase database for simplification //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2016. – P. 143-148.

[27] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weis, R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine learning in Python // The Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825-2830.
[28] Qian D. D. Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective //Language learning. – 2002. – Vol. 52. – №. 3. – P. 513-536.
[29] Reynolds R. 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. – 2016. – P. 289-300.
[30] Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine //MLMTA. – 2003. – Vol. 2003. – P. 273.
[31] Sennrich R., Volk M. MT-based sentence alignment for OCR-generated parallel texts. – 2010.
[32] Smith E. A., Senter R. J. Automated readability index //AMRL-TR. Aerospace Medical Research Laboratories (US). – 1967. – P. 1-14.
[33] Sharoff S. K. S., Hartley A. Seeking needles in the web haystack: Finding texts suitable for language learners //8th Teaching and Language Corpora Conference. TaLC-8. – 2008.
[34] Sibirtseva V. G., Karpov N.V. Automatic adaptation of the texts for electronic textbooks. Problems and perspectives (on an example of Russian). [Avtomaticheskaya adaptaciya tekstov dlya elektronnyh uchebnikov. Problemy i perspektivy (na primere russkogo yazyka)] // Nová rusistika. – Vol. VII, číslo 1 – 2014. – P. 19-33.
[35] Siddharthan A. A survey of research on text simplification //ITL-International Journal of Applied Linguistics. – 2014. – Vol. 165. – №. 2. – P. 259-298.
[36] Štajner S. et al. Sentence alignment methods for improving text simplification systems //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2017. – P. 97-102.
[37] Flesch R. A new readability yardstick //Journal of applied psychology. – 1948. – T. 32. – №. 3. – P. 221.
[38] Xu W., Callison-Burch C., Napoles C. Problems in current text simplification research: New data can help //Transactions of the Association for Computational Linguistics. – 2015. – Vol. 3. – P. 283-297.

Using RuGPT3-XL Model for RuNormAS competition

Anton Emelyanov[1,2] **Oleh Shliazhko**[1] **Nadezhda Katricheva**[1] **Tatiana Shavrina**[1,3,4]
login-const@mail.ru, olehshliazhko@gmail.com
n.katricheva@gmail.com, rybolos@gmail.com
[1]SberDevices, Sberbank, Moscow, Russia
[2]Moscow Institute of Physics and Technology, Moscow, Russia
[3]National Research University Higher School of Economics, Moscow, Russia
[4]ANO «AI Research Institute», Moscow, Russia

Abstract

The paper presents a fine-tuning methodology of the RuGPT3-XL (Generative Pretrained Transformer-3 for Russian) language model for the normalization of text spans task. The solution is presented in a competition for two tasks: Normalization of Named Entities (Named entities) and Normalization of a wider class of text spans, including the normalization of different parts of speech (Generic spans).
The best solution has achieved 0.9645 accuracy on the Generic spans task and 0.9575 on the Named entities task.
The presented solutions are in the public domain at https://github.com/RussianNLP/RuNormAS-solution

Keywords: text normalization, text generation, evaluation track, ruGPT-3, generative pretrained transformer
DOI: 10.28995/2075-7182-2021-20-204-212

Использование RuGPT3-XL модели для соревнования RuNormAS

Антон Емельянов[1,2] (login-const@mail.ru) Олег Шляжко[1] (olehshliazhko@gmail.com)
Надежда Катричева[1] (n.katricheva@gmail.com) Татьяна Шаврина[1,3,4] (rybolos@gmail.com)

[1]SberDevices, Сбербанк, Москва, Россия
[2]Московский физико-технический институт, Москва, Россия
[3]НИУ «Высшая Школа Экономики», Москва, Россия
[4]АНО «Институт Искусственного Интеллекта», Москва, Россия

Аннотация

В статье представлена методология дообучения языковой модели RuGPT3-XL (Generative Pretrained Transformer-3 для русского языка) для задачи нормализации спанов текста. Решение представлено на конкурсе по двум задачам: Нормализация именованных сущностей (Named entity) и Нормализация более широкого класса фрагментов текста, включая нормализацию различных частей речи (Generic spans).
Лучшее решение достигло точности 0.9645 для задачи нормализации фрагментов текста и 0.9575 для именованных сущностей.
Представляемые решения находятся в открытом доступе по адресу https://github.com/RussianNLP/RuNormAS-solution
Ключевые слова: нормализация текстов, генерация текстов, ruGPT-3, generative pretrained transformer

1 Introduction

The task of normalization is indispensable in Natural Language Processing (NLP) because it allows both to obtain a connection between the wordforms of the same paradigm and to reduce vocabulary size while preserving lexical meaning. Text classification, clusterization, topic modeling, style detection, and many more NLP tasks depend on normalization as a basic stage in the text processing pipeline. Regarding an isolating, fusional or agglutinative morphology type, normalization comes in two basic

wordform procedures: stemming or lemmatization. As a more simplistic approach, stemming only chops word endings from the stem, and thus it often attributes the same stem to cognates or different stems to the same lexeme. Lemmatization, in contrast, aims to bring tokens to lexemes. There are other types of normalization, too, such as expanding contractions and abbreviations, but in this paper, we understand normalization differently due to the entities it is applied to. To normalize a named entity or a phrase means to reduce it to its so-called «initial» form representing the semantic core which stays the same no matter what inflections its constituent parts may adopt to provide the syntactic integrity of a sentence. Most of the time it includes lemmatization, like in the case of the named entity (and noun phrase) "группы компаний ЛУКОЙЛ", which is normalized into "группа компаний ЛУКОЙЛ": the head of the noun phrase, "группа компаний ЛУКОЙЛ", becomes a lexeme after normalization. Proper nouns as parts of named entities can be normalized and plural at the same time, like in the case of "Сердца России", and there are other proper nouns which remain inflected and should not be changed through normalization.

Normalization methods include the usage of lexical databases, where word forms are linked to their lexemes. The result improves if part-of-speech (POS) tags are attributed to the word forms. Another common approach is rule-based, and of course words, as well as named entities and phrases, can be normalized using Neural Networks (NNs) (not only through using NNs for POS-tagging).

This paper is structured as follows: in section 2, we present the already existing research works related to the topic under discussion; section 3 gives a general overview of the competition; section 4 is devoted to our solution of the RuNormAS competition; section 5 provides error analysis, and the paper is concluded in section 6.

2 Previous Work

The English language was traditionally the first to undergo normalization algorithms, in particular, became the object for the first stemmer algorithm [1]. The analytical morphological structure was the best suited for this type of algorithms (for example, [2]), which, together with the growing needs of information retrieval, pushed their development — this happened, in particular, in the works [3], [4]. Nevertheless, it was the fusional and agglutinative languages, with their more productive morphology, that pushed normalization technologies to new levels and made them a subject of competition. Thus, the CONLL competition held in 2016-2018 [5,6, 7] set the task of complete grammatical annotation, from raw text to syntax, which comprised lemmatization for 103 languages, including "surprise languages" in the private test set. For the Russian language, the quality of word inflexion in context achieved 94.4% accuracy.

As for the Russian language separately, normalization technologies are also actively developing for it as a language with a developed morphology. The needs for information retrieval [9] prompted the use of the rich heritage of morphological description [8].

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 rare words [10]. The participants achieved 98.1% accuracy on lemmatization, the test set being not very large. At the MorphoRuEval-2017 shared task [11], a 96.91% accuracy score in lemmatization was achieved on a balanced set of data from various sources (news, social networks, fiction, etc.). And in the GramEval-2020 shared task [12] the track became even more complicated since data from social media, poetry and historical texts of the 17th century were added to the test sample: the best overall lemmatization score being 98% on fiction texts, 98.2% on the news, 95.3% on poetry, 96% on social media, 93% on wiki and 78.3% on historical texts. It became manifest that it is technically possible for the Russian language to pose more complex challenges, especially for notoriously "difficult-to-process" groups of words and lexical categories.

3 Dialog Evaluation 2021 Track

Within the framework of the RuNormAS (Russian Normalization of Annotated Spans) competition [13], the normalization problem is proposed — bringing a part of the text (a named entity, a phrase) to its

normal (initial) form. The main part of the task is to correctly normalize the words from the group that need normalization without changing the other ones (dependent, etc.) while using the given context to the benefit of this task. The latter is especially important since the initial form for many words can be determined only in context — for example, the the word "Иванова", depending on the surrounding context, can have either the normal form "Иванова"or "Иванов".
The competition offers two tracks:

1. Normalizing Named Entities
2. Normalization of a wider class of text spans, including the normalization of different parts of speech.

The data for the first track were collected from the articles of the «ВЗГЛЯД»newspaper, for the second one — from the documents of the Ministry of Economic Development. Both samples were labeled manually.

The quality metric for the task is the percentage of exact matches between the normalization result and the reference.

3.1 Dataset

Both tasks have the same data format. The `text_and_ann` folder contains files with texts (`.txt`) and files with span markup (`.ann`). In the file with the markup, the indices of the beginning and end of the entity in the text are written on each line. If the entity has breaks, then one line is written with the start and end indices for each chunk (and the chunks may be unordered). For example, if an entity has two breaking chunks, then the annotations on the corresponding line will contain `start1 end1 start2 end2` or `start2 end2 start1 end1`. In the folder norm, on each line, there is the result of normalization of the corresponding span. The match is made by the filename up to a dot. Also, for best model additional data was used. We add the "lenta news" dataset to the train data. This is a corpus of Russian News for the year 2019. That corpus was annotated automatically and is a part of Taiga corpus [14].

4 Approach

4.1 Baseline

The competition presents a baseline obtained using normalization tools from the Natasha library[1]. This solution is completely rule-based.

4.2 Neural Language modeling

The idea of finetuning a pretrained Language Model (LM) is at the core of our approach. All the experiments were carried out using RuGPT3XL[2]. The main difference is connected with data preparation for the RuGPT3XL LM finetuning procedure and model inference strategy. We do not separate data for two tasks and train one model at each approach on the whole set of train data.
The main algorithm for making predictions consists of three steps (all of them are described below):

1. Prepare data for LM using one of data preparation approaches;
2. Make predictions with LM using one of the inference strategies;
3. Apply the post-processing pipeline;

Each approach differs from the other one only in a specific template for generation, which is fed to the input of the LM. We tested the following approaches of data preparation (for the first step):

1. Model0 — only left context LM;
2. Model1 — only left context LM with `<start>` special token;
3. Model2 — left and right contexts LM with `<start>` and `<end>` special tokens;
4. Model3 — left and right contexts LM with `<start>` and `<end>` special tokens and additional training data;

For each approach, we apply two inference strategies:

[1] `https://github.com/natasha/natasha`
[2] `https://huggingface.co/sberbank-ai/rugpt3xl`

- **“argmax“ inference strategy** is the decoding strategy of LM. We select the next token by applying ‘argmax‘ operation over probability distribution that is produced by LM on each decoding step.
- **“beam search“ inference strategy** is the standard beam search algorithm with the number of beams equal to 10.

For each approach, we apply the same post-processing pipeline.

4.2.1 Post-processing pipeline

The post-processing pipeline should correct errors that occur while generating with LM (after the second step). We have categorized the errors as follows:

1. extra special tokens — model generates extra special tokens that should be removed;
2. letters case errors — model generates words in different cases;
3. extra or removed punctuation — model generate additional punctuation marks or remove some punctuation;
4. different word count in annotation and prediction;
5. symbol intersection error — this error occurs if the following condition is met:
   $\frac{|set(annotation) \cap set(generated)|}{|set(generated)|} < 0.6$
   here, the annotation and generation are strings; the 0.6 parameter is selected with some greed search on a subsample of the training data.

For steps 4-5 of this pipeline, we get prediction from the baseline model if errors occurred. Other steps of post-processing are also implemented in our repository.

4.2.2 Model0 — only left context LM

For each line in files with span markup (`.ann`), we find a substring in the text that should be normalized. For example, we have a text in the file of the test set with the name “723362“ at Named Entities task:

“Пушков назвал выход Греции из еврозоны «сильнейшим ударом по ЕС» Греция — ключ к будущему ЕС, ее выход из еврозоны станет сильнейшим ударом по ЕС за всю его историю, заявил глава комитета Госдумы по международным делам Алексей Пушков. «Что бы ни говорили в Берлине, выход Греции из еврозоны станет сильнейшим ударом по ЕС за всю его историю. Сегодня Греция — ключ к будущему ЕС», — написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле допускают выход Греции из еврозоны при необходимости. Позже заместитель официального представителя кабмина ФРГ Георг Штрайтер заявил, что позиция Германии по вопросу выхода Греции из еврозоны не изменилась. не изменилась“.

For the 12th line, in the annotation file we extract the subtext that needs to be normalized: “Вольфганг Шойбле“. For training LM, we construct a training record with the help of the following template:

```
<s>{left_context}{to_norm}<answer>{norm}</s>
```

Here the `<s>` token denotes beginning of text; `left_context` denotes all text before subtext that should be normalized (to_norm); the `<answer>` token separates input text prefix and answer that LM should learn; norm is the normalized text; and `</s>` token denotes the end of text. For our previous example, we have the following training record:

<s>Пушков назвал выход Греции из еврозоны «сильнейшим ударом по ЕС» Греция — ключ к будущему ЕС, ее выход из еврозоны станет сильнейшим ударом по ЕС за всю его историю, заявил глава комитета Госдумы по международным делам Алексей Пушков. «Что бы ни говорили в Берлине, выход Греции из еврозоны станет сильнейшим

ударом по ЕС за всю его историю. Сегодня Греция — ключ к будущему ЕС», — написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле<answer>Вольфганг Шойбле</s>

For the inference phase, we construct the following template that is passed to LM:

<s>{left_context}{to_norm}<answer>.

After making predictions, we correct the output with the post-processing pipeline that is described later.

4.2.3 Model1 — only left context LM with the <start> special token

We use the following template in this data preparation approach:

<s>{left_context}<start>{to_norm}<answer>{norm}</s>

The main difference from the previous template is the token <start> which denotes the beginning of the subtext that should be normalized. For our previous example, we have the following training record:

<s>Пушков назвал выход Греции из еврозоны «сильнейшим ударом по ЕС» Греция — ключ к будущему ЕС, ее выход из еврозоны станет сильнейшим ударом по ЕС за всю его историю, заявил глава комитета Госдумы по международным делам Алексей Пушков. «Что бы ни говорили в Берлине, выход Греции из еврозоны станет сильнейшим ударом по ЕС за всю его историю. Сегодня Греция — ключ к будущему ЕС», — написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны<start>Вольфганг Шойбле<answer>Вольфганг Шойбле</s>

For inference phase, we construct the following template that is passed to LM:

<s>{left_context}<start>{to_norm}<answer>.

After prediction we correct output with post-processing pipeline that described later.

4.2.4 Model2 — left and right contexts LM with the <start> and <end> special tokens

We use the following template in this data preparation approach:

<s>{left_context}<start>{to_norm}<end>{right_context}<answer>{norm}</s>

Here, the <s> token denotes the beginning of the text; left_context denotes the text before the subtext that should be normalized (to_norm); the token <start> denotes the beginning of the subtext that should be normalized; the <end> token denotes the end of the subtext that should be normalized; right_context denotes the text after the subtext that should be normalized (to_norm); the <answer> token separates the input text prefix and the answer that LM should learn; norm is the normalized text; and the </s> token denotes the end of the text. For our previous example, we have the following training record:

<s>написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству

Model name	Generic spans	Named entities
Model0 + argmax + not_fixed	0.6953	0.7513
Model0 + argmax	0.7507	0.7891
Baseline	0.7732	0.8881
Model0 + beam search	0.8454	0.8828
Model1 + argmax	0.9059	0.9306
Model1 + beam search	0.9483	0.9455
Model2 + beam search	0.9592	0.9570
Model3 + beam search + not_fixed	0.9636	0.9522
Model3 + beam search	**0.9645**	**0.9575**

Table 1: Evaluation results

Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны <start> Вольфганг Шойбле <end> допускают выход Греции из еврозоны при необходимости. Позже заместитель официального представителя кабмина ФРГ Георг Штрайтер заявил, что позиция Германии по вопросу выхода Греции из еврозоны не изменилась.<answer>Вольфганг Шойбле</s>

`left_context` and `right_context` are the texts that are limited to 40 words taken before `to_norm` and 40 words after `to_norm`. The parameter of window 40 is selected with some greed search on a subsample of training data.
For the inference phase, we construct the following template that is passed to LM:

```
<s>{left_context}<start>{to_norm}<end>{right_context}<answer>.
```

After prediction, we correct the output with the post-processing pipeline described below.

4.2.5 Model3 — left and right contexts LM with the <start> and <end> special tokens and additional training data

The main difference from the previous approach is using additional training data. We add the "lenta news" corpus with normalization markup and finetune the model on this corpus joint with the training data. After that, we finetune the model only on the training data. The data for LM finetuning was prepared as described in the previous section.

4.2.6 Training details

Each model was trained on 16 GPU with distributed training for around 12 hours. We use the Adam optimizer from [18] with the decoupled weight decay regularization 1e-2 [19]. We use a constant learning rate, 0.000015 on 20000 train iterations with fp16 precision and deepspeed code optimizations[20]. The final perplexity on all models is around 1.0002-1.0005.

5 Error Analysis and Results

5.1 Results

The results of our experiments on test set are presented in Table 1. The best result (*Accuracy Generic spans = 0.9645* and *Accuracy Named entities = 0.9575*) was obtained for "Model3 — left and right contexts LM with <start> and <end> special tokens and additional training data" approach with the beam search inference strategy.

The fourth approach "Model3 — left and right contexts LM with <start> and <end> special tokens and additional training data" with the beam search inference strategy obtains the best accuracy for the RuGPT3XL model in this competition. Also, we can see the difference provided by the post-processing pipeline on "Model0" and "Model3". For the last model, the impact is minor because the LM model has very strong results and sees more data.

5.2 Error analysis

5.2.1 Evaluation errors

Here, we describe errors that are connected with the incorrect data in the evaluation set and markup. We have categorized errors into the following classes:

1. word count errors — these errors denote different counts of words in gold prediction and annotation. For example: "Костромская область"and "областях", here our best model predicted "области".
2. titled errors — these errors denote difference between word cases in gold prediction and annotation. For example: "Генпрокуратура Украины"and "генпрокуратура Украины", here our best model predicted "генпрокуратура Украины".
3. symbol errors — these errors denote the difference between some symbols in gold prediction and annotation. For example: "город Антрацит"and "город Антрацит", here our best model predicted "город Антрацит".
4. punctuation errors — these errors denote the difference between the punctuation in gold prediction and annotation. For example: "ООО «Первая топливная компания»"and "ООО Первая топливная компания», here our best model predicted "ООО Первая топливная компания»".
5. word start errors — these errors denote the difference between the starting symbols in gold prediction and annotation. For example: "расти"and "будет расти", here our best model predicted "будет расти". Also these errors denote encoding mismatch or truncated markup.

If these errors are not taken into account, then the model obtained **0.9767** accuracy on the Generic spans task and **0.9810** accuracy on the Named entities task.

5.2.2 Model errors

Here we describe model errors. We divide the errors into categories:

1. word count errors. An example of prediction and gold prediction: "дорога Артемовск-Луганское-Дебальцево"and "дорога Артемовск-Луганское-Лозовое-Дебальцево".
2. word position errors. An example of prediction and gold prediction: "Киевская городская государственная администрация" and "Киевская государственная городская администрация".
3. word ending errors. An example of prediction and gold prediction: "Верховная рада"and "Верховая рада".
4. word case errors. An example of prediction and gold prediction: "«взрослый» Арктический Совет"and "«взрослый» Арктический совет".
5. consistency errors. An example of prediction and gold prediction: "Южно-Русский газоконденсатный месторождение"and "Южно-Русское газоконденсатное месторождение".
6. errors with foreign words. An example of prediction and gold prediction: "Укрнафта"and "Укртанснафта".
7. errors with POS tags mismatches. An example of prediction and gold prediction: "Новороссийский"and "Новороссийск".

Some of the described errors can be avoided by finetuning the model with more extra data.

6 Conclusion and Future Work

We present the results of our participation in the DE2021: RuNormAS (Russian Normalization of Annotated Spans) task. The implemented methods in both subtracks are based on RuGPT3XL LM. As future work, we plan to finetune RuGPT3XL LM on more extra data.

The best model was presented in the paper is available open-source. We hope that our developments will be useful to the community since all the presented prototypes are easily portable to new domains and tasks.

References

[1] Lovins, Julie Beth (1968). "Development of a Stemming Algorithm" (PDF). Mechanical Translation and Computational Linguistics. 11: 22–31.
[2] Dawson, J. L. (1974); Suffix Removal for Word Conflation, Bulletin of the Association for Literary and Linguistic Computing, 2(3): 33–46.
[3] Frakes, W. B. (1984); Term Conflation for Information Retrieval, Cambridge University Press.
[4] Frakes, W. B. (1992); Stemming algorithms, Information retrieval: data structures and algorithms, Upper Saddle River, NJ: Prentice-Hall, Inc.
[5] Cotterell R. et al. The SIGMORPHON 2016 shared task—morphological reinflection //Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. – 2016. – . 10-22.
[6] Cotterell R. et al. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages //arXiv preprint arXiv:1706.09031. – 2017.
[7] Cotterell R., Kirov Ch., Sylak-Glassman J., et al. (2018) The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In Proceedings of CoNLL–SIGMORPHON 2018.
[8] Zaliznyak, A. A. (1977) Grammatical Dictionary of Russian [Grammaticheskij slovar' russkogo yazyka]. Moscow.
[9] 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.
[10] Lyashevskaya, Olga, Irina Astaf'eva, Anastasia Bonch-Osmolovskaya, Anastasia Garejshina, Julia Grishina, Vadim D'jachkov, Maxim Ionov, Anna Koroleva, Maxim Kudrinsky, Anna Lityagina, Elena Luchina, Eugenia Sidorova, SvetlanaToldova, Svetlana Savchuk, and Sergej Koval' (2010) NLP evaluation: Russianmorphological parsers [Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka]. In: Computational linguistics and in-tellectual technologies. Proceedings of International Workshop Dialogue'2010. Vol. 9 (16), 2010. Pp. 318–326
[11] Sorokin, A., Shavrina, T., Lyashevskaya, O., Bocharov, B., Alexeeva, S., Droganova, K., ... Granovsky, D. (2017). MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian.
[12] Lyashevskaya O., Shavrina T., Trofimov I., Vlasova N. A. GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing, in: Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 17 июня — 20 июня 2020 г.) / Под общ. ред.: В. Селегей. Вып. 19(26). М. : Изд-во РГГУ, 2020. P. 553-569.
[13] Zolotukhin, Denis and Smurov, Ivan (2021). RuNormAS-2021: a Shared Task on Russian Normalization of Annotated Spans. Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue".
[14] Shavrina, Tatiana and Shapovalova, Olga (2017). TO THE METHODOLOGY OF CORPUS CONSTRUCTION FOR MACHINE LEARNING:"TAIGA". SYNTAX TREE CORPUS AND PARSER. Corpus linguitics-2017. Pp 78.
[15] Ivanin, V. A., et al. "Rurebus-2020 shared task: Russian relation extraction for business." (2020).
[16] Starostin, A. S., et al. "FactRuEval 2016: evaluation of named entity recognition and fact extraction systems for Russian." (2016).
[17] Malykh, Kalaidin. "HEADLINE GENERATION SHARED TASK ON DIALOGUE'2019." Компьютерная лингвистика и интеллектуальные технологии. 2019.
[18] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[19] Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.
[20] Minjia Zhang, Yuxiong He. (2020) Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping. arXiv:2010.13369.

Oculomotor everyday communication: How to pick a good metric

Olga V. Fedorova
Interdisciplinary Scientific and Educational School of Moscow University
"Brain, Cognitive Systems, Artificial Intelligence", Moscow, Russia
olga.fedorova@msu.ru

Abstract

This paper contributes to the research field of bimodal linguistics that explores two modalities involved in everyday communication – vocal and kinetic. When exploring almost any scientific phenomenon, one addresses two opposite issues: individual differences, on the one hand, and general patterns, on the other. We have focused on the individual differences and proposed a "portrait" approach to communication. We are faced with a difficult task to find a good metric for analyzing oculomotor behavior of people in everyday communication. In previous papers, starting from [14], the authors were looking for oculomotor patterns, but their results depend critically on the metric used. In this paper, we compared the most common metrics and showed that individual differences have a much more serious weight than general patterns. We then identified four coefficients that determine these individual differences: k_{aside}, k_{vip}, k_{chain}, and dur_{75}. By comparing these Core Oculomotor Portraits, we were able to make these individual differences more clear. However, a fact is a fact: there are far more individual differences than general patterns between our Narrators behavior. The proposed coefficients, in our opinion, clearly show (and even explain and predict) the observed individual differences.

Keywords: bimodal communication; eye tracking; gaze; fixation; metric
DOI: 10.28995/2075-7182-2021-20-213-226

Окуломоторное повседневное общение: как выбрать хорошую метрику

О.В. Федорова
Междисциплинарная научно-образовательная школа Московского университета
«Мозг, когнитивные системы, искусственный интеллект», Москва, Россия
olga.fedorova@msu.ru

Аннотация

Данная работа вносит вклад в исследовательскую область бимодальной лингвистики, в которой исследуются две модальности повседневной коммуникации – вокальная и кинетическая. Исследуя практически любой феномен, мы сталкиваемся с двумя противоположными явлениями: индивидуальными различиями, с одной стороны, и общими закономерностями, с другой. В данной работе мы сосредоточились на индивидуальных различиях и предложили «портретный» подход к коммуникации. Мы поставили сложную задачу найти хорошую метрику для анализа окуломоторного поведения людей в повседневном общении. В предыдущих работах, начиная с [14], авторы искали окуломоторные паттерны, но их результаты критическим образом зависели от используемой метрики. В данной работе мы сравнили наиболее распространенные метрики и показали, что индивидуальные различия имеют гораздо более серьезный вес, чем общие закономерности. Затем мы ввели четыре коэффициента, определяющих эти индивидуальные различия: k_{aside}, k_{vip}, k_{chain} и dur_{75}. Сравнив базовые окуломоторные портреты, мы смогли сделать наблюдаемые индивидуальные различия более ясными. Однако факт остается фактом: между поведением испытуемых гораздо больше индивидуальных различий, чем общих паттернов. Предложенные коэффициенты, на наш взгляд, ясно показывают (и даже объясняют и предсказывают) наблюдаемые индивидуальные различия.

Ключевые слова: бимодальная коммуникация, регистрация движений глаз; взгляд; фиксация; метрика

1 Introduction. Bimodal communication: Oculomotor component

This paper contributes to the research field of bimodal linguistics. Bimodal linguistics explores two modalities involved in everyday communication – vocal and kinetic, see Fig. 1[1]. Vocal modality (from the perspective of an addresser; or auditory modality from the perspective of an addressee) consists of the segmental verbal structure and non-segmental prosody. Kinetic modality (from the perspective of an addresser; or visual modality from the perspective of an addressee) includes all kinds of movements – with eyes, face, head, hands, etc. Since only two modalities are included into consideration in contemporary research (but cf. [23] on the touch modality), we consider the widely circulated notion of multimodality an overstatement and prefer the notion of bimodality (for multimodality see [15], [22], [8], [24], [5], [9], [11], inter alia).

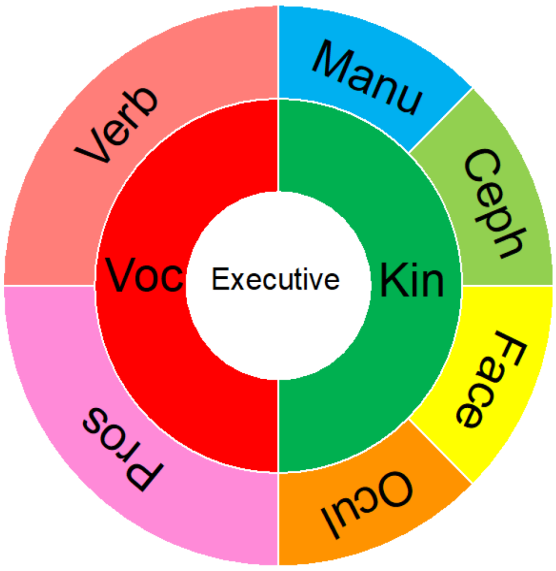


Figure 1: Bimodal communication from the addresser's perspective

In this paper, we consider "oculomotor" component of the kinetic modality, i.e. eye movements[2] ([14], [1], [21], [12], [10], [13], [3], [19]). Studying eye movements provides unique insights into what the participant found interesting or important, that is, what drew his/her attention, and provide a clue as to how he/she perceived the scene he/she was viewing. (Note that although eye movements and the visual attention are closely related, the nature of this relationship is not yet fully understood; see e.g. [26]).

Our previous fine-grained qualitative and quantitative analyses of a 10-min fragment of communication between three interlocutors showed that the use of different metrics of oculomotor analysis – the number and duration of fixations or the number and duration of gazes[3] – gives fundamentally different results ([6]). In this paper we address individual differences and propose a "portrait" approach to the oculomotor component of bimodal communication, see section 3. We address this issue with the help of the bimodal corpus "Russian Pear Chats & Stories" ([16], https://multidiscourse.ru), see section 2.

[1] The "Executive" is the central controlling component of the system (cf. similar executive components in theoretical models such as in [2] or [20]).

[2] When we look at a scene our eyes move around continually, locating some definite points. Rapid movements of the eyes are known as saccades. Saccades normally take about 20-150 ms, depending on their amplitude. Little or no actual visual processing occurs during saccades. Between the saccades, our eyes remain relatively still during fixations for about 100-1500 ms.

[3] Gaze typically consists of several fixations within an area of interest (AOI) and may include some short saccades between these fixations. A fixation occurring outside the AOI marks the beginning of a different gaze. AOIs are defined by the researcher, not by the participant. For example, if we describe a person, it is possible to draw separate AOIs around his/her body, his/her face, and his/her hands, see below.

2 The corpus "Russian Pear Chat & Stories"

2.1 Recording set-up

For collecting the corpus the well known Pear Film (Chafe ed. 1980) is used. Each session involved four participants with fixed roles: three main interlocutors – the Narrator (N), the Commentator (C), and the Reteller (R) – and the Listener (L). At the very beginning N and C each watched the film, trying to memorize the plot as precisely as possible. Then the main stages began. First, N told the R about the plot of the film; this is a monologic stage – "*First Telling*". During the subsequent interactive stage – "*Conversation*" – C added details and corrected the N's story where necessary, and R checked her/his understanding of the plot, asking questions to both interlocutors. Then L joined the group and another monologic stage – "*Retelling*" – followed, during which R was retelling the plot of the film to L. Finally, L wrote down the content of the film.

2.2 Recording software

The participants' speech was recorded with the help of a six-channel recorder ZOOM H6 Handy Recorder (96 kHz / 24 bit). Three industrial video cameras JAI GO (100 frames per second and 1392x1000 pixels) recorded three participants, shooting individually from a frontal perspective. In addition, the camera GoPro Hero was used to record the whole scene.

In order to record eye gaze, two head-mounted eye trackers were used (Tobii Glasses II, 50 Hz and 1920x1080 pixels); N and R were wearing eyetrackers. The eye trackers provide two types of data: videofiles produced by an inbuilt scene camera and data files representing eye movements. The screenshots in Fig. 2 result from an overlay of videofiles from the scene camera and the gaze coordinates from the data files; the circles are generated by the eye trackers and indicate the targets of interlocutors' gaze.

a. From the N's eye tracker

b. From the R's eye tracker

Figure 2: Screenshots of video scenes from eye trackers

2.3 Participants and corpus size

The full corpus includes 40 recordings with 160 Russian native participants aged 18–36. Our subcorpus includes seven recordings (##04, 06, 16, 21, 22, 23, 24) with 28 Russian native participants, 9 men and 19 women, recruited from the Moscow population. The subcorpus consists of 2 hours 40 minutes of recording and about 50,000 words. The distribution of the recordings' duration by stages see in Fig. 3.

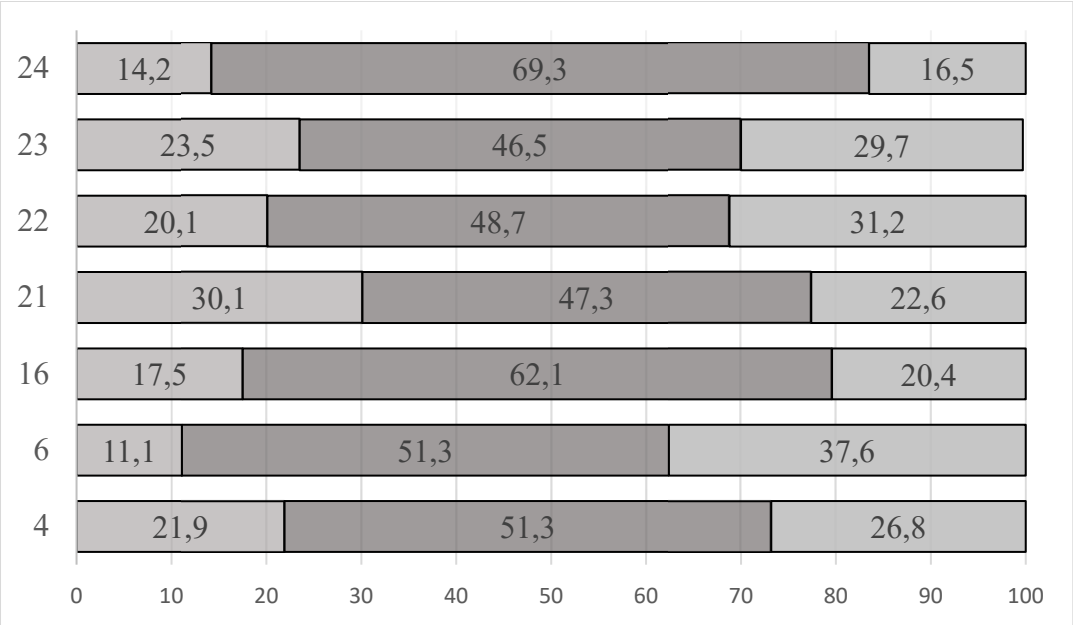


Figure 3: The distribution of the recordings' duration by stages ("*First Telling*" / "*Conversation*" / "*Retelling*"), in %

2.4 Annotations

The vocal annotation used in the project follows the principles previously developed for spoken Russian discourse (https://spokencorpora.ru; [18]). For the kinetic annotation see https://multidiscourse.ru/annotation/?en=1.

The oculomotor annotation scheme includes five tiers:

(1) a fixations number *-oFixation

(2) "Interlocutor", with five possible values:

- "N" (fixation on N)
- "R" (fixation on R)
- "C" (fixation on C)
- "L" (fixation on L)

"Surroundings" (=Sur, fixation on another object)

(3) "Locus", with four possible values:

- "Face" (fixation on the face of the participant)
- "Hands" (fixation on the hands of the participant)
- "Body" (fixation on the body of the participant)
- "Surroundings" (=Sur, fixation on another body part of the participant, e.g. legs)

(4) a gaze number *-oGaze

(5) "Gaze", with five possible values: "N", "R", "C", "L", "Surroundings".

The oculomotor annotation was carried out in MS Excel. With the help of Tobii Pro Glasses Analyzer software, we automatically extracted information about the time base of all fixations and then manually applied the five-tier annotation scheme described above.

3 How to pick a good metric: Oculomotor Portrait

Eye tracking is a relatively simple measure, but the tricky challenge is what to do with the data it provides. In this section, we present new data called "Oculomotor portraits" obtained by seven Narrators at the monological stages of our subcorpus, i.e. the "*First Telling*" and "*Retelling*" stages. Thus, the analysis was performed on the basis of 14 fragments with a total duration of 1 hour 15 minutes.

The most common oculomotor metrics include:

(1) Number:
- Number of fixations, overall
- Number of gazes, overall
- Number of fixations on each AOI
- Number of gazes on each AOI

(2) Duration:
- Duration of fixations (=dwell time), overall
- Duration of gazes (=dwell time), overall (=Duration of fixations, overall)
- Duration of fixations (=dwell time) on each AOI
- Duration of gazes (=dwell time) on each AOI (=Duration of fixations on each AOI)

(3) Mean duration:
- Mean fixation duration, overall
- Mean gaze duration, overall
- Mean fixation duration on each AOI
- Mean gaze duration on each AOI

(4) %:
- Fixation % (ratio, proportion of number) on each AOI
- Gaze % (ratio, proportion of number) on each AOI
- Fixation % (ratio, proportion of time) on each AOI
- Gaze % (ratio, proportion of time) on each AOI

(5) Rate:
- Fixation rate, overall (fixations / seconds)
- Gaze rate, overall (gazes / seconds)
- Fixation rate on each AOI (fixations / seconds)
- Gaze rate on each AOI (gazes / seconds)

(6) Scan path, i.e. the spatial arrangement of a sequence of fixations or gazes

(7) Heatmaps, i.e. visualizations which show the general distribution of gaze points[4]. Red, yellow, and green colors represent in descending order the amount of gaze points that were directed towards some parts of the image

(8) Time to the first fixation on each AOI

(9) The first fixation duration on each AOI

(10) Regressions. During reading, readers often move their eyes forward to process new information. However, not all eye movements take the eyes forward in the text. About 15% of eye movements move backwards to reprocess information ([25]).

In our study the first four types of metrics are used, see the Oculomotor Portrait for N 04 in Table 1. Number of fixations or gaze[5], as well as (mean) duration could reflect the importance of a particular AOI. We have calculated also the minimal and maximal durations (overall and on each AOI), as well as 25%, 50%, and 75% quantiles. Basic comparisons were made for 75% quantiles (italicized in Table 1). We called the presented data “Full Oculomotor Portrait” (the preliminary ideas on this topic see [17]).

In [7] we, based on the number and overall duration of fixations, found that the typical listener looks at the speaker with long fixations, broken by brief fixations to the surroundings, while the typical speaker alternates long fixations at the listener with brief fixations to the surroundings. However, let’s look at the Full Oculomotor Portraits for our seven Narrators more closely (for all seven portraits see Appendix). We can see that they are very different in all respects, that is, individual differences are very large. To be able to compare these data, we have introduced the following coefficients (highlighted in bold):

[4] Gaze points show what the participant is looking at. Our eye tracker collects data with a sampling rate of 50 Hz, thus we have 50 gaze points per second.

[5] Counting the number of gazes (i.e., successive fixations within the same AOI) is often considered more meaningful than counting the number of individual fixations.

Overall			
duration		1157.167	
R duration, ratio		849.006, 0.73, $\mathbf{k_{aside}}$ **0.3**	
Sur duration, ratio		267.584, 0.23	
		fixation	gaze
number		2190	554, $\mathbf{k_{chain}}$ **4**
mean duration, std		0.528, 0.664	2.089, 3.341
min, 25, 50, 75, max		0.06, 0.16, 0.28, ***0.6***, 10.477	0.06, 0.4, 0.979, 2.500, 26.974
R $\mathbf{k_{vip}}$ **3.8**	number, ratio	1048, 47.9	251, 45.3, $\mathbf{k_{chain}}$ **4.2**
	mean, std	0.81, 0.845	3.382, 4.461
	min, 25, 50, *75*, max	0.06, 0.24, 0.48, *1.14*, 10.477	0.1, 0.91, 1.86, 3.72, 26.974
Sur	number, ratio	1014, 46.3	260, 46.9, $\mathbf{k_{chain}}$ **3.9**
	mean, std	0.14, 0.216	1.018, 1.164
	min, 25, 50, *75*, max	0.06, 0.14, 0.2, *0.3*, 2.22	0.06, 0.275, 0.56, 1.26, 6.517
First telling			
duration		235.472	
R duration, ratio		144.547, 0.62, $\mathbf{k_{aside}}$ **0.6**	
Sur duration, ratio		90.465, 0.38	
		fixation	gaze
number		536	166, $\mathbf{k_{chain}}$ **3.2**
mean duration, std		0.439, 0.547	1.419, 1.22
min, 25, 50, 75, max		0.06, 0.14, 0.22, ***0.46***, 3.56	0.06, 0.4, 1.16, 2.155, 5.78
R $\mathbf{k_{vip}}$ **5**	number, ratio	161, 0.3	80, 0.48, $\mathbf{k_{chain}}$ **2**
	mean, std	1.321, 0.795	1.807, 1.188
	min, 25, 50, *75*, max	0.06, 0.24, 0.64, *1.4*, 3.56	0.1, 0.93, 1.499, 2.555, 5.78
Sur	number, ratio	372, 0.69	83, 0.5, $\mathbf{k_{chain}}$ **4.5**
	mean, std	0.243, 0.17	1.091, 1.15
	min, 25, 50, *75*, max	0.06, 0.134, 0.2, *0.28*, 1.18	0.06, 0.31, 0.6, 1.49, 5.337
Retelling			
duration		332.55	
R duration, ratio		313.421, 0.94, $\mathbf{k_{aside}}$ **0.1**	
Sur duration, ratio		19.129, 0.06	
		fixation	gaze
number		497	74, $\mathbf{k_{chain}}$ **6.7**
mean duration, std		0.669, 0.894	4.494, 6.916
min, 25, 50, 75, max		0.06, 0.2, 0.32, ***0.76***, 10.477	0.077, 0.345, 0.81, 6.078, 26.974
R $\mathbf{k_{vip}}$ **3.2**	number, ratio	418, 0.84	36, 0.49, $\mathbf{k_{chain}}$ **11.6**
	mean, std	0.75, 0.949	8.706, 8.001
	min, 25, 50, *75*, max	0.08, 0.22, 0.4, *0.9*, 10.477	0.14, 1.639, 6.577, 14.247, 26.974
Sur	number, ratio	79, 0.16	38, 0.51, $\mathbf{k_{chain}}$ **2.1**
	mean, std	0.242, 0.211	0.503, 0.455
	min, 25, 50, *75*, max	0.06, 0.12, 0.2, *0.28*, 1.26	0.077, 0.239, 0.37, 0.575, 2.54

Table 1: Full Oculomotor Portrait for N 04 (durations shown in seconds)

(1) k_{aside} denotes how often N looks away compared to his R's fixations or gazes; = (Sur's duration) / (R's duration).
(2) k_{vip} denotes how much R is more important for N compared to Sur; = (mean R's duration) / (mean Sur's duration).
(3) k_{chain} denotes how many fixations are included in one N's gaze; = (number of fixations) / (number of gazes).

Compare now the "Core Oculomotor Portraits" of our seven Narrators, including k_{aside}, k_{vip}, k_{chain}, and dur_{75}, i.e. mean durations of fixation for 75% quantiles, separately for the *"First Telling"* (Table 2) and the *"Retelling"* (Table 3) stages.

k / Ns	4	6	16	21	22	23	24
k_{aside}	0.6	1.2	0.8	0.3	3.8	0.4	0.2
k_{vip}	5	3.5	5.5	5.8	2	1.5	5.4
k_{chain}	3.2	2.9	3.9	2.7	6.5	5.7	3.4
dur_{75}	0.46	0.69	0.36	0.62	0.72	0.58	0.5

Table 2: Comparison of Core Oculomotor Portraits, the "*First Telling*" stage

k / Ns	4	6	16	21	22	23	24
k_{aside}	0.1	0.1	0.1	0.1	0	0.2	0
k_{vip}	3.2	2.8	3	4.1	3.6	1.6	6.5
k_{chain}	6.7	5.6	8.5	3	14.6	6.2	11.9
dur_{75}	0.76	0.7	0.74	1.22	1.45	0.64	1.61

Table 3: Comparison of Core Oculomotor Portraits, the "*Retelling*" stage

(1) k_{aside}
At the *"First Telling"* stage, we observe a classical continuum from 0.2 to 1.2. N22 (k_{aside}=3.8) distinguished herself from the others. At the *"Retelling"* stage, the coefficient is almost the same for all Ns.
(2) k_{vip}
As can be seen from the tables, R is always more important, the question is how much. Ns 04, 06, 16 and 21 have higher coefficients for the *"First Telling"* stage, while Ns 22 and 24, on the contrary, for the *"Retelling"* stage. N23 distinguished herself from the others by both the similarity of the coefficient k_{vip} for the *"First Telling"* and the *"Retelling"* stages and a small difference between R's and Sur's durations.
(3) k_{chain}
The values of the coefficient are distributed between 2.7 and 14.6. All Ns have higher coefficients for the *"Retelling"* stage, but for Ns 21 and 23 the difference is minimal. At the same time Ns 22 and 24 distinguished from the others by high values of k_{chain}.
(4) dur_{75}
All Ns have higher coefficients for the *"Retelling"* stage, but for N6 the difference is minimal. N 21 (dur_{75}=1.22), N 22 (dur_{75}=1.45), and N 24 (dur_{75}=1.61) have the duration for the *"Retelling"* stage more than 1 second.

What can we say about the differences between Ns based on the Core Oculomotor Portraits? We can assume that N 04 and N 16 behave the same way in terms of Core Oculomotor Portrait, for both the *"First Telling"* and the *"Retelling"* stages. N 6 is similar to this pair, but she has k_{aside} more than 1. Four other Ns are unique, each in his own way; N 22 is particularly unique.

4 Conclusion

When exploring almost any scientific phenomenon, one addresses two opposite issues: individual differences, on the one hand, and general patterns, on the other. In this paper we've focusing on the individual differences and proposed a "portrait" approach to bimodal communication. We are faced with a difficult task to find a good metric for analyzing oculomotor behavior of people in everyday communication. In previous papers, starting from [14], the authors were looking for oculomotor patterns, but their results depend critically on the metric used. In this paper, we compared the most common metrics and showed that individual differences have a much more serious weight than general patterns. We then identified four main coefficients that determine these individual differences: k_{aside}, k_{vip}, k_{chain}, and dur_{75}. By comparing these Core Oculomotor Portraits, we were able to make these individual differences more clear. However, a fact is a fact: there are far more individual differences than general patterns between

our Ns behavior. The proposed coefficients, in our opinion, clearly show (and even explain and predict) the observed individual differences.

Acknowledgements

This study is supported by Russian Foundation for Basic Research (project #19-012-00626).

References

[1] Abele A. Functions of gaze in social interaction: Communication and monitoring // Journal of Nonverbal Behavior. —1986. — Vol. 10. — №2. — P. 83–101.
[2] Baddeley A.D. Working memory, thought, and action. — Oxford: Oxford University Press, 2007.
[3] Brône G., Oben B. (eds.) Eye-tracking in Interaction: Studies on the role of eye gaze in dialogue. — John Benjamins, 2018.
[4] Chafe W. (ed.) The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. — Norwood: Ablex, 1980.
[5] Church R.B., Alibali M.W., Kelly S.D. (eds.) Why gesture? How the hands func.tion in speaking, thinking and communicating. — Amsterdam: John Benjamins, 2017.
[6] Fedorova O.V. On the communicative function of the gaze // Trudy Instituta russkogo yazyka im. V.V. Vinogradova — 2019. — Vol. 21 — P. 222–241.
[7] Fedorova O.V. Visual attention of the speaker and listener at the monological stages of natural communication: developing Kendon's ideas. — Submitted.
[8] Goldin-Meadow S. Widening the lens: What the manual modality reveals about language, learning, and cognition // Philosophical Transactions of the Royal society. — 2014. — Vol. 369.
[9] Grishina E.A. Russian gestures from a linguistic perspective [Russkaya zhestikulyatsiya s lingvisticheskoy tochki zreniya] — Moscow: Jazyki slavyanskoy kul'tury, 2017.
[10] Holler J., Kendrick K.H. Unaddressed participants' gaze in multi-person interaction: Optimizing recipiency // Frontiers in Psychology. — 2015. — Vol. 6. — №98.
[11] Holler J., Levinson S.C. Multimodal language processing in human communication // Trends in Cognitive Sciences — 2019. — Vol. 23. — №8. — P. 639–652.
[12] Horsley M., Eliot M., Knight B.A., Reilly R. Current Trends in Eye Tracking Research. — Springer, 2014.
[13] Jording M., Hartz A., Bente G., Schulte-Rüther M., Vogeley K. The "Social Gaze Space": A Taxonomy for Gaze-Based Communication in Triadic Interactions // Frontiers in Psychology. — 2018. — Vol. 9. — P. 226.
[14] Kendon A. Some functions of gaze-direction in social interaction // Acta Psychologica — 1967. — Vol. 26. — P. 22–63.
[15] Kendon A. Gesture. Visible action as utterance. — Cambridge 2004.
[16] Kibrik A.A., Fedorova O.V. An empirical study of multichannel communication: Russian Pear Chats and Stories, Psikhologiya // Zhurnal Vysshey shkoly ekonomiki. — 2018. — Vol. 15. — №2. — P. 191–200.
[17] Kibrik A.A., Fedorova O.V. A «portrait» approach to multichannel discourse // Eleventh International Conference on Language Resources and Evaluation (LREC). — Japan, 5 – 12 May 2018.
[18] Kibrik A.A., Podlesskaya V.I. (eds.) Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyakh: korpusnoye issledovaniye russkogo ustnogo diskursa]. — Moscow: Jazyki slavyanskikh kul'tur: 2009.
[19] Klein C., Ettinger U. (eds.) Eye Movement Research. — Springer, 2019.
[20] Levelt W.J.M. Speaking: From intention to articulation. — MIT Press, 1989.
[21] Liversedge S.P., Gilchrist I.D., Everling S. The Oxford Handbook of Eye Movements. — OUP: 2011.
[22] McNeill D.Gesture and thought. — Chicago, 2005.
[23] Mondada L. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction // Journal of Pragmatics. — 2019. — Vol. 145. — P. 47–62.
[24] Müller C., Fricke E., Cienki A., McNeill D. (eds.) Body – Language – Communication: An international handbook on multimodality in human interaction. — Berlin: Mouton de Gruyter, 2014.
[25] Rayner K., Pollatsek A. The psychology of reading. — Englewood Cliffs: Prentice Hall, 1989.
[26] Smith D.T., Schenk T. The premotor theory of attention: time to move on? // Neuropsychologia. — 2012. — Vol. 50. — P. 1104–1114.

Appendix. Full Oculomotor Portraits

Overall			
duration		1041.781	
R duration, ratio		696.375, 0.67, $\mathbf{k_{aside}}$ **0.3**	
Sur duration, ratio		219.956, 0.21	
		fixation	gaze
number		1656	402, $\mathbf{k_{chain}}$ **4.1**
mean duration, std		0.629, 1.198	2.592, 6.098
min, 25, 50, 75, max		0.06, 0.18, 0.3, ***0.6***, 22.493	0.06, 0.505, 1.05, 2.215, 95.298
R $\mathbf{k_{vip}}$ **3**	number, ratio	655, 39.6	171, 0.43, $\mathbf{k_{chain}}$ **3.8**
	mean, std	1.063, 1.734	4.072, 8.837
	min, 25, 50, *75*, max	0.06, 0.26, 0.5, *1.14*, 22.493	0.2, 0.7, 1.36, 3.29, 95.298
Sur	number, ratio	665, 40.2	183, 0.46, $\mathbf{k_{chain}}$ **3.6**
	mean, std	0.33, 0.447	1.202, 1.323
	min, 25, 50, *75*, max	0.06, 0.16, 0.24, *0.38*, 9.197	0.6, 0.37, 0.78, 1.629, 9.377
First telling			
duration		104.554	
R duration, ratio		47.908, 0.46, $\mathbf{k_{aside}}$ **1.2**	
Sur duration, ratio		56.646, 0.54	
		fixation	gaze
number		184	64, $\mathbf{k_{chain}}$ **2.9**
mean duration, std		0.568, 0.623	1.634, 1.51
min, 25, 50, 75, max		0.08, 0.2, 0.37, ***0.685***, 4.380	0.08, 0.695, 1.2, 2.09, 9.157
R $\mathbf{k_{vip}}$ **3.5**	number, ratio	40, 0.22	32, 0.5, $\mathbf{k_{chain}}$ **1.3**
	mean, std	1.198, 0.917	1.497, 1.21
	min, 25, 50, *75*, max	0.16, 0.555, 0.839, *1.655*, 4.380	0.2, 0.695, 1.249, 1.895, 5.277
Sur	number, ratio	144, 0.78	32, 0.5, $\mathbf{k_{chain}}$ **4.5**
	mean, std	0.393, 0.354	1.77, 1.769
	min, 25, 50, *75*, max	0.08, 0.16, 0.28, *0.48*, 1.937	0.08, 0.675, 1.159, 2.335, 9.157
Retelling			
duration		430.138	
R duration, ratio		382.512, 0.89, $\mathbf{k_{aside}}$ **0.1**	
Sur duration, ratio		47.626, 0.11	
		fixation	gaze
number		550	98, $\mathbf{k_{chain}}$ **5.6**
mean duration, std		0.782, 1.39	4.389, 11.055
min, 25, 50, 75, max		0.06, 0.22, 0.329, ***0.7***, 18.593	0.06, 0.469, 1.03, 1.97, 95.298
R $\mathbf{k_{vip}}$ **2.8**	number, ratio	388, 0.71	43, 0.44, $\mathbf{k_{chain}}$ **9**
	mean, std	0.986, 1.605	8.896, 15.633
	min, 25, 50, *75*, max	0.08, 0.26, 0.43, *0.94*, 18.593	0.24, 1.13, 1.98, 12.317, 95.298
Sur	number, ratio	162, 0.29	55, 0.56, $\mathbf{k_{chain}}$ **3**
	mean, std	0.289, 0.234	0.866, 0.791
	min, 25, 50, *75*, max	0.06, 0.14, 0.22, *0.34*, 1.68	0.06, 0.36, 0.6, 1.17, 4.12

Table 4: Full Oculomotor Portrait for N 06 (durations shown in seconds)

Overall			
duration		1389.722	
R duration, ratio		988.166, 0.71, **k_{aside} 0.3**	
Sur duration, ratio		285.072, 0.21	
		fixation	gaze
number		2817	684, **k_{chain} 4.1**
mean duration, std		0.493, 0.703	2.031, 5.435
min, 25, 50, 75, max		0.06, 0.12, 0.22, ***0.52***, 10.496	0.06, 0.34, 0.86, 1.865, 72.628
R **k_{vip} 5**	number, ratio	1150, 0.41	282, 0.41, **k_{chain} 4.1**
	mean, std	0.859, 0.946	3.502, 8.1
	min, 25, 50, *75*, max	0.06, 0.205, 0.55, *1.2*, 10.496	0.08, 0.72, 1.49, 2.855, 72.628
Sur	number, ratio	1351, 0.48	329, 0.48, **k_{chain} 4.1**
	mean, std	0.211, 0.175	0.859, 1.003
	min, 25, 50, *75*, max	0.06, 0.1, 0.16, *0.24*, 2.16	0.06, 0.24, 0.56, 1.06, 7.4
First telling			
duration		223.446	
R duration, ratio		126.713, 0.57, **k_{aside} 0.8**	
Sur duration, ratio		96.733, 0.43	
		fixation	gaze
number		609	157, **k_{chain} 3.9**
mean duration, std		0.367, 0.459	1.423, 1.236
min, 25, 50, 75, max		0.06, 0.12, 0.2, ***0.36***, 3.62	0.08, 0.56, 1.1, 1.96, 6.217
R **k_{vip} 5.5**	number, ratio	127, 0.21	77, 0.49, **k_{chain} 1.7**
	mean, std	1.278, 0.654	1.646, 1.16
	min, 25, 50, *75*, max	0.06, 0.47, 0.96, *1.32*, 3.62	0.14, 0.88, 1.26, 2.34, 6.217
Sur	number, ratio	482, 0.79	80, 0.51, **k_{chain} 6**
	mean, std	0.201, 0.147	1.209, 1.276
	min, 25, 50, *75*, max	0.06, 0.1, 0.16, *0.24*, 1.66	0.08, 0.295, 0.73, 1.585, 6.1
Retelling			
duration		318.993	
R duration, ratio		303.122, 0.95, **k_{aside} 0.1**	
Sur duration, ratio		15.871, 0.05	
		fixation	gaze
number		501	59, **k_{chain} 8.5**
mean duration, std		0.637, 0.889	5.407, 13.186
min, 25, 50, 75, max		0.06, 0.14, 0.28, ***0.74***, 9.44	0.06, 0.21, 0.52, 1.81, 72.628
R **k_{vip} 3**	number, ratio	436, 0.87	27, 0.46, **k_{chain} 16.1**
	mean, std	0.695, 0.936	11.227, 17.956
	min, 25, 50, *75*, max	0.06, 0.14, 0.34, *0.838*, 9.44	0.1, 0.52, 1.88, 12.03, 72.628
Sur	number, ratio	65, 0.13	32, 0.54, **k_{chain} 2**
	mean, std	0.244, 0.211	0.496, 0.713
	min, 25, 50, *75*, max	0.06, 0.08, 0.18, *0.28*, 0.957	0.06, 0.16, 0.29, 0.67, 4.037

Table 5: Full Oculomotor Portrait for N 16 (durations shown in seconds)

Overall			
duration		913.07	
R duration, ratio		738.513, 0.81, $\mathbf{k_{aside}}$ **0.2**	
Sur duration, ratio		131.265, 0.14	
		fixation	gaze
number		1344	406, $\mathbf{k_{chain}}$ **3.3**
mean duration, std		0.679, 0.984	2.254, 4.042
min, 25, 50, 75, max		0.06, 0.16, 0.28, ***0.74***, 10.2	0.08, 0.36, 0.94, 2.34, 39.411
R $\mathbf{k_{vip}}$ **5.6**	number, ratio	654, 0.49	193, 0.48, $\mathbf{k_{chain}}$ **3.4**
	mean, std	1.129, 1.246	3.826, 5.297
	min, 25, 50, *75*, max	0.06, 0.26, 0.72, *1.559*, 10.2	0.12, 1.04, 2.017, 4.337, 39.411
Sur	number, ratio	530, 0.39	192, 0.47, $\mathbf{k_{chain}}$ **2.8**
	mean, std	0.248, 0.174	0.682, 0.847
	min, 25, 50, *75*, max	0.06, 0.14, 0.2, *0.28*, 1.46	0.08, 0.22, 0.397, 0.805, 6.3
First telling			
duration		275.979	
R duration, ratio		207.695, 0.75, $\mathbf{k_{aside}}$ **0.3**	
Sur duration, ratio		68.284, 0.25	
		fixation	gaze
number		507	187, $\mathbf{k_{chain}}$ **2.7**
mean duration, std		0.544, 0.672	1.476, 1.63
min, 25, 50, 75, max		0.06, 0.16, 0.28, ***0.62***, 5.458	0.08, 0.37, 0.9, 1.779, 11.617
R $\mathbf{k_{vip}}$ **5.8**	number, ratio	226, 0.45	94, 0.5, $\mathbf{k_{chain}}$ **2.4**
	mean, std	0.919, 0.856	2.21, 1.848
	min, 25, 50, *75*, max	0.06, 0.285, 0.66, *1.24*, 5.458	0.16, 0.91, 1.58, 3.053, 11.617
Sur	number, ratio	281, 0.55	93, 0.5, $\mathbf{k_{chain}}$ **3**
	mean, std	0.243, 0.155	0.734, 0.903
	min, 25, 50, *75*, max	0.1, 0.13, 0.19, *0.215*, 0.8	0.08, 0.22, 0.48, 0.88, 6.3
Retelling			
duration		231.025	
R duration, ratio		217.405, 0.94, $\mathbf{k_{aside}}$ **0.1**	
Sur duration, ratio		13.02, 0.06	
		fixation	gaze
number		239	80, $\mathbf{k_{chain}}$ **3**
mean duration, std		0.967, 1.235	2.889, 4.514
min, 25, 50, 75, max		0.08, 0.2, 0.44, ***1.22***, 6.979	0.08, 0.295, 0.59, 4.295, 22.894
R $\mathbf{k_{vip}}$ **4.1**	number, ratio	192, 0.8	39, 0.49, $\mathbf{k_{chain}}$ **4.9**
	mean, std	1.132, 1.324	5.574, 5.281
	min, 25, 50, *75*, max	0.08, 0.22, 0.597, *1.48*, 6.979	0.56, 1.39, 4.82, 6.839, 22.894
Sur	number, ratio	46, 0.19	40, 0.5, $\mathbf{k_{chain}}$ **1.2**
	mean, std	0.283, 0.142	0.329, 0.18
	min, 25, 50, *75*, max	0.08, 0.2, 0.26, *0.36*, 0.78	0.08, 0.22, 0.29, 0.385, 1.12

Table 6: Full Oculomotor Portrait for N 21 (durations shown in seconds)

Overall			
duration		898.282	
R duration, ratio		472.258, 0.53, $\mathbf{k_{aside}}$ **0.4**	
Sur duration, ratio		180.083, 0.2	
		fixation	gaze
number		1149	185, $\mathbf{k_{chain}}$ **6.2**
mean duration, std		0.782, 0.922	4.89, 11.241
min, 25, 50, 75, max		0.06, 0.2, 0.42, ***0.98***, 7.1	0.1, 0.79, 1.84, 4.355, 118.712
R $\mathbf{k_{vip}}$ **2.9**	number, ratio	422, 0.37	70, 0.38, $\mathbf{k_{chain}}$ **6**
	mean, std	1.119, 1.11	6.776, 16.604
	min, 25, 50, *75*, max	0.08, 0.32, 0.72, *1.5*, 7.1	0.24, 1.039, 2.499, 5.218, 118.712
Sur	number, ratio	418, 0.36	74, 0.4, $\mathbf{k_{chain}}$ **5.7**
	mean, std	0.431, 0.455	2.428, 4.176
	min, 25, 50, *75*, max	0.06, 0.16, 0.26, *0.52*, 3.38	0.1, 0.46, 1.17, 2.315, 21.514
First telling			
duration		164.723	
R duration, ratio		34.088, 0.21, $\mathbf{k_{aside}}$ **3.8**	
Sur duration, ratio		130.635, 0.79	
		fixation	gaze
number		301	46, $\mathbf{k_{chain}}$ **6.5**
mean duration, std		0.547, 0.573	3.581, 4.978
min, 25, 50, 75, max		0.06, 0.18, 0.34, ***0.72***, 3.38	0.14, 0.825, 2.1, 3.405, 21.514
R $\mathbf{k_{vip}}$ **2**	number, ratio	35, 0.12	23, 0.5, $\mathbf{k_{chain}}$ **1.5**
	mean, std	0.974, 0.768	1.482, 1.228
	min, 25, 50, *75*, max	0.12, 0.369, 0.8, *1.25*, 2.94	0.24, 0.66, 0.96, 2.067, 5.277
Sur	number, ratio	266, 0.88	23, 0.5, $\mathbf{k_{chain}}$ **11.6**
	mean, std	0.491, 0.518	5.68, 6.321
	min, 25, 50, *75*, max	0.06, 0.16, 0.32, *0.635*, 3.38	0.14, 2.1, 2.5, 6.088, 21.514
Retelling			
duration		297.55	
R duration, ratio		293.15, 0.99, $\mathbf{k_{aside}}$ **0**	
Sur duration, ratio		4.08, 0.01	
		fixation	gaze
number		278	19, $\mathbf{k_{chain}}$ **14.6**
mean duration, std		1.07, 1.142	15.633, 30.355
min, 25, 50, 75, max		0.08, 0.28, 0.6, ***1.454***, 7.1	0.1, 0.33, 0.66, 16.157, 118.712
R $\mathbf{k_{vip}}$ **3.6**	number, ratio	262, 0.94	9, 0.47, $\mathbf{k_{chain}}$ **29.1**
	mean, std	1.119, 1.158	32.572, 38.207
	min, 25, 50, *75*, max	0.1, 0.3, 0.66, *1.495*, 7.1	0.66, 11.779, 18.757, 29.811, 118.712
Sur	number, ratio	14, 0.05	9, 0.47, $\mathbf{k_{chain}}$ **1.6**
	mean, std	0.291, 0.172	0.396, 0.376
	min, 25, 50, *75*, max	0.1, 0.15, 0.21, *0.415*, 0.62	01, 0.18, 0.34, 0.4, 1.34

Table 7: Full Oculomotor Portrait for N 22 (durations shown in seconds)

Overall			
duration		721.661	
R duration, ratio		384.117, 0.53, $\mathbf{k_{aside}}$ **0.4**	
Sur duration, ratio		157.436, 0.22	
		fixation	gaze
number		1721	259, $\mathbf{k_{chain}}$ **6.6**
mean duration, std		0.419, 0.644	2.763, 4.629
min, 25, 50, 75, max		0.06, 0.12, 0.24, ***0.46***, 11.977	0.08, 0.5, 1.18, 2.789, 46.231
R $\mathbf{k_{vip}}$ **1.7**	number, ratio	643, 0.37	96, 0.37, $\mathbf{k_{chain}}$ **6.7**
	mean, std	0.597, 0.879	3.939, 6.034
	min, 25, 50, *75*, max	0.06, 0.149, 0.34, *0.66*, 11.977	0.08, 0.935, 2.049, 5.432, 46.231
Sur	number, ratio	446, 0.26	117, 0.45, $\mathbf{k_{chain}}$ **3.8**
	mean, std	0.353, 0.531	1.346, 1.795
	min, 25, 50, *75*, max	0.06, 012, 0.22, *0.4*, 7.097	0.08, 0.34, 0.72, 1.48, 9.637
First telling			
duration		166.881	
R duration, ratio		119.693, 0.72, $\mathbf{k_{aside}}$ **0.4**	
Sur duration, ratio		46.768, 0.28	
		fixation	gaze
number		370	65, $\mathbf{k_{chain}}$ **5.7**
mean duration, std		0.451, 0.426	2.567, 2.885
min, 25, 50, 75, max		0.06, 0.16, 0.32, ***0.579***, 2.9	0.08, 0.74, 1.68, 3.379, 18.557
R $\mathbf{k_{vip}}$ **1.5**	number, ratio	244, 0.66	31, 0.48, $\mathbf{k_{chain}}$ **7.9**
	mean, std	0.491, 0.461	3.861, 3.399
	min, 25, 50, *75*, max	0.06, 0.18, 0.35, *0.66*, 2.9	0.94, 1.74, 2.64, 5.477, 18.557
Sur	number, ratio	123, 0.33	32, 0.49, $\mathbf{k_{chain}}$ **3.8**
	mean, std	0.38, 0.342	1.462, 1.646
	min, 25, 50, *75*, max	0.06, 0.15, 0.3, *0.44*, 2.317	0.12, 0.5, 0.84, 1.755, 7.817
Retelling			
duration		239.37	
R duration, ratio		198.605, 0.83, $\mathbf{k_{aside}}$ **0.2**	
Sur duration, ratio		40.685, 0.17	
		fixation	gaze
number		363	59, $\mathbf{k_{chain}}$ **6.2**
mean duration, std		0.659, 1.124	4.066, 7.317
min, 25, 50, 75, max		0.06, 0.14, 0.28, ***0.64***, 11.977	0.08, 0.41, 0.98, 5.817, 46.231
R $\mathbf{k_{vip}}$ **1.6**	number, ratio	286, 0.79	30, 0.51, $\mathbf{k_{chain}}$, **9.5**
	mean, std	0.694, 1.131	6.638, 9.407
	min, 25, 50, *75*, max	0.06, 0.14, 0.32, *0.715*, 11.977	0.08, 0.675, 4.2, 7.799, 46.231
Sur	number, ratio	76, 0.21	28, 0.47, $\mathbf{k_{chain}}$ **2.7**
	mean, std	0.535, 1.102	1.453, 2.22
	min, 25, 50, *75*, max	0.06, 0.1, 0.2, *0.445*, 7.097	0.08, 0.335, 0.69, 1.185, 9.637

Table 8: Full Oculomotor Portrait for N 23 (durations shown in seconds)

Overall			
duration		1241.489	
R duration, ratio		991.884, 0.8, $\mathbf{k_{aside}}$ **0.1**	
Sur duration, ratio		114.308, 0.09	
		fixation	gaze
number		2326	388, $\mathbf{k_{chain}}$ **6**
mean duration, std		0.534, 0.803	3.209, 9.97
min, 25, 50, 75, max		0.06, 0.14, 0.24, ***0.56***, 8.597	0.06, 0.36, 0.88, 2.478, 112.653
R $\mathbf{k_{vip}}$ **2.9**	number, ratio	1506, 0.65	158, 0.41, $\mathbf{k_{chain}}$ **9.5**
	mean, std	0.659, 0.94	6.278, 15.033
	min, 25, 50, *75*, max	0.06, 0.14, 0.3, *0.74*, 8.597	0.08, 0.8, 2.009, 4.899, 112.653
Sur	number, ratio	515, 0.22	164, 0.42, $\mathbf{k_{chain}}$ **3.1**
	mean, std	0.222, 0.154	0.719, 0.851
	min, 25, 50, *75*, max	0.06, 0.14, 0.18, *0.26*, 1.84	0.06, 0.2, 0.38, 0.77, 4.917
First telling			
duration		182.361	
R duration, ratio		146.57, 0.8, $\mathbf{k_{aside}}$ **0.2**	
Sur duration, ratio		35.371, 0.19	
		fixation	gaze
number		333	99, $\mathbf{k_{chain}}$ **3.4**
mean duration, std		0.548, 0.77	1.887, 2.416
min, 25, 50, 75, max		0.06, 0.159, 0.22, ***0.5***, 4.997	0.1, 0.34, 1.1, 2.31, 15.857
R $\mathbf{k_{vip}}$ **5.4**	number, ratio	167, 0.5	49, 0.49, $\mathbf{k_{chain}}$ **3.4**
	mean, std	0.878, 0.975	3.083, 2.892
	min, 25, 50, *75*, max	0.06, 0.18, 0.46, *1.39*, 4.997	0.22, 1.24, 2.08, 4.38, 15.857
Sur	number, ratio	163, 0.49	49, 0.49, $\mathbf{k_{chain}}$ **3.3**
	mean, std	0.217, 0.121	0.722, 0.813
	min, 25, 50, *75*, max	0.6, 0.14, 0.18, *0.26*, 0.9	0.1, 0.2, 0.34, 1.077, 3.977
Retelling			
duration		238.287	
R duration, ratio		235.527, 0.99, $\mathbf{k_{aside}}$ **0**	
Sur duration, ratio		2.76, 0.01	
		fixation	gaze
number		203	17, $\mathbf{k_{chain}}$ **11.9**
mean duration, std		1.174, 1.525	14.017, 32.909
min, 25, 50, 75, max		0.06, 0.19, 0.48, ***1.61***, 8.597	0.18, 0.3, 0.62, 2.86, 112.653
R $\mathbf{k_{vip}}$ **6.5**	number, ratio	191, 0.94	9, 0.53, $\mathbf{k_{chain}}$ **21.2**
	mean, std	1.233, 1.553	26.17, 42.578
	min, 25, 50, *75*, max	0.06, 0.2, 0.58, *1.76*, 8.597	0.36, 0.76, 2.86, 25.434, 112.653
Sur	number, ratio	12, 0.6	8, 0.47, $\mathbf{k_{chain}}$ **1.5**
	mean, std	0.23, 0.113	0.345, 0.21
	min, 25, 50, *75*, max	0.08, 0.18, 0.21, *0.27*, 0.44	0.18, 0.195, 0.26, 0.41, 0.72

Table 9: Full Oculomotor Portrait for N 24 (durations shown in seconds)

Text Simplification with Autoregressive Models

Alena Fenogenova
Sberbank, SberDevices
Moscow, Russia
alenush93@gmail.com

Abstract

Text Simplification is the task of reducing the complexity of the vocabulary and sentence structure of the text while retaining its original meaning with the goal of improving readability and understanding. We explore the capability of the autoregressive models such as RuGPT3 (Generative Pre-trained Transformer 3 for Russian) to generate high quality simplified sentences. Within the shared task RuSimpleSentEval we present our solution based on different usages of RuGPT3 models. The following setups are described: 1) few-shot unsupervised generation with the RuGPTs models 2) the effect of the size of the training dataset on the downstream performance of fine-tuned model 3) 3 inference strategies 4) the downstream transfer and post-processing procedure using pre-trained paraphrasers for Russian. This paper presents the second-place solution on the public leaderboard and the fifth-place solution on the private leaderboard. The proposed method is comparable with the novel state-of-the-art approaches. Additionally, we analyze the performance and discuss the flaws of RuGPTs generation.

Keywords: text simplification, RuGPT3, text generation, paraphrase generation
DOI: 10.28995/2075-7182-2021-20-227-234

Упрощение текстов с помощью авторегрессионных моделей

Алена Феногенова
Сбербанк, SberDevices
Москва, Россия
alenush93@gmail.com

Аннотация

Упрощение текста — задача автоматического получения упрощенного предложения из сложного. В работе представлена методика упрощения текстов на основе авторегрессионных моделей, в частности RuGPT3 (Generative Transformer 3 for Russian). Решение представлено в рамках соревнования RuSimpleSentEval, которое заняло второе место на публичном лидерборде по метрике SARI и пятое место на приватном лидерборде. В работе рассмотрены следующие подходы: 1) генерация упрощенного текста с помощью техники few-shot, 2) изучение влияния размера обучающей выборки и параметров на целевое качество моделей, обученных с помощью метода fine-tuning, 3) сравнение трех инференс стратегий и постбработки, 4) применение предобученных моделей для генерации парафразов на русском языке на целевой задаче и в качестве компонента пост-обработки сгенерированных упрощенных текстов.

Ключевые слова упрощение текстов, RuGPT3, генерация парафразов

1 Introduction

The task of text simplification (TS) aims to reduce its linguistic complexity in order to improve readability and understanding. Text complexity criteria include the presence of complex grammatical structures, participial and adverbial constructions, subordinate sentences, the presence of infrequent and ambiguous words. Recent research on TS has been of keen interest, especially after the development of automatic approaches which have led to the transition from manually defined rules to automatic simplification

using neural networks. Simplification has a variety of important applications. For example, in socio-psychological respect, it increases the information accessibility for those with cognitive disorders such as aphasia, dyslexia, and autism, as well as for non-native speakers. Furthermore, automatic text simplification could improve performance on other NLP tasks, such as paraphrasing, summarization, information extraction, semantic role labeling, and machine translation.

Existing methods have been predominantly designed for English due to the availability of high-quality text corpora which contain aligned complex and simplified sentences such as Newsela[1] [24] and Turk Corpus [25]. WikiLarge constructed from Wikipedia and Simple Wikipedia is a very common dataset for English as well. However, the construction of such datasets for new language is expensive, and no attempts have been made to create a TS dataset for the Russian language. To this end, the shared task RuSimpleSentEval-2021 [16] aims to fill this gap and facilitate the development of automatic TS methods for Russian. This paper describes the submission to the shared task and proposes the TS method based on RuGPT3, and details the experiments with the autoregressive models for Russian. We explore the RuGPT3[2] models capabilities in a full compliance with the competition rules, study the effect of the size of the training dataset on the model performance, combine different inference strategies and post-processing techniques. The method has achieved the second place on the RuSimpleSentEval public leaderboard and the fifth place on the private leaderboard.

The remainder is organized as follows: Section 2 briefly describes the prior research in the field; Section 3 outlines the data used in the experiments; Section 4 provides the description of the experiments; we discuss the results and provide the analysis of the proposed method and generated abilities of the best model in Section 5, section 6 concludes the paper.

2 Related Work

The task of TS is similar in nature to other sequence-to-sequence NLP tasks such as machine translation, paraphrase generation [21, 17] and most to text summarization. It can be considered as text summarization which can involve selecting sentences from the input text (extractive) or re-writing the input text (abstractive) in order to preserve most of the meaning [7]. In contrast to text summarization, simplification methods do not necessarily "compress" the input text and thus can produce longer texts, e.g. when generating term explanations. Whereas text summarization predominantly aims at filtering out the redundant text segments, TS approaches preserve the structure of the text. Despite this, a number of studies have explored the combinations of the approaches by integrating TS methods into summarization systems [27, 19].

The survey [6] provides a comprehensive overview of TS approaches, including a brief description of the earlier attempts to solve the task, discussion of various aspects of simplification (lexical, semantic, and syntactic), and the latest techniques being utilized in the field. Recent research in the field has clearly shifted towards utilizing deep learning techniques to perform TS, with a specific focus on developing solutions to combat the lack of data available for simplification. [18] is another review of the most significant studies in TS. It highlights more than 300 studies of the last three decades in the field of TS. The paper covers the corpora and evaluation metrics, for example, BLEU [13] and the most reliable metric for the sentence simplification task SARI [25].

The state-of-the-art results on TS task for English on a Turk Corpus are demonstrated by the following models:

1. DMASS & DCSS [29] is a combination of Deep Memory Augmented Sentence Simplification (DMASS) model and Deep Critic Sentence Simplification (DCSS) that has achieved 40.45 SARI.
2. ACCESS [10] by Facebook has obtained 72.54 BLEU and 41.87 SARI. The method shows that explicitly conditioning the sequence-to-sequence models on control tokens such as length, amount of paraphrasing, lexical complexity and syntactic complexity, increases the results of generation.
3. MUSS [11] has received the highest scores 78.17 (BLEU) and 42.53 (SARI). The method incorporates leveraging unsupervised data to train TS systems in multiple languages using the controllable

[1] https://newsela.com/data
[2] https://github.com/sberbank-ai/ru-gpts

generation mechanisms and pre-training.

Another line of research is focused on approaches based on reinforcement learning [28]. Transformer-based language models [23] have been applied to the sequence-to-sequence tasks for Russian, ranging from text summarization [9] to news generation[4]. The large scale pre-trained transformers represent a promising direction in the field of TS and comparable to the state-of-the-art methods. GPT-3 has achieved competitive performance on text summarization and simplification tasks [12, 22, 20]. In line with these works, we focus on the applicability of the autoregressive models, namely RuGPT3, for TS.

3 Data

The TS datasets contain parallel pairs of complex sentences (source) and their corresponding simplified versions (source).

The organizers of the RuSimpleSentEval-2021 shared task have introduced a TS dataset constructed by automatic translation and post-processed WikiLarge corpus [25]. The resulting dataset was split into train, dev and test sets. The additional dev, public and private test sets were created via crowd-sourcing using Yandex.Toloka[3]. The training set contains inappropriate examples due to being automatically constructed. Consider an example, where the sentences are likely to refer to the same town but the target sentence contains extra information which can not be derived from the source sentence: Город также является центром производства сахара и промышленности. => В 2002 году общая численность населения муниципалитета составляла 77 698 человек: 38 093 мужчины и 39 605 женщин. There are also some cases where the translation is only partially done: Belleview находится по адресу. ==> Бельвью - город во Флориде в США. Another problem is sentences where the target sentence contains more information, which is a crucial case because it contradicts the definition of simplification. The sentence is not simplified, instead it is complicated: Некоторые могут проявлять миксотроию. ==> Некоторые могут проявлять миксотрофию при использовании смешанных источников энергии. As we see further the data for training is a primary issue for the prominent performance of the TS methods. Thus, we make an attempt to overcome these issues and conduct the experiments in the following data settings: 1) all the data provided by the organizers (further in the text "$data\_all$") 2) all cleaned data ("$clean\_all$") 3) a 10000 examples subset of cleaned data (''$clean\_subset$"). The cleaning procedure of proposed data contains the following filtration steps:

- Discarding examples with less than two lemmas in the intersection between the lemmatized source and target sentences. We removed the stopwords during this step and lemmatize the sentences with pymorphy2 tagger[4];
- Discarding examples where the source sentence is a substring of the target one and the length is greater than of the source one.

4 Experimental Setup

The shared task is evaluated with SARI (System output Against References and against the Input sentence) released in EASSE[1][5]. The baseline of the competition is a multilingual BART (mBART) [8] which is commonly used for the summarization task including the Russian language [5]. The model was fine-tuned on the train set and achieved the 30.15 SARI score on the public leaderboard. We now describe the experiments conducted in this work.

Downstream transfer using pre-trained paraphrasers The motivation behind this setting is that the TS task is similar to paraphrase generation. To this end, we use the pre-trained paraphrasers for Russian and evaluate them on the task without fine-tuning [3][6]. We used mt5-base and RuGPT3 paraphrasers and the following generation hyperparameters: temperature 1, top_k repetition_penalty 1, top_p 0.9, max length 100 and the probability threshold of 0.8.

[3] https://toloka.yandex.ru/

[4] `https://github.com/kmike/pymorphy2`

[5] https://github.com/feralvam/easse

[6] https://github.com/RussianNLP/russian_paraphrasers

Fine-tuning Another approach includes fine-tuning of the following models:

1. mT5[26] - Multilingual T5 (mT5) by Google is a massively multilingual pre-trained text-to-text transformer model trained on the mC4 corpus in 101 languages including Russian.
2. RuGPT3-Large is a Russian open source analogue of GPT-3[2]. RuGPT3-Large[7] (760 millions of parameters) was trained on Internet text on 1024 context length with transformers on 80 billion tokens around 3 epochs, and then was fine-tuned on 2048 context.
3. RuGPT3-XL[8] was trained with 512 sequence length using Deepspeed and Megatron code by Sber-Devices team, on 80B tokens dataset for 4 epochs. After that the model was finetuned 1 epoch with sequence length 2048.

Since the best performance on the public leaderboard was achieved with the RuGPT3-XL model, we used it in a series of further experiments.

Exploring the effect of the training data size on the downstream performance In this setting, we first experiment with different sizes of the training data and the filtration procedure described in Section 3. Second, we explore the following setup:

1. **Few-shot method** with a pre-trained RuGPT3-XL model. We feed the model with 5 examples from the dev set combined with "prompts" and generate the output for the test examples. An example of the "prompt" is presented in Figure 1. For each test example, we generate 5 candidates and rank them by the lowest perplexity score.
2. **Fine-tuning and decoding methods**: we fine-tune the RuGPT3-XL model and experiment with greedy decoding, top-k and top-p sampling, and beam search methods.
3. **Post-processing** of the generated output using heuristic-based approach and re-writing the output with a pre-trained paraphraser. First, we check the appropriateness of the punctuation marks and casing of the named entities. Second, we consider the generated output to be inappropriate if: (a) the length of the output is too short, (b) there is no lemmas in the intersection of the source and target sentences, (c) the source sentence is a sub-string of the generated sentence, or the Levenshtein distance between the sentences is less than 5: in this case we rewrite the output with using the paraphraser.

Упрости: "Агрессия, как со стороны пациентов так и направленная против них, обычно случается в контексте сложных социальных взаимодействий в семье, а также является проблемой в условиях клиники и по месту жительства больного." ==> Конфликты в медучреждении - следствии конфликтов в семье.
Упрости: "Алма-Ата — лёгкая и пищевая промышленность, машиностроение; и Тараз — машиностроение, химическая и пищевая промышленность. " ==> В Алма-Ате развита лёгкая и пищевая промышленность, машиностроение; а в Таразе - машиностроение, химическая и пищевая промышленность.
Упрости: "14 декабря 1944 года рабочий посёлок Ички был переименован в рабочий посёлок Советский, после чего поселковый совет стал называться Советским." ==> 14 декабря 1944 года рабочий посёлок Ички переименован в Советский.
Упрости: "Автор гола в ворота сборной Англии, получившего название «Гол столетия», и признанного лучшим голом в истории чемпионатов мира; в той же игре забил мяч рукой, этот случай известен как «Рука Бога»." ==>

Figure 1: The "prompt" example for the few-shot technique with the RuGPT3-XL model.

All the experiments were conducted and measured on the public leaderboard. The results are presented in Table 1. During the public competition phase, the best submissions achieved with the RuGPT3-XL model trained on 10k cleaned training examples using greedy decoding and with the RuGPT3-XL model

[7] https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2
[8] https://github.com/sberbank-ai/ru-gpts/tree/master

trained on all the cleaned data with sampling. We assumed that the combination of the two configurations was the best option. Sentence transformers[9] [15] were used to compare the generated sentences from the two configurations and to choose the best one. The library provides the multilingual model for paraphrase identification "paraphrase-xlm-r-multilingual-v1"[14]. The source sentence embeddings were compared with generated sentence embeddings from the two configurations and the one with a higher cosine similarity was kept as the final answer. Formally, the scheme of the final submission is presented in Figure 2.

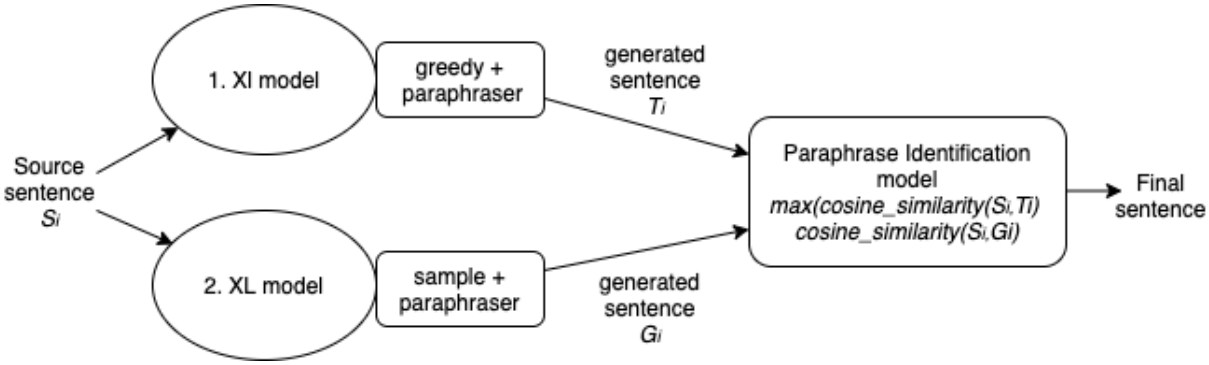


Figure 2: A graphical representation of the final method pipeline.

Method			SARI
Data	Model	Inference	
Ru/En wiki corpus	Baseline	-	30.15
Paraphraser corpus	paraphraser on RuGPT3	-	31.135
Paraphraser corpus	paraphraser on mT5	-	31.2
$data\_all$	mT5 base fine-tune	-	32.12
$clean\_all$	RuGPT3 large fine-tune	-	35.09
$nodata$	RuGPT3-XL few-shot	-	28.05
$clean\_all$	RuGPT3-XL fine-tune	beam + postproc	37.04
$clean\_subset$	RuGPT3-XL	greedy + postproc	38.10
$clean\_all$	RuGPT3-XL	sampling + postproc	38.01
$clean\_all/clean\_subset$	RuGPT3-XL	greedy/sampling + postproc	**38.87**

Table 1: Results on the public test set. **Data** represents the data on which the model was fine-tuned. **Model** shows the model, and **Inference** refers to the post-processing and decoding method.

Method			SARI
Data	Model	Inference	
$clean\_all$	RuGPT3-XL	beam + postproc	36.68
$clean\_all$	RuGPT3-XL	sample + postproc	36.78
$clean\_subset$	RuGPT3-XL	greedy + postproc	**37.82**
$clean\_all/clean\_subset$	RuGPT3-XL	greedy/sample + postproc	37.80

Table 2: Results on the private test set. **Data** represents the data on which the model was fine-tuned. **Model** shows the model, and **Inference** refers to the post-processing and decoding method.

[9] https://github.com/UKPLab/sentence-transformers

Method	Source	Generated	Estimate
few-shot	А один из ссыльных, Оскар Александрович Энгберг, изготовил жениху и невесте обручальные кольца из медного пятака.	а Один осужденный сделал молодожёнам обрученные перстни из пятикопеечной монеты	Good paraphrase, the meaning is the same
	Авиационные линии поддерживают связь крупнейших городов Италии со многими городами Европы, а также других континентов.	Крупнейшие авиарейсы позволяют экономить время для путешествий пассажиров внутри города	The thema is the same, the meaning is different.
	Актёру были близки и фарсовая, балаганная стилистика картины, и главный герой — меняющий маски и одежды подпольщик Бродский.	Актеру была близка сатирическая манера отображения происходящего его героем-подпольщиком; именно этого он сам пытался достичь за счёт изменения образа жизни (изменил имя). Неизменным оставалось лишь содержание произведения искусства - ирония!	The start of the generated sentence is good simplification, after - too many extra information.
fine-tune	В зависимости от части в сюжетной линии в прошлом времени игрок может управлять одним или двумя протагонистами.	В зависимости от части в сюжетной линии игрок может управлять одним или двумя персонажами.	Good simplification
	Аспирин — это лечение первой линии для лихорадки и симптомов суставной боли при остром ревматизме.	Аспирин обычно используется для лечения лихорадки, боли в суставах и мышечных судорог.	Good simplification
	Алкоголь может увеличить риск рождения ребёнка с врождёнными аномалиями нервной системы и обусловить задержку роста.	Алкоголь в умеренных количествах не влияет на рост и развитие плода.	Thema is the same, but the meaning is opposite
	Алма-Ата становится одним из торговых, ремесленных и сельскохозяйственных центров на Великом Шёлковом пути, имеющих монетный двор.	В Алма-Ате появляется монетный двор.	Too short simplification, the meaning is cut

Figure 3: Examples of the simplified sentences produced by the RuGPT3-XL model.

5 Results and discussion

The RuGPT3-XL models show results comparable with other methods submitted to the leaderboard. The fine-tuned models achieve results about 37 SARI depending on the inference strategy. Beam search shows worse results compared with greedy and sampling decoding methods. The performance without the post-processing component is lower across all configurations, indicating that the Russian paraphrasers are a valuable tool for simplification procedure.

We have obtained different results for the generation-based methods. First, the few-shot method is beneficial due to its simplicity. For the best result, the developer needs to investigate the prompts and choose the most optimal one. Without any fine-tuning, the RuGPT3-XL model generates a number of appropriate simplified sentences. We manually validated 50 examples produced by this method: 16% are

appropriately simplified sentences, 41% are semantically inappropriate sentences but on the same topic as the original sentence, and 43% are fully inappropriate. The examples are provided in Figure 3.

The fine-tuning approach receives reasonable performance. However, there is room for improvements. One can see that the meaning of the produced sentence can be opposite despite being simplified. Another case is that the sentence gets overly "compressed" thus losing the relevant information. The best combined solution has achieved the 38.87 SARI score, as we tried to increase the score based on the best performing submissions. However this approach has not been proved to be the best option on the public leaderboard. After the competition, when the submissions were no longer limited, we discovered that the greedy decoding with post-processing shows better results. Thus, the best configuration is the RuGPT3-XL model fine-tuned on all clean_subset with greedy decoding and paraphraser post-processing that achieves a 37.82 SARI score. We observe the performance drops between the public and private test sets (from 38.10 to 37.82). A possible reason is the effect of the different generation hyperparameters for both the RuGPT3-XL model and the paraphraser, shifts in the test distributions or the model overfitting.

6 Conclusion

In this paper, we present the submission to the RuSimpleSentEval 2021 shared task devoted to the problem of text simplification. The method combines the autoregressive transformer, namely the RuGPT3-XL model, and pre-trained paraphrasers for the Russian language. The experiments are conducted using various method configurations, ranging from the few-shot and fine-tuning approaches to heuristic-based data pre-processing and post-processing procedures. The results demonstrate that the proposed method can simplify sentences with and without any fine-tuning, solely based on the prompts fed as little supervision. Our approach has achieved second place on the public leaderboard and fifth place on the private leaderboard reaching the 38.87 and 37.8 SARI score, respectively. The qualitative and quantitative analysis shows that there is still room for improvements which we consider an exciting direction for future work. Another line includes the applicability of the approach to the English language and comparison between languages. We hope that our method will be served as a prototype in the applications where text simplification is required, or used as a strong baselines for development of more sophisticated text simplification systems for Russian.

Acknowledgements

I would like to express my deepest appreciation to Vladislav Mikhailov (Sberbank, SberDevices), the best human-simplificator I know. No GPT can handle simplification tasks better than you ;).

References

[1] Fernando Alva-Manchego et al. "EASSE: Easier automatic sentence simplification evaluation". In: *arXiv preprint arXiv:1908.04567* (2019).

[2] Tom B Brown et al. "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165* (2020).

[3] Alena Fenogenova. "Russian Paraphrasers: Paraphrase with Transformers". In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. 2021, pp. 11–19.

[4] Daniil Gavrilov, Pavel Kalaidin and Valentin Malykh. "Self-attentive model for headline generation". In: *European Conference on Information Retrieval*. Springer. 2019, pp. 87–93.

[5] Ilya Gusev. "Dataset for Automatic Summarization of Russian". In: *arXiv preprint arXiv:2006.11063* (2020).

[6] Behrooz Janfada and Behrouz Minaei-Bidgoli. "A Review of the Most Important Studies on Automated Text Simplification Evaluation Metrics". In: *2020 6th International Conference on Web Research (ICWR)*. IEEE. 2020, pp. 271–278.

[7] Chandra Khatri, Gyanit Singh and Nish Parikh. "Abstractive and extractive text summarization using document context vector and recurrent neural networks". In: *arXiv preprint arXiv:1807.08000* (2018).

[8] Yinhan Liu et al. “Multilingual denoising pre-training for neural machine translation”. In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 726–742.

[9] Valentin Malykh, Denis Porplenko and Elena Tutubalina. “Generating Sport Summaries: A Case Study for Russian”. In: *Analysis of Images, Social Networks and Texts*. Ed. by Wil M. P. van der Aalst et al. Cham: Springer International Publishing, 2021, pp. 149–161. ISBN: 978-3-030-72610-2.

[10] Louis Martin et al. “Controllable sentence simplification”. In: *arXiv preprint arXiv:1910.02677* (2019).

[11] Louis Martin et al. “Multilingual unsupervised sentence simplification”. In: *arXiv preprint arXiv:2005.00352* (2020).

[12] Takumi Maruyama and Kazuhide Yamamoto. “Extremely Low Resource Text simplification with Pre-trained Transformer Language Model”. In: *2019 International Conference on Asian Language Processing (IALP)*. IEEE. 2019, pp. 53–58.

[13] Kishore Papineni et al. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[14] Nils Reimers and Iryna Gurevych. “Making monolingual sentence embeddings multilingual using knowledge distillation”. In: *arXiv preprint arXiv:2004.09813* (2020).

[15] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084* (2019).

[16] Andrey Sakhovskiy et al. “RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian”. In: *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”*. Vol. XX. 2021, pp. xx–xx.

[17] Advaith Siddharthan. “A survey of research on text simplification”. In: *ITL-International Journal of Applied Linguistics* 165.2 (2014), pp. 259–298.

[18] Punardeep Sikka et al. “A Survey on Text Simplification”. In: *arXiv preprint arXiv:2008.08612* (2020).

[19] Sara Botelho Silveira and António Branco. “Combining a double clustering approach with sen-tence simplification to produce highly informative multi-document summaries”. In: *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*. IEEE. 2012, pp. 482–489.

[20] Neha Srikanth and Junyi Jessy Li. “Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification”. In: *arXiv preprint arXiv:2010.10035* (2020).

[21] Suha S Al-Thanyyan and Aqil M Azmi. “Automated text simplification: A survey”. In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–36.

[22] Hoang Van, David Kauchak and Gondy Leroy. “AutoMeTS: The Autocomplete for Medical Text Simplification”. In: *arXiv preprint arXiv:2010.10573* (2020).

[23] Ashish Vaswani et al. “Attention is all you need”. In: *arXiv preprint arXiv:1706.03762* (2017).

[24] Wei Xu, Chris Callison-Burch and Courtney Napoles. “Problems in current text simplification research: New data can help”. In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 283–297.

[25] Wei Xu et al. “Optimizing statistical machine translation for text simplification”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 401–415.

[26] Linting Xue et al. “mT5: A massively multilingual pre-trained text-to-text transformer”. In: *arXiv preprint arXiv:2010.11934* (2020).

[27] David Zajic et al. “Multi-candidate reduction: Sentence compression as a tool for document summarization tasks”. In: *Information Processing & Management* 43.6 (2007), pp. 1549–1570.

[28] Xingxing Zhang and Mirella Lapata. “Sentence simplification with deep reinforcement learning”. In: *arXiv preprint arXiv:1703.10931* (2017).

[29] Sanqiang Zhao et al. “Integrating transformer and paraphrase rules for sentence simplification”. In: *arXiv preprint arXiv:1810.11193* (2018).

Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models

Alena Fenogenova[1] **Maria Tikhonova**[1,2] **Vladislav Mikhailov**[1,2]
Tatiana Shavrina[1,2,3] **Anton Emelyanov**[1,4] **Denis Shevelev**[1]
Alexandr Kukushkin[5] **Valentin Malykh**[6,7] **Ekaterina Artemova**[2,6]

[1]SberDevices, Sberbank, Moscow, Russia
[2]National Research University Higher School of Economics, Moscow, Russia
[3]ANO «AI Research Institute», Moscow, Russia
[4]Moscow Institute of Physics and Technology, Moscow, Russia
[5]Alex Kukushkin Lab, Moscow, Russia
[6]Huawei Noah's Ark lab, Moscow, Russia
[7]Kazan Federal University, Kazan, Russia

Abstract

In the last year, new neural architectures and multilingual pre-trained models have been released for Russian, which led to performance evaluation problems across a range of language understanding tasks.

This paper presents Russian SuperGLUE 1.1, an updated benchmark styled after GLUE for Russian NLP models. The new version includes a number of technical, user experience and methodological improvements, including fixes of the benchmark vulnerabilities unresolved in the previous version: novel and improved tests for understanding the meaning of a word in context (RUSSE) along with reading comprehension and common sense reasoning (DaNetQA, RuCoS, MuSeRC). Together with the release of the updated datasets, we improve the benchmark toolkit based on `jiant` framework for consistent training and evaluation of NLP-models of various architectures which now supports the most recent models for Russian. Finally, we provide the integration of Russian SuperGLUE with a framework for industrial evaluation of the open-source models, MOROCCO (MOdel ResOurCe COmparison), in which the models are evaluated according to the weighted average metric over all tasks, the inference speed, and the occupied amount of RAM. Russian SuperGLUE is publicly available at `https://russiansuperglue.com/`.

Keywords: model evaluation, natural language understanding, benchmarks, NLP models, language modelling, general language understanding evaluation

DOI: 10.28995/2075-7182-2021-20-235-245

Russian SuperGLUE 1.1: пересматривая невыученные уроки русскоязычных NLP-моделей

Алена Феногенова[1] Мария Тихонова[1,2] Владислав Михайлов[1,2]
Татьяна Шаврина[1,2,3] Антон Емельянов[1,4] Денис Шевелев[1]
Александр Кукушкин[5] Валентин Малых[6,7] Екатерина Артемова[2,6]

[1]SberDevices, Сбербанк, Москва, Россия
[2]НИУ «Высшая школа экономики», Москва, Россия
[3]АНО «Институт Искусственного Интеллекта», Москва, Россия
[4]Московский физико-технический институт, Москва, Россия
[5]Лаборатория Александра Кукушкина, Москва, Россия
[6]Huawei Noah's Ark lab, Москва, Россия
[7]Казанский (Приволжский) федеральный университет, Казань, Россия

Аннотация

В прошлом году на русскоязычном материале были обучены новые нейронные архитектуры, в том числе мультиязычные NLP-модели, что привело к новым вызовам в оценке качества решений задач понимания естественного языка.

В этой статье представлен Russian SuperGLUE 1.1, бенчмарк на основе GLUE для оценки языковых моделей для русского языка. Новая версия включает в себя ряд технических обновлений, улучшение пользовательского опыта и устранение методологических уязвимостей версии 1.0., в том числе создание новых тестовых сетов и улучшение датасетов на понимание смысла слова в контексте (RUSSE), машинное чтение и здравый смысл (DaNetQA, RuCoS, MuSeRC). Кроме того, представлены технические обновления бенчмарка на основе фреймворка jiant для консистентного обучения и оценки NLP-моделей различных архитектур, включая самые последние модели для русского языка. Помимо обновления основного бенчмарка, мы представляем интеграцию бенчмарка Russian SuperGLUE с фреймворком для промышленной оценки моделей с открытым исходным кодом – MOROCCO (MOdel ResOurCe COmparison), в котором модели оцениваются по средневзвешенной метрике всех заданий, скорости быстродействия и занимаемого объема оперативной памяти. Материалы Russian SuperGLUE доступны по адресу https://russiansuperglue.com/.

Ключевые слова: оценка моделей, понимание естественного языка, бенчмарки, NLP-модели

1 Introduction

In the last years, new architectures and methods for model pre-training and transfer learning have driven striking performance improvements across a range of language understanding tasks. Complex benchmark approaches are being developed for testing general intellectual "abilities" of NLP models on a wide range of natural language understanding (NLU) tasks. The tasks range from identifying causal relations in texts (NLI) to common sense, world knowledge, and logic. The central benchmarks in the field are GLUE [1] and SuperGLUE [2] projects for English, they include versatile tasks and allow competitive evaluation of the models on a public leaderboard. Recently, analogous general language understanding evaluation benchmarks have been developed for Chinese [3], French [4], Polish [5] and Russian [6]. RussianSuperGLUE provides nine novel Russian NLU tasks, a public leaderboard, count-based and transformer-based baselines, and human solver evaluation.

This work presents Russian SuperGLUE 1.1, a new release of the benchmark that provides multiple updates and improvements of the previous version. First, we updated the following datasets: 1) RUSSE: expansion of the dataset and construction of a novel test set; 2) DaNetQA: increasing the size of the dataset and creation of a new test set; 3) RuCoS: doubling the size of the validation and test sets, cleaning typos and inaccuracies; 4) MuSeRC: the expansion of the dataset, cleaning typos and inaccuracies. Second, we provide an improved Russian SuperGLUE toolkit based on `jiant` framework [7] for consistent training and evaluation of NLP models for Russian, which now supports the novel transformer-based models such as RuGPT[1]. Furthermore, we introduce an enhanced web interface of the benchmark that includes bug fixes and new features: the model evaluation by individual task (one can get the score for a specific task), a better notification procedure, and a new leaderboard based upon the model performance evaluation. Finally, Russian SuperGLUE has been integrated with MOROCCO, a framework for industrial evaluation of model performance. Models submitted to the leaderboard can be additionally estimated by inference speed and memory footprint.

The remainder is organized as follows. Section 2 briefly describes the benchmark tasks. Section 3 outlines the new release, namely the dataset updates and improvements of the leaderboard interface. Section 4 provides the description of MOROCCO framework and the performance evaluation metrics. We compare a number of novel models for Russian with English ones in Section 5 and conclude in Section 6.

2 Previous Work

Russian-based NLP-systems have a long history of benchmarking within various tasks. Starting with ROMIP Seminar in 2003[2], then Dialog Evaluation tracks starting from 2008[3] have continued the prolific

[1] https://github.com/sberbank-ai/ru-gpts/tree/master

[2] `http://romip.ru/ru/2003/index.html`

[3] `http://www.dialog-21.ru/evaluation/`

Task	Task Type	Task Metric	Train	Val	Test
TERRa	NLI	Accuracy	2616	307	3198
RCB	NLI	Avg. F1 / Accuracy	438	220	438
LiDiRus	NLI & diagnostics	MCC	0	0	1104
RUSSE	Common Sense	Accuracy	19845	8508	18892
PARus	Common Sense	Accuracy	400	100	500
DaNetQA	World Knowledge	Accuracy	1749	821	805
MuSeRC	Machine Reading	F1 / EM	500	100	322
RuCoS	Machine Reading	F1 / EM	72193	7 577	7257
RWSD	Reasoning	Accuracy	606	204	154

Table 1: Russian SuperGLUE task description. Train/Val/Test include number of samples for each set; MCC stands for Matthews Correlation Coefficient; EM - Exact Match.

tradition of yearly system evaluation on the most technically relevant problems, including morphological and syntactic parsing, text classification, spell check, named entity recognition, and many more. RUSSE'2018[4], word sense induction and disambiguation for the Russian shared task, is definitely worth mentioning as well. Last but not least, SberSQuAD [8] QA-system leaderboard completes a series of traditional single-task benchmarks.

Russian SuperGLUE benchmark first introduced a multi-task benchmark for Russian, providing a stable updated leaderboard with all the systems ranged by their average performance on 9 complex tasks.

2.1 Russian SuperGLUE Tasks

We continue our work on Russian SuperGLUE[5] [6] which follows the general language understanding evaluation methodology. Similarly to the English prototype, Russian benchmark includes a set of NLU tasks and a publicly available leaderboard. Namely, the benchmark comprises 9 tasks divided into 5 groups:

- **Textual Entailment & NLI**: TERRa, RCB, LiDiRus;
- **Common Sense**: RUSSE, PARus;
- **World Knowledge**: DaNetQA [9];
- **Machine Reading**: MuSeRC, RuCoS [10];
- **Reasoning**: RWSD.

Task Description We outline the information on the tasks by their type, metrics and partition sizes in Table 1.

TERRA Textual Entailment Recognition for Russian is aimed at capturing textual entailment in a binary classification form. Given two text fragments (premise and hypothesis), the task is to determine whether the meaning of the hypothesis is entailed from the premise. The dataset was sampled from the Taiga corpus [11].

RCB The Russian Commitment Bank is a 3-way classification task aimed at recognizing textual entailment (NLI). In contrast to TERRa, the premise in RCB may represent a textual segment rather than a single sentence. The corpus was filtered from Taiga with a number of pre-defined rules and labeled by crowd workers.

LiDiRus is a diagnostic set that tests models for a rich set of 33 linguistic features, commonsense, and world knowledge. The dataset was constructed as a translation from GLUE diagnostics with the preservation of all features. Thus, it provides an opportunity for evaluation of linguistic and semantic properties of language models in the setting of NLI task and for drawing comparisons between the

[4] https://russe.nlpub.org/2018/wsi/
[5] https://russiansuperglue.com/

languages.

RUSSE is a binary classification task that involves word sense disambiguation. Given a pair of sentences containing the same ambiguous word, the goal of the model is to recognize if the word is used in the same meaning. The dataset was constructed from RUSSE [12].

PARus is a binary classification task aimed to identify the most plausible alternative out of two for a given premise. It is a manually verified translation of the COPA dataset from SuperGLUE.

DaNetQA is a Russian QA dataset for yes/no questions which follows the BoolQ design [21]. Each sample consists of a Wikipedia paragraph and a human-generated question related to the paragraph. The task is to come up with a binary answer (yes or no) for the given question.

MuSeRC is a machine reading comprehension (MRC) task. Each sample consists of a text paragraph, multi-hop questions based on the paragraph, and possible answers for each question. The goal is to choose all correct answers for each question. The dataset was collected from publicly available sources across multiple domains (elementary school texts, news, summary of series, fiction stories, and fairy tales), and further annotated by crowd-workers.

RuCoS is an MRC task that involves commonsense reasoning and world knowledge. The dataset is a counterpart of ReCoRD [22] for English. Each example consists of a text paragraph, a query with a missing named entity, and a set of candidates for the answer. The task is to select one of the candidates that best fits the gap.

RWSD Russian Winograd Schema task is devoted to coreference resolution in a binary classification form. The corpus was created as a manually validated translation of the Winograd Schema Challenge[6].

3 New Release Features

Version 1.1 includes important methodological updates to the datasets, as well as the expansion of a number of "out-of-the-box" supported model architectures in the software, which is attached for the convenience of developers and a unified testing environment for all systems. Also, in addition to the main leaderboard, the model evaluation process was significantly supplemented by industrial metrics, which will be described below.

3.1 Improving the Tasks

The first version of the datasets has a number of drawbacks that revealed themselves after the initial release. We collected the feedback and fixed the shortcomings of the previous version of the benchmark. Among the main reasons for weaknesses are: 1) data leakage, 2) class distributions in test sets, 3) smaller size of the MRC datasets and the number of typos and inconsistencies in them. The latest problem is inevitable, RuCoS and MuSeRC are the most resources and time-consuming for collection and verification datasets - still their sizes were smaller than in their English analogs (ReCoRD and MultiRC respectively). As a result, four datasets were improved, for two of them (DaNetQA and RUSSE) completely new test sets were created. The results of the leaderboard were rescored - all baselines were measured again on the new datasets. Additionally, we asked the participants to resubmit on the new data. Thus, now only the latest version of the datasets and leaderboard are supported. In this section, we describe the procedure of the dataset improvements.

3.1.1 RUSSE

The update of RUSSE was motivated by extremely high scores of the language models which significantly outperformed the human benchmark.

For example, mBART [28] achieved almost 99% accuracy, while human performance was at the level of 75%. We believe that the conversion of the publicly available RUSSE test set to that of Russian SuperGLUE has possibly led to data leakage.

To this end, we constructed a completely new test set for the task in order to eliminate the potential leakage. First, we filtered the anchor words, discarded the most outdated and rare ones, and enriched the dataset with novel samples. Second, we collected sentences dating from the 2020 year using publicly

[6] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

available news sources such as Wikinews[7] and Lenta.ru[8]. Finally, we manually validated the meanings of the anchor words in the resulted sentences and annotated the answers. The total size of the novel test set is 18 892 examples.

For the obtained test set we re-scored the human benchmark using the same annotation procedure in Yandex.Toloka task as described in [6] but on the new subset of the data. The human performance achieved 80.5% accuracy, while the best model performance on the leaderboard 2 at present is 72.9% (RuBERT conversational).

3.1.2 DaNetQA

Originally, DaNetQA had a limited number of examples: 392, 295, 295 (train/val/test). We extended the dataset following the methodology described in [9], and converted a subset of MuSeRC into the yes/no QA setting, labeled by crowd-workers afterward. The new task contains 1750, 821, and 805 examples (train/val/test). In addition, we manually checked validation and test sets and balanced both sets by target class, as opposed to the previous version where the class distribution was 80/20% and changed the answer distribution by balancing the sets. The current class balance is 50/50% in contrary to the originally imbalanced data with 80% yes answers.

Since the test set has been changed completely we re-scored the performance of human solvers and models. While human performance gets 91 of accuracy score for the updated dataset, the language models (see table 2) are not greater than 65,7% (not 80% as it was before).

3.1.3 RuCoS

The new version of RuCoS involves the following updates. We doubled the size of the validation (7527 examples) and test (7257 examples) sets as described in [10]. We manually verified the crowd-worker annotations and corrected typos and annotation inconsistencies. Since the human performance was assessed on a subset of the test set, the results remain the same. The best-performing model is now RuGPT3-XL over a few-shot technique.

3.1.4 MuSeRC

As opposed to the English analogue MultiRC, MuSeRC was relatively small in size which we aimed to improve. However, MuSeRC consists only of multi-hop questions which makes the tasks more difficult in contrast to MultiRC which also includes one-hop questions. We extended the train set with more than 300 new samples containing novel multi-hop questions. As a result, the size of MuSeRC became comparable with MultiRC as the number of multi-hop questions is 5,228 and 5,825 respectively. Thus, in the new Russian SuperGLUE release we: expanded the train set; cleaned typos, grammar mistakes, and text inaccuracies in all the samples.

3.2 Infrastructure Advances

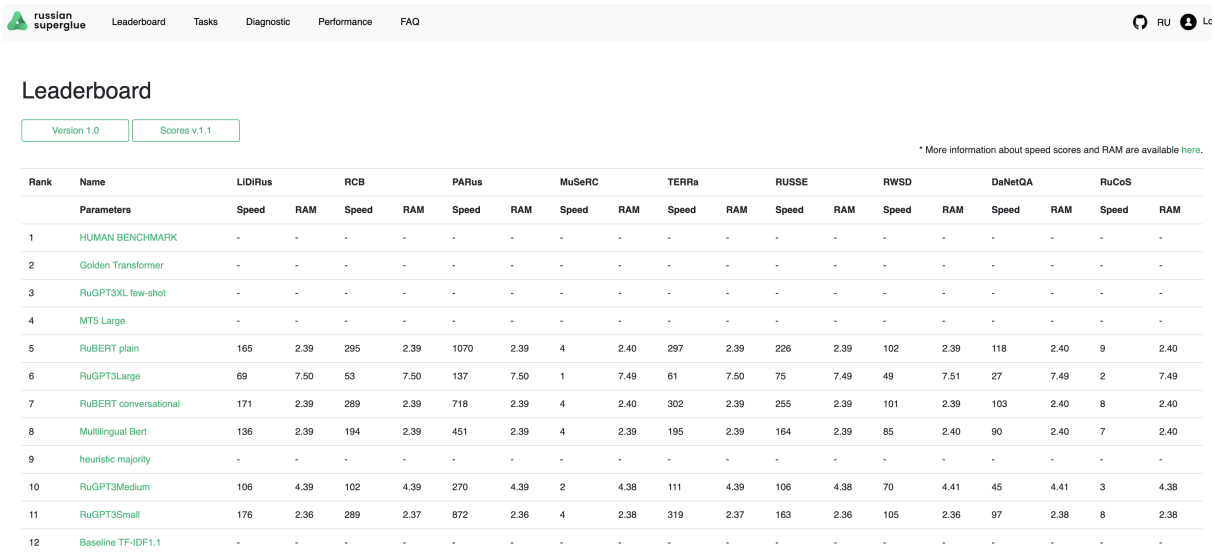
The interaction between the leaderboard participants and the benchmark interface is crucially important as the entry usage threshold directly affects the user experience and submission quantity. To this end, the new release contains bug fixes and presents new features of the leaderboard interface and infrastructure.

First, we improved the reliability of the model evaluation system on the website. We made it more strict and fixed some minor bugs. Furthermore, we enhanced the web interface and added new features: 1) The user can download their previous submissions and edit them; 2) The user can evaluate their model and upload submission both on a single task and the full set of tasks; 3) The user receives two email notifications after they make their submission public. The first one confirms the submission verification step, and the second one informs whether the submission was published on the leaderboard or rejected (and why); 4) The user guide is updated to provide a better leaderboard usage experience; 5) The user now can evaluate their model by industrial performance metrics, namely inference speed and memory footprint which we describe in Section 4. The performance leaderboard is developed as well (see Figure 1).

[7] https://www.wikinews.org/

[8] https://lenta.ru/

Besides, Russian SuperGLUE 1.1 involves minor bug fixes along with the support of the novel models for Russian: RuGPT3 models[9] included in the list of models by HuggingFace library[10].



Rank	Name	LiDiRus		RCB		PARus		MuSeRC		TERRa		RUSSE		RWSD		DaNetQA		RuCoS	
	Parameters	Speed	RAM	Speed	RAM	Speed	RAM	Speed	RAM	Speed	RAM	Speed	RAM	Speed	RAM	Speed	RAM	Speed	RAM
1	HUMAN BENCHMARK	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	Golden Transformer	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	RuGPT3XL few-shot	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	MT5 Large	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	RuBERT plain	165	2.39	295	2.39	1070	2.39	4	2.40	297	2.39	226	2.39	102	2.39	118	2.40	9	2.40
6	RuGPT3Large	69	7.50	53	7.50	137	7.50	1	7.49	61	7.50	75	7.49	49	7.51	27	7.49	2	7.49
7	RuBERT conversational	171	2.39	289	2.39	718	2.39	4	2.40	302	2.39	255	2.39	101	2.39	103	2.40	8	2.40
8	Multilingual Bert	136	2.39	194	2.39	451	2.39	4	2.39	195	2.39	164	2.39	85	2.40	90	2.40	7	2.40
9	heuristic majority	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	RuGPT3Medium	106	4.39	102	4.39	270	4.39	2	4.38	111	4.39	106	4.38	70	4.41	45	4.41	3	4.38
11	RuGPT3Small	176	2.36	289	2.37	872	2.36	4	2.38	319	2.37	163	2.36	105	2.36	97	2.38	8	2.38
12	Baseline TF-IDF1.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 1: Performance evaluation leaderboard in RussianSuperGLUE benchmark.

4 Computational Efficiency Evaluation

A number of closely related benchmarks consider only the task-specific performance of the models, leaving the computational efficiency unattended. To this end, Russian SuperGLUE is now integrated with MOROCCO [31], a framework for industrial evaluation of model performance by the following metrics: *memory footprint* and *inference speed* (see Section 4.1). The user can evaluate their model by submitting a Docker container which is expected to read the text from the standard input and channel the predictions to standard output. The container is run in a fixed isolated environment with limited running time, RAM, and CPU/GPU resources. We use Yandex.Cloud[11] platform where the following hardware is provided: 1 $\times$ Intel Broadwell CPU, 1 $\times$ NVIDIA Tesla V100 GPU. The Docker container OS is Ubuntu 20.04. The solution is run over several iterations to eliminate the dispersion, with the median values further computed. Along with the metrics, we also compute the task-specific metric based upon the Docker output to further aggregate the results into a final score for the submission.

4.1 Industrial Metrics

Memory footprint, or GPU RAM usage M is measured by running a Docker container with a single record as input and measuring the maximum M. We repeat the procedure 5 times and take the median value.

Inference speed, or *throughput* Tp is computed by running a Docker container with N records as input and optional batch size to measure T_N. Besides, we estimate the model initialization time T_{init} by running the container with a single record as input. The resulting *throughput* is computed as follows[12]:

$$Tp = \frac{N}{T_N - T_{\text{init}}}.$$

[9] RuGPT3-Small: `https://huggingface.co/sberbank-ai/rugpt3small_based_on_gpt2`, RuGPT3-Medium: `https://huggingface.co/sberbank-ai/rugpt3medium_based_on_gpt2`, RuGPT3-Large: `https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2`, RuGPT3-XL: `https://huggingface.co/sberbank-ai/rugpt3xl`

[10] `https://github.com/huggingface/transformers`

[11] `https://cloud.yandex.com/`

[12] During the evaluation process, the $N = 2000$ is used and the batch size is 32. We repeat the procedure 5 times and compute the median value.

We propose to use these three characteristics, namely Q, Tp, and M, in the following way: we comprise a 2-dimensional plot with the horizontal axis being a quality for a downstream task Q (this metric is specific to the task) and vertical axis being a throughput Tp for the model. To visualize memory footprint M, we propose to use circles of different sizes instead of a mere point on the plot. The scores that involve industrial performance metrics for the models from Russian SuperGLUE leaderboard are presented in Figure 2.

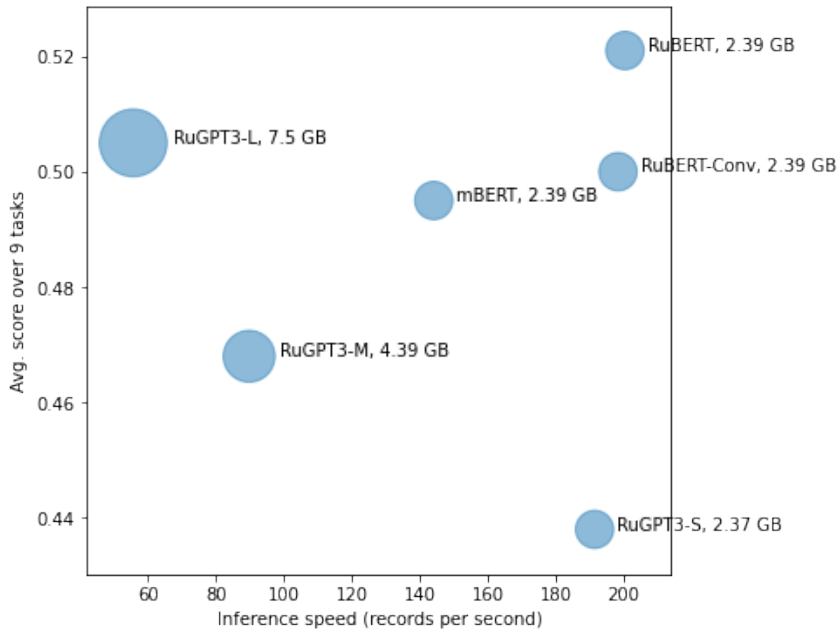


Figure 2: Models comparison on Russian SuperGLUE benchmark.

MOROCCO allows computing the performance by running the containers in other technical environments that best fit the user needs. Figure 2 presents an unexpected result: the dilemma of choosing between speed and performance for the Russian language is not at all a dilemma - the highest quality models are also the fastest (certainly, except for the simplest solutions and baselines like TF-IDF).

5 Results and Discussion

The current level of systems participating in the above-described language tests and competitions has certainly grown over the past decades: although the most common text-based tests are less than a dozen years old, we can trace the development trends and system evaluations based on simpler, unchanged technical criteria for their growth. The benchmark approach to the assessment of intelligent systems is currently dominant, allowing to combine the assessment of various intellectual abilities under the cumulative assessment of general intelligence. Intellectual tests, expressed through texts, constitute the main productive method of such an assessment, making it possible to formulate a variety of types of tasks and compare the level of systems with human intelligence, including the formation of sets of examples of tasks, to successfully solve skills or abilities that are not lower than human, but which do not have clear definitions within the framework of neuroscience: common sense, goal-setting, cause-and-effect relationships, knowledge about the world. The current version of the leaderboard version 1.1 is shown in Figure 3.

The existing problems of benchmark approaches, however, are subject to close research by the community. The discussion is provoked by a significant difference in the level of metrics observed in Russian SuperGLUE and in the English-language leaderboard. The best result so far is 67.9% overall for Russian (while the human level is 81.1%), while for English the best model performance is 90.4%, beating the human score of 89.8%.

A separate subject of discussion is the issue of limitations of the presented leaderboard and its methodological analogues in other languages. While the tasks set in the benchmark themselves are designed to test the human intellectual abilities or their imitation, we see that in some cases (for example, the English SuperGLUE), a result higher than human has already been achieved in a completely mechanical approach, using transformer models pretrained on large corpora and fine-tuned on accumulated task-specific data. At the time of this writing, the best result among Russian-based models was raised by 15% using simplistic ML-hackathon methods - model ensembling, automatic translation and training a meta-classifier for weighting the models in various tasks.

The growing popularity of multilingual benchmarks is promising: the extension of the testing methodology has led to a comprehensive multilingual assessment - the XTREME [29] and XGLUE [30] projects have combined the available materials for testing systems to their ability to reproduce human intellectual abilities on 40 and 19 typologically diverse languages accordingly. The Russian language is included as a part of the test material in both of these benchmarks, which is a promising prospect for the further integration of project data into multilingual X-benchmarks with both testing and training data provided. We anticipate that the speed of benchmark hacking will become lower than the speed of creating weighted, complex benchmarks if they strictly evaluate the modelling of the language and separate the language itself from all distinct abilities expressed with the language.

Rank	Name	Team	Link	Score	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
1	HUMAN BENCHMARK	AGI NLP	i	**0.811**	0.626	0.68 / 0.702	0.982	0.806 / 0.42	0.92	0.805	0.84	0.915	0.93 / 0.89
2	Golden Transformer	Avengers Ensemble	i	**0.679**	0.0	0.406 / 0.546	0.908	0.941 / 0.819	0.871	0.587	0.545	0.917	0.92 / 0.924
3	RuGPT3XL few-shot	sberdevices	i	**0.535**	0.096	0.302 / 0.418	0.676	0.74 / 0.546	0.573	0.565	0.649	0.59	0.67 / 0.665
4	MT5 Large	AGI NLP	i	**0.528**	0.061	0.366 / 0.454	0.504	0.844 / 0.543	0.561	0.633	0.669	0.657	0.57 / 0.562
5	RuBERT plain	DeepPavlov	i	**0.521**	0.191	0.367 / 0.463	0.574	0.711 / 0.324	0.642	0.726	0.669	0.639	0.32 / 0.314
6	RuGPT3Large	sberdevices	i	**0.505**	0.231	0.417 / 0.484	0.584	0.729 / 0.333	0.654	0.647	0.636	0.604	0.21 / 0.202
7	RuBERT conversational	DeepPavlov	i	**0.5**	0.178	0.452 / 0.484	0.508	0.687 / 0.278	0.64	0.729	0.669	0.606	0.22 / 0.218
8	Multilingual Bert	DeepPavlov	i	**0.495**	0.189	0.367 / 0.445	0.528	0.639 / 0.239	0.617	0.69	0.669	0.624	0.29 / 0.29
9	heuristic majority	ling_ling	i	**0.468**	0.147	0.4 / 0.438	0.478	0.671 / 0.237	0.549	0.595	0.669	0.642	0.26 / 0.257
10	RuGPT3Medium	sberdevices	i	**0.468**	0.01	0.372 / 0.461	0.598	0.706 / 0.308	0.505	0.642	0.669	0.634	0.23 / 0.224
11	RuGPT3Small	sberdevices	i	**0.438**	-0.013	0.356 / 0.473	0.562	0.653 / 0.221	0.488	0.57	0.669	0.61	0.21 / 0.204
12	Baseline TF-IDF1.1	AGI NLP	i	**0.434**	0.06	0.301 / 0.441	0.486	0.587 / 0.242	0.471	0.57	0.662	0.621	0.26 / 0.252
13	Random weighted	ling_ling	i	**0.385**	0.0	0.319 / 0.374	0.48	0.45 / 0.071	0.483	0.528	0.597	0.52	0.25 / 0.247

Figure 3: Models comparison on RussianSuperGLUE benchmark.

6 Conclusion

We present Russian SuperGLUE v1.1, an updated benchmark for evaluating general-purpose language understanding systems for Russian. As part of the development of the project, the tasks were updated:

- RUSSE (understanding word meaning in context) - a new test set has been compiled, the possibility of a dataset leak is excluded;
- DaNetQA (yes / no questions based on commonsense reasoning and machine reading) - a new test set was compiled, the composition of classes in the dev and test set was balanced;
- RuCoS (machine-reading and commonsense reasoning) - task dataset expanded and manually corrected, the composition of the classes in the dev and test set has been balanced;
- MuSeRC (machine-reading and information retrieval) - task dataset expanded and manually corrected.

A prominent direction for future work is to expand the existing tasks, provide the support for upcoming models for Russian, and improve the user experience of the MOROCCO framework, specifically by supporting models not released as a part of HuggingFace library such as ELMo[13]. Overall, we believe that Russian SuperGLUE provides the research community with a challenging frontier and further natural language understanding progress for Russian.

[13] http://docs.deeppavlov.ai/en/master/features/pretrained_vectors.html#elmo

References

[1] Wang, Alex, et al. “GLUE: A multi-task benchmark and analysis platform for natural language understanding.” arXiv preprint arXiv:1804.07461 (2018).

[2] Wang, Alex, et al. “Superglue: A stickier benchmark for general-purpose language understanding systems.” arXiv preprint arXiv:1905.00537 (2019).

[3] Xu L. et al. Clue: A chinese language understanding evaluation benchmark //arXiv preprint arXiv:2004.05986. – 2020.

[4] Le H. et al. Flaubert: Unsupervised language model pre-training for french //arXiv preprint arXiv:1912.05372. – 2019.

[5] Rybak P. et al. KLEJ: comprehensive benchmark for polish language understanding //arXiv preprint arXiv:2005.00630. – 2020.

[6] Shavrina, Tatiana, et al. “RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark.” arXiv preprint arXiv:2010.15925 (2020).

[7] Pruksachatkun, Yada, et al. "jiant: A software toolkit for research on general-purpose text understanding models." arXiv preprint arXiv:2003.02249 (2020).

[8] Efimov P. et al. SberQuAD–Russian reading comprehension dataset: Description and analysis //International Conference of the Cross-Language Evaluation Forum for European Languages. – Springer, Cham, 2020. – pp. 3-15.

[9] Glushkova, Taisia, et al. “DaNetQA: a yes/no Question Answering Dataset for the Russian Language.” arXiv preprint arXiv:2010.02605 (2020).

[10] Fenogenova, Alena, Vladislav Mikhailov, and Denis Shevelev. “Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian.” Proceedings of the 28th International Conference on Computational Linguistics. 2020.

[11] Shavrina, Tatiana, and Olga Shapovalova. “To the methodology of corpus construction for machine learning: “Taiga” syntax tree corpus and parser.” Proceedings of “CORPORA-2017” International Conference. 2017.

[12] Panchenko, Alexander, et al. “RUSSE’2018: a shared task on word sense induction for the Russian language.” arXiv preprint arXiv:1803.05795 (2018).

[13] Sharoff, Serge, and Joakim Nivre. “The proper place of men and machines in language technology." Processing Russian without any Linguistic Knowledge. Computational Linguistics and Intelligent Technologies 10.17 (2011): 657-670.

[14] Conneau, Alexis, and Douwe Kiela. “Senteval: An evaluation toolkit for universal sentence representations.” arXiv preprint arXiv:1803.05449 (2018).

[15] McCann, Bryan, et al. “The natural language decathlon: Multitask learning as question answering.” arXiv preprint arXiv:1806.08730 (2018).

[16] Eichler, Max, Gözde Gül Şahin, and Iryna Gurevych. “LINSPECTOR WEB: A multilingual probing suite for word representations.” arXiv preprint arXiv:1907.11438 (2019).

[17] Şahin, Gözde Gül, et al. “Linspector: Multilingual probing tasks for word representations.” Computational Linguistics 46.2 (2020): 335-385.

[18] Wolf, Thomas, et al. “HuggingFace’s Transformers: State-of-the-art natural language processing.” arXiv preprint arXiv:1910.03771 (2019).

[19] Gauen, Kent, et al. “Low-power image recognition challenge.” 2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC). IEEE, 2017.

[20] Panchenko, Alexander, et al. “Russe: The first workshop on russian semantic similarity.” arXiv preprint arXiv:1803.05820 (2018).

[21] Clark, Christopher, et al. “BoolQ: Exploring the surprising difficulty of natural yes/no questions.” arXiv preprint arXiv:1905.10044 (2019).

[22] Zhang, Sheng, et al. “Record: Bridging the gap between human and machine commonsense reading comprehension.” arXiv preprint arXiv:1810.12885 (2018).

[23] Dagan, Ido, Oren Glickman, and Bernardo Magnini. “The pascal recognising textual entailment challenge.” Machine Learning Challenges Workshop. Springer, Berlin, Heidelberg, 2005.

[24] Haim, R. Bar, et al. “The second pascal recognising textual entailment challenge.” Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment. 2006.

[25] Giampiccolo, Danilo, et al. “The third pascal recognizing textual entailment challenge.” Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing. 2007.

[26] Bentivogli, Luisa, et al. “The Fifth PASCAL Recognizing Textual Entailment Challenge.” TAC. 2009.

[27] Min, Sewon, et al. “NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned.” arXiv preprint arXiv:2101.00133 (2021).

[28] Liu, Yinhan, et al. "Multilingual denoising pre-training for neural machine translation." Transactions of the Association for Computational Linguistics 8 (2020): 726-742.

[29] Hu J. et al. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation //International Conference on Machine Learning. – PMLR, 2020. – pp. 4411-4421.

[30] Liang Y. et al. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation //arXiv preprint arXiv:2004.01401. – 2020.

[31] V.Malykh, A.Kukushkin, E.Artemova, V.Mikhailov, M.Tikhonova, T.Shavrina. MOROCCO: Model Resource Comparison Framework. //arXiv preprint arXiv:2104.14314. – 2021.

7 Appendix

7.1 Appendix 1. Russian SuperGLUE Leaderboard

Table 2: Russian SuperGLUE v 1.1 Leaderboard

Rank	Name	Score	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
1	HUMAN BENCHMARK	0.811	0.626	0.68 / 0.702	0.982	0.806 / 0.42	0.92	0.805	0.84	0.915	0.93 / 0.89
2	RuGPT3XL few-shot	0.535	0.096	0.302 / 0.418	0.676	0.74 / 0.546	0.573	0.565	0.649	0.59	0.67 / 0.665
3	MT5 Large	0.528	0.061	0.366 / 0.454	0.504	0.844 / 0.543	0.561	0.633	0.669	0.657	0.57 / 0.562
4	RuBERT plain	0.521	0.191	0.367 / 0.463	0.574	0.711 / 0.324	0.642	0.726	0.669	0.639	0.32 / 0.314
5	RuGPT3Large	0.505	0.231	0.417 / 0.484	0.584	0.729 / 0.333	0.654	0.647	0.636	0.604	0.21 / 0.202
6	RuBERT conv	0.5	0.178	0.452 / 0.484	0.508	0.687 / 0.278	0.64	0.729	0.669	0.606	0.22 / 0.218
7	mBert	0.495	0.189	0.367 / 0.445	0.528	0.639 / 0.239	0.617	0.69	0.669	0.624	0.29 / 0.29
8	heuristic majority	0.468	0.147	0.4 / 0.438	0.478	0.671 / 0.237	0.549	0.595	0.669	0.642	0.26 / 0.257
9	RuGPT3 Medium	0.468	0.01	0.372 / 0.461	0.598	0.706 / 0.308	0.505	0.642	0.669	0.634	0.23 / 0.224
10	RuGPT3 Small	0.438	-0.013	0.356 / 0.473	0.562	0.653 / 0.221	0.488	0.57	0.669	0.61	0.21 / 0.204
11	Baseline TF-IDF1.1	0.434	0.06	0.301 / 0.441	0.486	0.587 / 0.242	0.471	0.57	0.662	0.621	0.26 / 0.252
12	Random weighted	0.385	0.0	0.319 / 0.374	0.48	0.45 / 0.071	0.483	0.528	0.597	0.52	0.25 / 0.247
13	majority class	0.374	0.0	0.217 / 0.484	0.498	0.0 / 0.0	0.513	0.587	0.669	0.503	0.25 / 0.247

Traditional Machine Learning and Deep Learning Models for Argumentation Mining in Russian Texts

Fishcheva I. N.
Vyatka State University,
Kirov, Russia
fishchevain@gmail.com

Goloviznina V. S.
Vyatka State University,
Kirov, Russia
golovizninavs@gmail.com

Kotelnikov E. V.
Vyatka State University,
Kirov, Russia;
ITMO University,
Saint Petersburg, Russia
kotelnikov.ev@gmail.com

Abstract

Argumentation mining is a field of computational linguistics that is devoted to extracting from texts and classifying arguments and relations between them, as well as constructing an argumentative structure. A significant obstacle to research in this area for the Russian language is the lack of annotated Russian-language text corpora. This article explores the possibility of improving the quality of argumentation mining using the extension of the Russian-language version of the Argumentative Microtext Corpus (ArgMicro) based on the machine translation of the Persuasive Essays Corpus (PersEssays). To make it possible to use these two corpora combined, we propose a Joint Argument Annotation Scheme based on the schemes used in ArgMicro and PersEssays. We solve the problem of classifying argumentative discourse units (ADUs) into two classes – "pro" ("for") and "opp" ("against") using traditional machine learning techniques (SVM, Bagging and XGBoost) and a deep neural network (BERT model). An ensemble of XGBoost and BERT models was proposed, which showed the highest performance of ADUs classification for both corpora.

Keywords: argumentation mining; machine translation; deep learning; BERT
DOI: 10.28995/2075-7182-2021-20-246-258

Модели традиционного машинного обучения и глубокого обучения для анализа аргументации русскоязычных текстов

Фищева И. Н.
Вятский государственный университет,
Киров, Россия
fishchevain@gmail.com

Головизнина В. С.
Вятский государственный университет,
Киров, Россия
golovizninavs@gmail.com

Котельников Е. В.
Вятский государственный университет,
Киров, Россия;
Университет ИТМО,
Санкт-Петербург, Россия
kotelnikov.ev@gmail.com

Аннотация

Анализ аргументации – это область компьютерной лингвистики, которая посвящена извлечению из текстов и классификации аргументов и связей между ними, а также построению аргументационной структуры. Существенным препятствием исследованиям в этой области для русского языка является недостаток аннотированных русскоязычных текстовых корпусов. В настоящей статье исследуется возможность повышения качества анализа аргументации при помощи расширения русскоязычной версии Argumentative Microtext Corpus (ArgMicro) на основе машинного перевода Persuasive Essays Corpus (PersEssays). Для возможности совместного применения двух корпусов мы предлагаем объединенную схему разметки на основе схем, используемых в ArgMicro и PersEssays. Мы решаем задачу классификации аргументативных дискурсивных единиц (ADUs) на два класса – "за" и "против" с использованием традиционных методов машинного обучения (SVM, Bagging

и XGBoost) и глубокой нейросетевой модели BERT. Был предложен ансамбль моделей XGBoost и BERT, который и показал наивысшее качество классификации ADUs для обоих корпусов.

Ключевые слова: анализ аргументации; машинный перевод; глубокое обучение; BERT

1 Introduction

Argumentation (or argument) mining is a field of computational linguistics that is devoted to extracting from texts and classifying arguments and relations between them, as well as constructing an argumentation structure [16], [19]. This area is seeing an influx of research activity – for example, since 2014, seven workshops on the analysis of arguments have already been held[1]. Besides being of academic interest, argumentation mining is in the focus of attention due to a wide range of applications, in particular, when studying user opinions based on social media analysis [1], [17], analyzing legal texts [18], scientific texts [9], political debates [25], news articles [3] and student essays [29].

The main text element used in the argumentation mining is an argumentative discourse unit (ADU) – a piece of text that has a single argumentation value [30, p. 63]. As a rule, ADUs are most often individual sentences, but in some cases ADU is a part of a sentence or several sentences.

In ADU-based argumentation mining, the tasks are as follows [30, p. 6]:

1) identifying text fragments containing argumentation;
2) segmenting the text into ADUs;
3) identifying the central (or major) claim (usually among ADUs; but there can also be implicit central claims);
4) classification of ADUs – the main classes are supporting and rebutting ADUs;
5) establishing relations between ADUs;
6) building an argumentation structure;
7) assessing the argumentation quality.

There is also a stance detection task, related to the argumentation mining. This task is to determine the point of view of the text's author and is often solved independently, without identifying arguments [13].

To successfully solve the above mentioned tasks, annotated text corpora are required. Currently, there is a fairly large number of corpora with a variety of argumentative annotation – Lawrence and Reed [16] estimate the known corpora at 2.2 million words. The largest open database of text corpora with argumentative annotation is AifDB[2] [15], which contains more than 14,000 texts. However, most of these corpora are in English.

Fishcheva and Kotelnikov [7] showed that the machine translation of the English-language Argumentative Microtext Corpus (ArgMicro) [24], [27] into Russian allows obtaining the performance of ADUs classification that is not inferior to human translation. In this paper, following [7], we investigate the possibility of improving the performance of ADUs classification based on the extension of the Russian version of the ArgMicro corpus through machine translation of the Persuasive Essays Corpus (PersEssays) [29]. To classify ADUs, we use traditional machine learning techniques (Support Vector Machines – SVM, Bagging and Gradient Boosting – XGBoost implementation), the deep neural network (BERT model [4]), and the XGBoost and BERT ensemble.

When considering argument annotated corpora combined, one of the important problems is the difference in annotation schemes [16]. We propose the Joint Argument Annotation Scheme based on those used in ArgMicro and PersEssays.

The contribution of this paper is as follows:

- the Joint Argument Annotation Scheme that takes into account the peculiarities of ArgMicro and PersEssays annotation schemes is proposed;
- a new Russian-language corpus with argumentative annotation is created. This corpus is formed by expanding the existing Russian-language version of the ArgMicro corpus with the machine translation version of PersEssays corpus. The new corpus is made publicly available;

[1] https://argmining2020.i3s.unice.fr.

[2] http://corpora.aifdb.org.

- for the new corpus the performance scores of the ADUs classification into two classes – “pro” and “opp” were obtained based on the traditional machine learning techniques (SVM, Bagging and XGBoost), the neural network model (BERT), as well as the XGBoost and BERT ensemble;
- the effect of expanding the training dataset and the influence of various groups of features on the classification performance was investigated.

The paper is structured as follows. The second section provides an overview of previous work, including existing argument annotation schemes, papers on cross-lingual argumentation mining and argumentation mining for the Russian language. The third section describes the proposed Joint Argument Annotation Scheme. The fourth section is devoted to text corpora and machine learning models for argumentation mining used in this work. In the fifth section the experimental results are presented and discussed. The sixth section provides conclusions and suggests directions for further research.

2 Previous work

In this section, firstly, the existing argument annotation schemes are considered, then papers in the field of cross-lingual argumentation mining are given, in conclusion, papers on the Russian-language argumentation mining are indicated.

2.1 Argument annotation schemes

The well-known argument annotation schemes are described in [30], as well as in [20]. Almost every corpus uses its own version of the annotation scheme, since different goals were laid down when creating the corpus.

The microtext scheme is based on Freeman's theory [8] and is described in detail by Peldszus and Stede [23]. This scheme was used in the annotation of the ArgMicro corpus [24]. In the microtext scheme, an argument structure is seen as a collection of multiple interconnected arguments. A claim (or conclusion) can be supported by premises or attacked by counterarguments. The main types of relations within this scheme are shown in Figure 1.

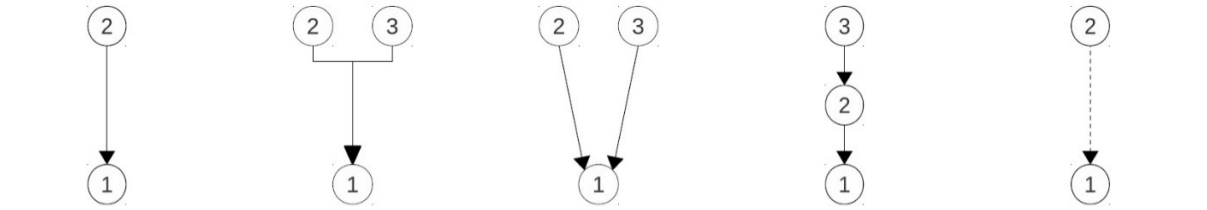


Figure 1: Main types of relations in the microtext scheme [23]

Figure 1 shows the following relations:

- basic argument – one claim is supported by one premise;
- linked support – premises are to be connected before conclusion;
- multiple support – two arguments converging to the same argument;
- serial support – argument can be the premise and the conclusion simultaneously: (2) is a premise for (1), and (2) is a conclusion for (3);
- example support – supporting argument, which is the example.

There are rebutting arguments in this annotation that attack the conclusion or premise. The relation between any arguments can also be questioned (undercut). In this case, rebutting arguments can attack other rebutting arguments.

The persuasive essay scheme was used to mark up the PersEssays corpus and is described in detail by Stab and Gurevych [29]. This annotation scheme includes argument components and argumentative relations between the components. One of the argument components called *Major Claim* – it expresses

the stance of the essay's author on the topic under discussion. *Claim* arguments support or attack the *Major Claim*. An attribute of the *Claim* ("for" or "against") indicates the polarity related to the *Major Claim*. *Premise* arguments support or attack the *Claims* or another *Premise*. The argumentative relations are defined by the discourse structure.

The science scheme is annotation scheme on a fine-grained level in scientific journal articles from the educational domain [11]. It uses a graph of arguments, which links support, attack, detail, and the undirected sequence relations.

The Modified Toulmin Scheme. Habernal and Gurevych [10] used the Toulmin model as a basis [32]. They analyzed user-generated web discourse. Within this scheme, there is no need to explicitly annotate any relations between the nodes.

The Cornell eRulemaking Scheme. Niculae et al. [21] considered a corpus of user comments on government rule making. It turned out that a lot of comments in the corpus could not be tagged using the microtext or the persuasive essay schemes. Therefore, specific argumentative role labels and new relation types were introduced.

Within our study, we use two existing corpora – ArgMicro and PersEssays. To use them together, we propose the Joint Argument Annotation Scheme based on microtext and persuasive essay schemes.

2.2 Cross-lingual argumentation mining

Currently, due to quite a large array of English-language text corpora, annotated by argumentation, and at the same time, the lack of such corpora for other languages, a range of works has emerged where the argumentation annotation of a corpus in one language is used to annotate a corpus in another language.

The current situation, when there is a sufficiently large array of English-language text corpora, annotated by argumentation, and at the same time, there is a lack of such corpora for other languages, has led to the emergence of a range of works on transfer of the argumentation annotation from a corpus in one language to a corpus in another language.

Aker and Zhang [2] created the first annotated Chinese corpus using the existing English corpora and manually matching claims and premises with parallel Chinese texts.

Eger et al. [5] created corpora in German, French, Spanish, and Chinese using human and machine translations of the PersEssays corpus. Eger et al. also compare the annotation projection and direct transfer strategy.

Sliwa et al. [28] created the first annotated corpus for Arabic and the Balkan language group using parallel corpora and annotation transfer of the English version of the corpus based on classifier training.

Eger et al. [6] examined cross-lingual transfer solving two tasks: sentence-level argumentation mining and automatic morphological tagging. They combined two cross-lingual approaches – direct transfer and projection, eliminating the shortcomings of both methods and combining their strengths.

Toledo-Ronen et al. [31] explored the potential of transfer learning using the multilingual BERT model for argumentation mining in non-English languages, based on English datasets and machine translation. They fine-tuned BERT for argumentation mining tasks using training on a corpus that includes both the original English-language texts and those translated into several languages.

In our study, in contrast to [2] and [28], we obtained performance scores of ADUs classification into two classes – "pro" and "opp" based on traditional machine learning techniques (SVM, Bagging and XGBoost), neural network model (BERT) and their ensemble. In contrast to [5] and [6], we explore the effect of expanding the training dataset and the importance of various groups of features. Unlike [31], we work with the BERT version for one language (Russian) and classify corpora in one language (Russian).

2.3 Argumentation mining in Russian

There is very little research on argumentation mining for the Russian language, as opposed to English.

Fishcheva and Kotelnikov [7] created the first annotated corpus for the Russian language based on the translation of the ArgMicro corpus. Also, an automated classification of the "pro" and "opp" sentences was carried out.

Kononenko et al. [12] studied argumentation using the comparative analysis of discourse structures. Various types of argument structures were considered. In order to automatically extract argumentative

relations, the analysis of the rhetorical and argumentative annotations was carried out. The experiment was carried out on a corpus of 11 popular science articles from Ru-RSTreebank.

Salomatina et al. [26] described a combined approach to partial extraction of the argumentative structure of text, which can be used if there are no sufficient annotated data to effectively apply machine learning techniques for the direct detection of arguments and their relationships.

In this paper, we develop the approach proposed in [7]. We expand the existing Russian-language corpus with argumentation annotation based on machine translation of the persuasive essays corpus. In contrast to [7], to classify ADUs in the new extended version of the corpus, we use a deep neural network (BERT model) along with traditional machine learning techniques, and also explore the effect of expanding the training dataset.

3 Joint Argument Annotation Scheme

The Join Argument Annotation Scheme (JAAS) was developed to enable combined processing of ArgMicro and PersEssays corpora. This annotation scheme is based on those used in ArgMicro and PersEssays.

There are three types of objects in the argument annotation schemes: a topic, a node and an edge [30].

1. The topic is a matter dealt with in a text; with or without indicating point of view (stance).
2. The node is a vertex of the argumentation graph. There are three types of nodes: major claims, regular nodes and neutral nodes.
    - The major claim (“mcl”) is a node of the argumentation graph which expresses some point of view related to the topic (conclusion). There may be one (ArgMicro) or two (PersEssays) nodes labeled as “mcl”. If there are two major claims then both reflect the same stance.
    - Regular nodes are the nodes of the argumentation graph which provide arguments (“pro” or “opp”) related to the major claim.
    - Neutral nodes are the sentences which are not members of the argumentation graph in Persuasive Essays Corpus.
3. The edge is a unit which determines a connection between two nodes. There are five types of edges:
    - support (“sup”) – a source node supports a target node;
    - additional support (“add”) – two or more source nodes support a target node only if they are taken together;
    - example (“exa”) – a source node serves as an example of the support of a target node;
    - rebuttal (“reb”) – a source node rebuts a target node;
    - undercut (“und”) – a source node attacks the connection (edge) between some two nodes.

In the ArgMicro annotation scheme, the graph nodes represent the propositions: the proponent’s nodes and the opponent’s nodes. The edges connecting the nodes represent different supporting and attacking moves.

In PersEssays annotation scheme, the sentences are classified as major claims, claims, premises and neutral. The PersEssays annotation scheme has unlabeled connections. We convert unlabeled edge types of PersEssays into three types of edges: “sup”, “reb” and “exa”. Types of edges in ArgMicro are more variable. Thus, types of edges in JAAS are equivalent to ArgMicro edges.

For illustration purposes, Figures 2 and 3 give an example of graph conversion from ArgMicro and PersEssays to JAAS.

4 Materials and Methods

4.1 Text corpora

Within this study, we used the Argumentative Microtext Corpus (ArgMicro) [24], [27] and the Persuasive Essay Corpus (PersEssays) [29]. Fishcheva and Kotelnikov [7] showed that the best result among

the Google Translate, Yandex.Translate and Promt systems was demonstrated by Google Translate in the machine translation of the ArgMicro corpus in English into Russian. Therefore, the PersEssays corpus was also translated into Russian using Google Translate[3]. Then the annotations of both corpora were converted to JAAS. A specific issue when converting PersEssays annotation to JAAS was identifying the "example" ("exa") edge types. To address this issue a two-stage procedure was used. At the first stage, an automatic search was carried out for ADUs containing template phrases such as "for example", "for instance", etc. At the second stage, the selected ADUs were manually checked. If the presence of the "example" relation type in the target corpus is not essential, this procedure can be omitted.

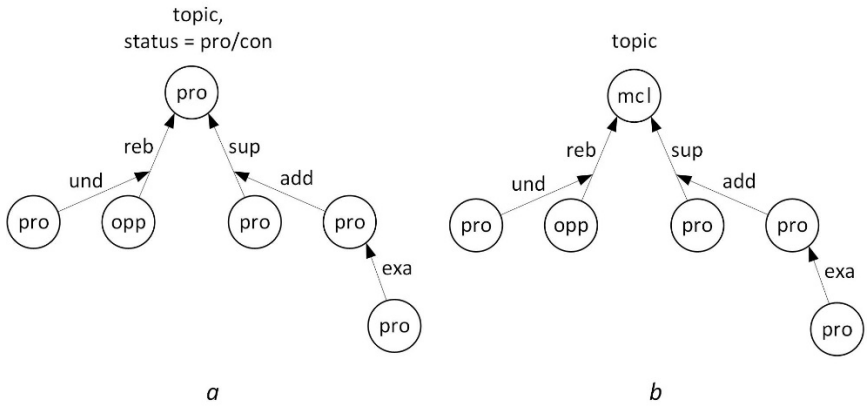


Figure 2: Equivalent graph representation of argumentation structure: a – ArgMicro; b – JAAS

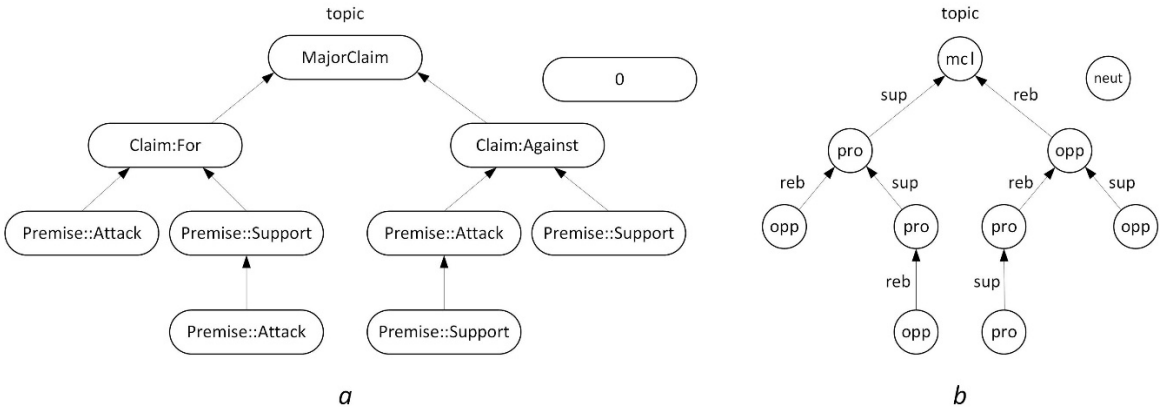


Figure 3: Equivalent graph representation of argumentation structure: a – PersEssays; b – JAAS

Thus, we used the Russian-language versions of the ArgMicro and PersEssays corpora, annotated in accordance with JAAS. The number of texts and ADUs in converted corpora is shown in Table 1. Both individual sentences and parts of sentences can be ADUs in these corpora. For the ArgMicro corpus the inter-annotator agreement by Fleiss kappa is equal to 0.83 (three annotators) [24]. For the PersEssays corpus the inter-annotator agreement by Krippendorff $\alpha_U = 0.72$ for argument components and $\alpha = 0.81$ for argumentative relations [29].

The ArgMicro corpus consists of 1,537 edges of "seg" type, 730 – "sup", 245 – "reb", 140 – "und", 78 – "add", 32 – "exa"; the PersEssays corpus: 7277 – "seg", 5617 – "sup", 715 – "reb", 301 – "exa".

Figure 4 shows an example of text from the ArgMicro corpus in the JAAS, where ADUs, their types ("mcl", "opp", "pro") and relationships between them ("reb", "und", "sup", "add") are indicated.

[3] https://translate.google.ru.

Corpora	Texts	ADUs				
		pro	opp	mcl	neut	all
ArgMicro	283	983 (63.8%)	253 (16.4%)	301 (19.5%)	4 (0.3%)	1,541 (100%)
PersEssays	399	4,599 (63.2%)	703 (9.7%)	746 (10.2%)	1,229 (16.9%)	7,277 (100%)
ArgMicro +PersEssays	682	5,582 (63.3%)	956 (10.8%)	1,047 (11.9%)	1,233 (14.0%)	8,818 (100%)

Table 1: Characteristics of text corpora

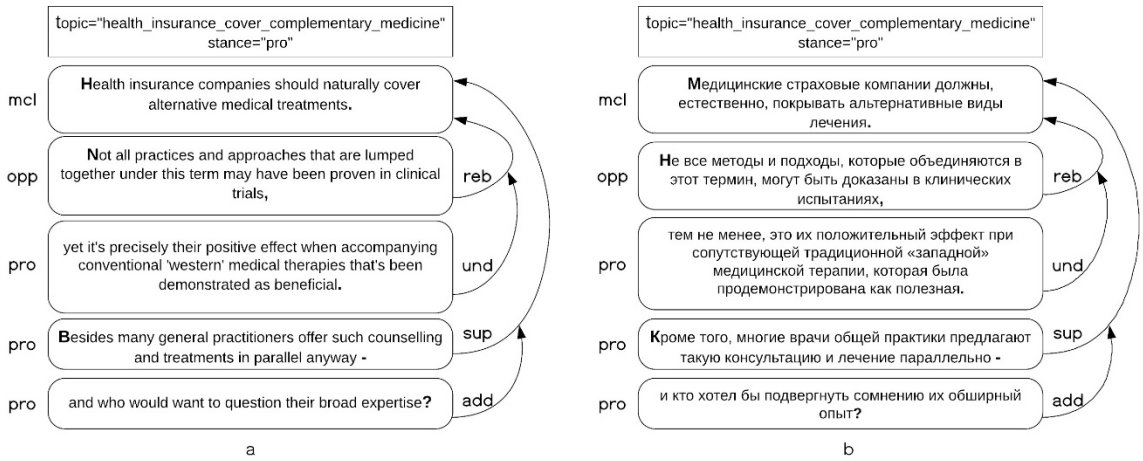


Figure 4: An example of text from the ArgMicro corpus:
a – original English-language variant; *b* – Russian-language variant (machine translation)

4.2 Features

Fishcheva and Kotelnikov [7] argued that when using traditional machine learning classifiers, the TF.IDF features, the word2vec features and the location of the sentence in the text do not improve the performance of the classifiers. Therefore, in this study, only the following three types of features were considered:

- lexical features – discourse markers ("consequently", "I think", "eventually", etc.) and modal words ("need", "maybe", "necessarily", etc.), including negations, 255 features in total;
- punctuation features – comma, colon, semicolon, question and exclamation marks, 5 features in total;
- morphosyntactic features – N-grams based on parts of speech (nouns, pronouns, verbs, adjectives and adverbs), N = {2, 3, 4}, and grammatical features of verbs: tense, mood, person, 783 features in total.

The preprocessing was carried out on the basis of tokenization and removal of stop words using *nltk*[4], as well as lemmatization using *mystem*[5]. For each ADU, a single vector was formed based on the concatenation of all feature types for the current ADU, as well as all feature types for the previous and next ADUs (if they present) in order to take the context into account.

[4] https://www.nltk.org.
[5] https://yandex.ru/dev/mystem.

4.3 Traditional machine learning techniques

For training, we used classifiers that gave the best results in [7] – linear Support Vector Machines (SVM), Bagging classifier and Gradient Boosting. The hyperparameters of the latter two classifiers were selected from the following ranges:

- Bagging: number of trees = [50, 100, 200, 500];
- Gradient Boosting: number of trees = [150]; maximum depth of a tree = [2, 8, 20, 30].

We used the SVM and Bagging implementation in *scikit-learn* [22], and for the gradient boosting we used the *XGBoost* library[6].

4.4 BERT

Bidirectional Encoder Representations from Transformers (BERT) [4] is a deep neural network language model based on the Transformer architecture [33]. The model is trained on a large text corpus using two tasks: masked words and next sentence prediction. The model is then fine-tuned to solve particular natural language processing tasks. BERT allows bi-directional context-dependent text processing. The model accepts a sequence of tokens (subwords) as input, which is then advanced through several layers of the encoder. The number of layers is 12 ($\text{BERT}_{\text{BASE}}$) or 24 ($\text{BERT}_{\text{LARGE}}$). Each layer applies self-attention mechanism and passes the results to the feed-forward network, after which the output of the current layer is fed to the input of the next layer. The encoder output is used as input to a linear classifier with a SoftMax function.

Within this study, the experiments were carried out using the RuBERT model from *DeepPavlov* [[14]]. RuBERT is a multilingual version of $\text{BERT}_{\text{BASE}}$ (12 layers, hidden size 768, feed-forward hidden size 3,072, and 12 self-attention heads), trained on the Russian-language Wikipedia and the news corpus.

The hyperparameters of the RuBERT model in our experiments were chosen from the following ranges:

- number of epochs = [3, 5, 7];
- learning rate = [10^{-3}, 10^{-4}, $5 \cdot 10^{-5}$, 10^{-5}, 10^{-6}, 10^{-7}];
- batch size = [4, 8, 16, 32].

5 Results and Discussion

5.1 Design of experiments

The experiments were carried out in order to get answers to the following questions:

- Q1: What performance of the binary classification of Russian-language ADUs into “pro” and “opp” can be achieved by modern machine learning models?
- Q2: Is it possible to improve the classification performance by expanding the training corpus?
- Q3: What is the significance of different types of features for traditional machine learning classifiers?

To conduct the experiments, we used the ArgMicro and PersEssays corpora, translated into Russian using Google Translate, annotated with JAAS. During the training, only “pro” and “opp” ADUs in these corpora were taken into account; “mcl” and “neut” were ignored (see Table 1).

We studied four variants to create training and test datasets:

- training on the ArgMicro, testing on the ArgMicro;
- training on the PersEssays, testing on the PersEssays;
- training on the ArgMicro and PersEssays, testing on the ArgMicro;
- training on the ArgMicro and PersEssays, testing on the PersEssays.

Due to the small size of the corpora, we used a 5-fold cross-validation for each of the four variants. The partitions were random, stratified by class, and kept the same for all experiments. In addition, in the

[6] https://xgboost.readthedocs.io.

experiments with RuBERT, with regard to the random initialization of the linear classifier weights, for each of the four variants five runs of the training procedure were carried out.

Because of the strong class imbalance of both corpora, the macro-averaged F1-score was used as the main performance metric. Accuracy and macro-averaged Precision and Recall were also calculated. The results obtained for each experiment were averaged, and the standard deviation for folds was computed.

Nested 3-fold cross-validation was used to fit the hyperparameters. The following hyperparameter values turned out to be optimal for the RuBERT model in most runs:

- training on the ArgMicro: number of epochs – 5, learning rate – 10^{-5}, batch size – 4;
- training on the PersEssays and training on the joint dataset (ArgMicro and PersEssays): number of epochs – 5, learning rate – 10^{-5}, batch size – 32.

5.2 Results

Table 2 shows the results for the XGBoost classifier, which turned out to be the best in all experiments among other traditional machine learning techniques (SVM and Bagging), as well as the results of RuBERT model. XGBoost results are presented for the full set of features – lexical, punctuation, and morphosyntactic (see Subsection 4.2).

Train dataset	Test dataset	Model	F_1-score	Precision	Recall	Accuracy
ArgMicro	ArgMicro	XGBoost	**0.7921 ±0.0309**	0.8567 ±0.0437	**0.7597 ±0.0324**	**0.8819 ±0.0166**
		RuBERT	0.7441 ±0.0537	0.7678 ±0.0414	0.7318 ±0.0598	0.8468 ±0.0228
ArgMicro+ PersEssays		XGBoost	0.7678 ±0.0203	**0.8583 ±0.0152**	0.7288 ±0.0204	0.8746 ±0.0081
		RuBERT	0.7349 ±0.0231	0.7691 ±0.0345	0.7159 ±0.0237	0.8429 ±0.0161
PersEssays	PersEssays	XGBoost	0.6308 ±0.0191	**0.7617 ±0.0433**	0.6009 ±0.0132	**0.8793 ±0.0073**
		RuBERT	**0.6715 ±0.0339**	0.7211 ±0.0292	**0.6469 ±0.0298**	0.8744 ±0.0088
ArgMicro+ PersEssays		XGBoost	0.6510 ±0.0165	0.7488 ±0.0303	0.6194 ±0.0120	0.8791 ±0.0066
		RuBERT	0.6665 ±0.0299	0.7250 ±0.0230	0.6398 ±0.0255	0.8752 ±0.0085

Table 2: Results of XGBoost and RuBERT:
macro-averaged F1-score, Precision, Recall and Accuracy (Mean ± Std Dev)

5.3 Discussion

The best result for the ArgMicro corpus (question Q1) was obtained using XGBoost (F_1-score=0.7921) when trained only on the ArgMicro. RuBERT lags far behind (F_1-score=0.7441): the ArgMicro corpus includes only 1,236 “pro” and “opp” ADUs, which are not enough for high-quality fine-tuning of the RuBERT, especially considering 5-fold cross-validation. Particularly low is the Precision for RuBERT relative to XGBoost (0.7678 vs. 0.8567).

For the PersEssays corpus, RuBERT produces the best result (F_1-score=0.6715). It outperforms XGBoost due to higher Recall (0.6469 vs. 0.6009). The PersEssays corpus (5,302 ADUs) is 4.3 times the size of ArgMicro, and RuBERT is able to train at a level that surpasses traditional machine learning techniques.

Expanding the training dataset (question Q2) by adding PersEssays to ArgMicro in the case of testing on ArgMicro worsens the results for XGBoost (by 0.024) and slightly decreases for RuBERT (by 0.009). Both classifiers lose performance due to macro-averaged Recall, which in turn is reduced due to Recall for the "opp" class: if the PersEssays corpus with a stronger class imbalance is added, it impairs the ability of classifiers to recognize minority class (see Table 3).

When ArgMicro is added to PersEssays, the results are diverse: for XGBoost, the performance improves by 0.02, for RuBERT – almost does not change (decreases by 0.005). When ArgMicro is added, the class imbalance is slightly reduced by increasing Recall for the "opp" class for XGBoost.

XGBoost produces more stable results: the standard deviation of results for folds is lower than for RuBERT (0.0217 vs. 0.0352 on average for all experiments).

The class imbalance in both corpora (the "pro" class is 79.5% in ArgMicro, 86.7% in PersEssays), leads to extremely uneven performance by class. Table 3 shows the performance metrics of the best models in Table 2 by classes.

Train dataset	Test dataset	Model	Macro F_1-score	Class	F_1-score	Precision	Recall
ArgMicro	ArgMicro	XGBoost	0.7921	pro (79.5%)	0.9286	0.8940	0.9664
				opp (20.5%)	0.6556	0.8193	0.5529
PersEssays	PersEssays	RuBERT	0.6715	pro (86.7%)	0.9296	0.9043	0.9565
				opp (13.3%)	0.4134	0.5379	0.3373

Table 3: Results of the best classifiers (XGBoost for ArgMicro and RuBERT for PersEssays) for "pro" and "opp" classes: F1-score, Precision and Recall

Table 4 contains the number of ADUs that were classified identically and differently by both classifiers. For example, the column "XGBoost – true, RuBERT – false" shows the number of ADUs that were correctly predicted by XGBoost and incorrectly – by RuBERT.

Test dataset	Class	XGBoost – true, RuBERT – true	XGBoost – true, RuBERT – false	XGBoost – false, RuBERT – true	XGBoost – false, RuBERT – false	Sum
ArgMicro	all	975	127	74	60	1,236
	pro	887	66	30	0	983
	opp	88	61	44	60	253
PersEssays	all	4,466	203	187	446	5,302
	pro	4,339	139	96	25	4,599
	opp	127	64	91	421	703

Table 4: Results of classification by number of ADUs

The analysis of Table 4 allows us to advance a hypothesis that the ensemble of both models will perform better than the models separately. The rule for predicting the ADU class in the ensemble is as follows: if at least one of the classifiers predicts a minority class “opp”, return “opp”; otherwise return “pro”. The ensemble results are shown in Table 5 along with the best classifier for the respective corpora.

Test dataset	Model	F_1-score	Precision	Recall	Accuracy
ArgMicro	XGBoost	0.7921±0.0309	**0.8567±0.0437**	0.7597±0.0324	**0.8819±0.0166**
	Ensemble	**0.8157±0.0305**	0.8022±0.0306	**0.8326±0.0312**	0.8738±0.0226
PersEssays	RuBERT	0.6715±0.0339	**0.7211±0.0292**	0.6469±0.0298	**0.8744±0.0088**
	Ensemble	**0.6901**±0.0138	0.7159±0.0185	**0.6723**±0.0114	0.8716±0.0068

Table 5: Results of ensemble of classifiers and the best classifiers (XGBoost for ArgMicro and RuBERT for PersEssays): F1-score, Precision and Recall (Mean ± Std Dev)

Table 5 shows that the use of the proposed ensemble allows improving the classification performance by 0.024 for ArgMicro and by 0.019 for PersEssays.

5.4 Feature importance

To answer question Q3 about the significance of various types of features, the dependence of the XGBoost classification performance on various combinations of features was investigated:

- lexical features – only lexical features in the previous, the current and the following ADUs;
- all without discourse markers – all features (lexical, punctuation and morphosyntactic) without discourse markers in the previous, the current and the following ADUs;
- all without features of previous ADUs – lexical, punctuation and morphosyntactic features only for the current and the following ADUs;
- all features – a full set of the features in the previous, the current and the following ADUs.

The results are shown in Table 6.

Train dataset	Test dataset	Lexical features	All without discourse markers	All without features of previous ADUs	All features
ArgMicro	ArgMicro	0.8092±0.0273	0.5440±0.0098	**0.8116±0.0355**	0.7921±0.0309
PersEssays	PersEssays	**0.6534±0.0184**	0.5140±0.0156	0.6284±0.0125	0.6308±0.0191

Table 6: Results of XGBoost for various combinations of feature types: macro-averaged F1-score (Mean ± Std Dev)

The discourse markers are the most important features, because without these features the performance of the classifier drops drastically. The previous ADU features are the most useless, since the exclusion of these features did not worsen the classifier performance, but allowed obtaining the best result for XGBoost (F_1-score=0.8116). Punctuation and morphosyntactic features are not very useful, because when these features are excluded (column “Lexical features”), the result is either close to the best (ArgMicro) or the best (PersEssays).

6 Conclusion

Thus, in order to use the ArgMicro and PersEssays corpora combined, the Join Argument Annotation Scheme based on the schemes used in ArgMicro and PersEssays has been proposed. The PersEssays corpus was translated into Russian using Google Translate and made publicly available[7]. We investigated the problem of classifying ADUs into two classes – "pro" and "opp". The experimental study made it possible to answer the questions posed:

Q1: What performance of the binary classification of Russian-language ADUs into "pro" and "opp" can be achieved by modern machine learning models? – The best performance (macro-averaged F_1-score) can be achieved by the proposed ensemble of XGBoost and RuBERT: for ArgMicro F_1-score=0.8157, for PersEssays F_1-score=0.6901. The performance for PersEssays is worse, firstly, due to the fact that the corpus is more imbalanced, and secondly, PersEssays ADUs are longer and more complex than in ArgMicro – the average ADU length in PersEssays is 18.6 tokens vs. 13.8 tokens in ArgMicro.

Q2: Is it possible to improve the performance of classification by expanding the training corpus? – Yes, if the imbalance of the extended corpus is reduced in comparison to the original one. If a less imbalanced ArgMicro corpus was added to a more imbalanced PersEssays, the performance for the XGBoost classifier was slightly higher. In other cases, the performance either did not increase or decreased.

Q3: What is the significance of different types of features for traditional classifiers? – Discourse markers turned out to be the most important features; features of previous ADUs have minimal impact on the performance of the classifier.

The urgent tasks to be solved in further research are, firstly, expanding the range of Russian-language corpora with argumentative annotation both based on the translation of existing corpora in other languages, and using annotation by people; secondly, the study of the performance of the argumentation mining for new corpora.

Acknowledgements

The reported study was jointly financed by the German Academic Exchange Service (DAAD) and the Ministry of Science and Higher Education of the Russian Federation within the "Michail Lomonosov" programme (2021).

References

[1] Addawood A.A., Bashir M.N. (2016), What is your evidence? A study of controversial topics on social media, Proceedings of the 3rd workshop on Argument Mining (ArgMining-2016), pp. 1–11.

[2] Aker A., Zhang H. (2017), Projection of Argumentative Corpora from Source to Target Languages, Proceedings of the 4th Workshop on Argument Mining, Copenhagen, Denmark, pp. 67–72.

[3] Baff R.E., Wachsmuth H., Al-Khatib K., Stein B. (2020), Analyzing the Persuasive Effect of Style in News Editorial Argumentation, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3154–3160.

[4] Devlin J., Chang M-W., Lee K., Toutanova K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 4171–4186.

[5] Eger S., Daxenberger J., Stab C., Gurevych I. (2018), Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need!, Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, pp. 831–844.

[6] Eger S., Ruckle A., Gurevych I. (2018), PD3: Better Low-Resource Cross-Lingual Transfer by Combining Direct Transfer and Annotation Projection, Proceedings of the 5th Workshop on Argument Mining, pp. 131–143.

[7] Fishcheva I., Kotelnikov E. (2019), Cross-Lingual Argumentation Mining for Russian Texts, Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST-2019), Springer, pp. 134–144.

[8] Freeman J.B. (2011), Argument Structure: Representation and Theory, Argumentation Library, Vol. 18. Springer.

[7] https://github.com/kotelnikov-ev/PersEssays_Russian.

[9] Green N.L. (2018), Towards mining scientific discourse using argumentation schemes, Argument & Computation, Vol. 9(2), pp. 121–135.
[10] Habernal I., Gurevych I. (2017), Argumentation mining in user-generated web discourse, Computational Linguistics, Vol. 43(1), pp. 125–179.
[11] Kirschner C., Eckle-Kohler J., Gurevych I. (2015), Linking the thoughts: Analysis of argumentation structures in scientific publications, Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1–11.
[12] Kononenko I., Sidorova E., Akhmadeeva I. (2020), The Study of Argumentative Relations in Popular Science Discourse, RCAI 2020: Artificial Intelligence, pp. 309–324.
[13] Küçük D., Can F. (2020), Stance Detection: A Survey, ACM Computing Surveys, Vol. 53(1), article no. 12.
[14] Kuratov Y., Arkhipov M. (2019), Adaptation of deep bidirectional multilingual transformers for Russian language, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2019», No. 18 (24), pp. 333–340.
[15] Lawrence J., Bex F., Reed C., Snaith M. (2012), AIFdb: infrastructure for the argument web, Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012), pp. 515–516.
[16] Lawrence J., Reed C. (2020), Argument Mining: A Survey, Computational Linguistics, Vol. 45(4), pp. 765–818.
[17] Liebeck M., Esau K., Conrad S. (2016), What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld, Proceedings of the 3rd workshop on Argument Mining (ArgMining-2016), pp. 144–153.
[18] Lippi M., Palka P., Contissa G., Lagioia F., Micklitz H.-W., Sartor G., Torroni P. (2019), CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service, Artificial Intelligence and Law, Vol. 27, pp. 117–139.
[19] Lytos A., Lagkas T., Sarigiannidis P., Bontcheva K. (2019), The evolution of argumentation mining: From models to social media and emerging tools, Information Processing and Management, Vol. 56, 102055.
[20] Macagno F., Walton D., Reed C. (2017), Argumentation Schemes. History, Classifications, and Computational Applications, Journal of Logics and their Applications, Vol. 4(8), pp. 2493–2556.
[21] Niculae V., Park J., Cardie C. (2017), Argument mining with structured SVMs and RNNs, Proceedings of the Association for Computational Linguistics, Vol. 1, pp. 985–995.
[22] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B. et al. (2011), Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.
[23] Peldszus A., Stede M. (2013), From argument diagrams to argumentation mining in texts: A survey, International Journal of Cognitive Informatics and Natural Intelligence, Vol. 7(1), pp. 1–31.
[24] Peldszus A., Stede M. (2015), An annotated corpus of argumentative microtexts, Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, pp. 801–815.
[25] Roush A., Balaji A. (2020), DebateSum: A large-scale argument mining and summarization dataset, Proceedings of the 7th Workshop on Argument Mining, pp. 1–7.
[26] Salomatina N., Kononenko I., Sidorova E., Pimenov I. (2021), Identification of connected arguments based on reasoning schemes “from expert opinion”, Journal of Physics: Conference Series, Vol. 1715.
[27] Skeppstedt M., Peldszus A., Stede M. (2018), More or less controlled elicitation of argumentative text: enlarging a microtext corpus via crowdsourcing, Proceedings of the 5th Workshop in Argumentation Mining, pp. 155–163.
[28] Sliwa A., Ma Y., Liu R., Borad N., Ziyaei S.F., Ghobadi M., Sabbah F., Aker A. (2018), Multilingual Argumentative Corpora in English, Turkish, Greek, Albanian, Croatian, Serbian, Macedonian, Bulgarian, Romanian and Arabic, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC’2018), Miyazaki, Japan, pp. 3908–3911.
[29] Stab C., Gurevych I. (2014), Annotating argument components and relations in persuasive essays, Proceedings of the International Conference on Computational Linguistics, pp. 1501–1510
[30] Stede M., Schneider J. (2018), Argumentation Mining, Synthesis Lectures on Human Language Technologies, San Rafael: Morgan and Claypool Publishers.
[31] Toledo-Ronen O., Orbach M., Bilu Y., Spector A., Slonim N. (2020), Multilingual Argument Mining: Datasets and Analysis, Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 303–317.
[32] Toulmin S.E. (1958), The Uses of Argument, Cambridge University Press.
[33] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N. et al. (2017), Attention is all you need, Advances in Neural Information Processing Systems, Vol. 30, pp. 5998–6008.

ruBTS: Russian Sentence Simplification Using Back-translation

Farit Galeev
Innopolis University
Innopolis, Russia

Marina Leushina
Innopolis University
Innopolis, Russia

Vladimir Ivanov
Innopolis University,
Innopolis, Russia
Kazan Federal University,
Kazan, Russia

{f.galeev,m.leushina,v.ivanov}@innopolis.ru

Abstract

Automatic text simplification is a crucial task enabling to reduce text complexity while preserving meaning. This paper presents our solution to the Russian Sentence Simplification Shared Task (RSSE) based on a back-translation technique. We show that applying the simple back-translation approach for sentence simplification can give competitive results with the other methods without fine-tuning or training.

Keywords: sentence simplification, Russian language, back-translation
DOI: 10.28995/2075-7182-2021-20-259-267

ruBTS: Упрощение предложений с использованием обратного перевода для русского языка

Фарит Галеев
Университет Иннополис

Марина Леушина
Университет Иннополис

Владимир Иванов
Университет Иннополис,
Казанский Федеральный Университет

f.galeev@innopolis.ru

m.leushina@innopolis.ru

v.ivanov@innopolis.ru

Аннотация

Автоматическое упрощение текста является важной задачей, позволяющей снизить сложность текста при сохранении его смысла. В статье представлено наше решение общей задачи по упрощению предложения на русском языке (RSSE), основанное на методе обратного перевода. Мы показываем, что применение простого обратного перевода для упрощения предложений может дать конкурентные результаты с другими методами без какой-либо тонкой настройки или обучения.

Ключевые слова: упрощение предложений, русский язык, обратный перевод

1 Introduction

Text simplification is used to reduce text complexity, either as part of natural language processing or as a stand-alone study area. Simplified texts are more accessible to people with reading and understanding issues, especially non-native speakers, people with related disorders. Text simplification implies modifications at different levels, including stylistic, grammatical, and lexical levels, preserving the text's original meaning. One of the most studied settings is simplifying a text sentence by sentence (sentence-level simplification), traditionally included paraphrasing as a subtask. However, with recent advances in natural language generation, sentence-level simplification combines transformer-based architectures and paraphrases and typically implies text generation as well [16].

Evaluation of modern methods for sentence simplification carried out using English corpora. Evaluation sentence simplification methods on Russian texts is a novel task that was proposed as a part of the

Dialog 2021 Evaluation initiative[1]. The organizers of the Russian Sentence Simplification Shared Task (RuSimpleSentEval, or RSSE) [14] prepared both train and test sets using a crowd-sourcing platform and translated texts from Simple Wikipedia. This paper presents our solution[2] to the proposed task based on a back-translation technique that does not require any fine-tuning and has shown competitive results.

2 Related Works

2.1 Text Simplification: Approaches

The most recent survey on text simplification [16] points out the connection between summarization and simplification and classifies text simplification approaches by the same categories: extractive and abstractive. Extractive approaches are based on selecting information from the text to preserve the most significant parts and drop less informative details. It is mainly used to simplify a significant amount of text and essentially solve text summarization tasks.

Abstractive approaches, contrary to extractive ones, use text generation for creating simplified text. These approaches can be divided into sentence-level and text-level simplification, or both. The main difference is that simplification can happen on only the lexical level by identifying complex words or phrases and replacing them with the more simple substitute. It can involve syntactic simplification, which may split complex grammatical constructions into simpler ones, delete or add information in the text.

However, these techniques can be used simultaneously. One can achieve it by learning simplification from data directly. A way to do that is the sequence to sequence modelling, the method for text-to-text generation, which was first applied for Text Simplification in 2017 by Nisioi et al. [12]. One of the best simplifications works, ACCESS [11], solves sentence simplification task using both lexical and syntactic approaches. AudienCe-CEntric Sentence Simplification aims to control attributes, which correspond to the text complexity: the amount of compression, amount of paraphrasing, lexical complexity, and syntactic complexity. Their solution is based on the transformer model [19], which is trained in a sequence-to-sequence manner.

The transformer is the model that was originally presented for Neural Machine Translation (NMT) [19]. Text Simplification can be considered Monolingual Translation, where the source text would be translated to the more straightforward text. Experiments with NMT techniques and Text Simplification first suggested [21] and conducted [20] in 2016. In the paper, Wang Tong et al. identify several differences between NMT and Text Simplification, which should be addressed while using NMT directly for simplification. It includes differences in vocabulary sizes, shared words in aligned sentences, and difficulties when splitting sentences in two. The LSTM-based model (Long Short-Term Memory) learned how to perform reversing, sorting, and replacement operations (for lexical and grammatical simplification).

Zhang et al. [25] suggest another model based on the monolingual translation and sequence to sequence approach. However, authors use reinforcement learning algorithms to encourage a variety of simplification tricks by rewarding simplicity, relevance, and fluency. The motivation behind this is that most used datasets for simplification contain many copies of the original text as simplifications, which creates an imbalance in the applied simplifications.

Another way to better comprehend simplification with the monolingual translation is to include external knowledge bases to increase learned simplification rules. Zhao et al. [26] suggest two modifications, one of which takes advantage of Simple PPDB (A Paraphrase Database for Simplification) [13], by encouraging the model to apply simplification rules, presented in Simple PPDB.

2.2 Text Simplification: Datasets

One downside of these approaches is that a lot of paired data is required to achieve good results. There are several widely used datasets for English:

[1] http://www.dialog-21.ru/evaluation/; https://github.com/dialogue-evaluation/RuSimpleSentEval

[2] Source code of solution is available at `https://github.com/HiGal/RSSE`

- Simple English Wikipedia: several datasets (Wikipedia - Simple Wikipedia [6], PWKP [27], SS Corpus [5]) constructed by parsing Simple English Wikipedia in pair with regular English Wikipedia to obtained paired sentences.
- Xu et al. [23] point at the problems in Simple Wikipedia, such as not aligned sentences between corresponding articles, target sentences that are not simpler than the source, or just noisy sentences. Thus, they present the Newsela dataset, which contains the data of 1130 news articles, where each article contains five versions (one original and four simplified versions), re-written by editors from Newsela company.
- Xu et al. also presented Turk dataset [24] in 2016, which was collected through a crowdsourced rewriting of English Wikipedia sentences on Amazon Mechanical Turk.

Nevertheless, for other languages, it presents a problem in obtaining such a dataset. For Russian, recent work by Gudkov et al. [2] presents a method for paraphrase generating based on the denoising procedure and resulting ParaPhraser Plus corpus. Authors show that automatically aligned and ranked datasets can generate paraphrasing, especially in low-resource languages. Another way to solve the Text Simplification problem for such languages is to turn to multilingual models of various transformer architectures, allowing them to join datasets of different languages to enlarge the training data.

2.3 Text Simplification Evaluation: Metrics and Tools

The primary evaluation metric used for the translation problem is BLEU (bilingual evaluation understudy). However, several works [18], [24] showed that this is not the best choice for text simplification due to its low correlation with grammaticality and meaning preservation and human evaluation of simplification. In 2016, Wei Xu et al. [24] use paraphrasing as the primary tool for Text Simplification and suggest two new metrics that stated solve these problems. New metrics, FK-BLEU and SARI, measure readability and goodness of word choice, respectively.

FK-BLEU represents a combination of the Flesch-Kincaid Index (FK)[7] and BLEU, allowing this metric to measure readability (from FK) and adequacy (from BLEU). Flesch-Kincaid Index is a readability metric that was suggested back in 1975. It is calculated based on the number of words in sentences and the number of syllables in words. Although this metric is easy to compute since it relies on average lengths of sentences, it does not reflect adequacy. It also does not reflect on meaning preservation since it does not compare sentences with any references of possible simplifications.

SARI, in turn, uses multiple references and input sentences to compare with the result. Authors [24] show that BLEU assigns a higher score to the samples with the same complexity level and not penalizes them as SARI does. Along with FK-BLEU, these metrics achieve a much higher correlation with humans' evaluation of simplicity, keeping in mind grammaticality and meaning preservation.

The Python package EASSE [1], Easier Automatic Sentence Simplification Evaluation, is a helpful tool for automatic evaluation of simplification quality. It can evaluate simplification using BLEU and SARI metrics using references. In addition, it can calculate reference-independent quality metrics: FK grade level[7], Levenshtein similarity [8], Lexical Complexity score[11], and compression level. Lexical Complexity score is computing the third-quartile of log-ranks (inverse frequency order) of all words in a sentence. Compression level refers to the character length ratio between the original sentence and its simplified version. Levenshtein similarity [8], Lexical Complexity score (referred to as WordRank), and compression level was used as the tokens which control the simplification process in the ACCESS model [11], one of the state-of-the-art models.

3 Proposed Approach and Models

In this work, we conduct experiments on three different approaches: training the transformer-based NMT model, fine-tuning the MBart model, and applying back-translation to inference pre-trained Text Simplification model. Section 3 describes motivation and details of conducted experiments, along with data preparation, and details and results of these experiments described in Section 4.

3.1 Data Preprocessing

We have chosen automatically translated WikiLarge Dataset [25] provided by organizers as a dataset. It already split into train, test, and validation sets. However, this translated corpus has some problems, such as repetitive target sentences that do not save the sentence's meaning, so that dataset needs additional preprocessing.

Initially, we remove repetitive sentences from the dataset because such samples can distract the model during training. Then, since it is hard for any model to process long sequences and force padding for every sentence to maximum length, we keep only the sentence pairs in which a complex sentence length does not exceed 350 words, and the length of the simple sentence does not exceed 300 words.

As an additional dataset, we use ParaPhraserPlus [2] (a dataset of paraphrased headlines for the Russian Language) without any additional preprocessing for more training samples. The effect of extending the training dataset with ParaPhraserPlus is described in Section 4.

3.2 NMT Transformer as Sentence Simplification model

As we discussed in Section 2, the task of Sentence Simplification can be interpreted as sequence-to-sequence modelling. There is also evidence that sentence simplification is close to the NMT task [21]. Due to this, our next experiment is to train the Sentence Simplification model as the NMT model.

State-of-the-art NMT models use sequence-to-sequence architectures consist of two parts encoder and decoder. The encoder codes information of the input sequence, while the decoder tries to generate a new sequence based on the input sequence. The input sequence in the NMT task is in the source language, and the target sequence is in the target language. According to the NMT task and its application in sentence simplification, the source sequence will be a complex sentence, and the target sequence will be a simplified sentence.

Recently transformers [19] showed promising results in the translation task. So, as an NMT model, we chose transformer architecture described in [19] with three encoder layers and three decoder layers. We observed that sinusoidal positional encoding influences the model convergence (in our case model did not converge), so we decided to replace it with positional embeddings [22]. The number of heads in multi-head self-attention was set to 8.

As an activation function, we use GeLU, and the remaining model parameters were the same as in transformer [19]. For tokenization, a pre-trained ruBERT tokenizer was used. To train the model, standard cross-entropy loss was selected with Adam optimizer and reduce on plateau scheduler with a learning rate equal to $3 \cdot 10^{-4}$.

3.3 Fine-Tuning MBart Model

We fine-tune the MBart model on sentence simplification as our next experiment. MBart is the multilingual model for sequence-to-sequence generation that showed state-of-the-art results on various text generation tasks, including NMT [9]. We fine-tune two different MBart models, one that was trained to translate text between different languages and one that was trained to summarize Russian news [3]. Summarization is a task close to the Text Simplification problem, so we wanted to see if using a model trained for this task will improve results compared to the model trained on classic MBart.

We fine-tune both pre-trained models, for translation and summarization, by the same algorithm. We freeze the encoder and positional embeddings of the MBart, and train in a sequence-to-sequence manner using cross-entropy loss.

3.4 Sentence simplification through back-translation

One of the techniques to get pseudo parallel corpora for context-aware NMT models is data augmentation using back-translation [17]. So, taking this approach, we assume that sentence simplification can be partially solved with the back-translation technique without fine-tuning to a downstream task or training a new model. This approach does not require additional computing power, which necessary to train modern models on large datasets.

The idea of the method is to first translate the source sentence from Russian to English and translate the sentence back to Russian. As our machine translation model, we chose MarianMT model [4] from hugging-face trained on different language pairs, including Russian-English and English-Russian. We leverage a machine translation system to perform a two-step approach: (1) translating forward $(RU \rightarrow EN)$ followed by (2) back translation $(EN \rightarrow RU)$. The assumption behind the approach is that the machine translation system will probably have a limited vocabulary and, therefore, will produce simplification as a part of translation; performing the back-translation can potentially further simplify the sentence.

The proposed two-step approach complicates the whole process of sentence simplification. More advanced techniques for simplifications at the sentence level exist, such as MUSS [10]. They were already tested for English, French and Spanish languages but required much computational power for training. Besides the bigger carbon footprint, one may find it challenging to participate in deep learning research due to the high cost of such computations. However, machine translation for English is already good enough and can be used 'out of the box'. This consideration justifies using the back-translation method for the sake of rational use of computing resources. In our experiments, we try to answer the question, is it worth complicating the process of sentence simplification using the back-translation in terms of the quality of the result?

4 Experiments and Results

We conduct experiments with training Transformer, fine-tuning the two MBarts (a model pre-trained for NMT and a model pre-trained for Russian Text Summarization). Finally, we test the Back-translation method. We should also mention that using the ParaPhraserPlus dataset, described in Section 2.2, for Transformer training does not give any improvements but increases training time drastically. Due to this fact, we do not use ParaPhraserPlus for MBart fine-tuning.

Represented in SARI score results of all methods are shown in Table 1. The Back-translation method shows the best result among the approaches that we applied. Thus, further we provide a more profound analysis of its result.

Method	SARI
MBart fine-tuned for translation	26.38
MBart pretrained on news summarization	32.32
Transformer	32.50
Back-translation (MarianMT-based model)	**37.08**

Table 1: Results on the sentence simplification task on public test set. Back-translation significantly outperforms other models with no training or fine-tuning on downstream task. Although, we should mention that the final score of the system calculated on the private test set was 36.94.

We use the EASSE package [1], which provides useful metrics to evaluate the result of the Back-translation method. Figure 1 shows plots of Levenshtein Similarity and Compression ratio between system output and reference sentences. Levenshtein similarity quantifies the extent to which the source sentence has been changed (through paraphrasing, adding, and deleting content). Compression ratio is the proportion of the number of characters between the source and target sentences. Both can be interpreted as indirect indicators of sentence simplification quality. We provide the results in Table 2, where identity baseline takes input sentence as system output; truncate baseline takes the first 80% of words of input as system output, and the reference takes randomly one of the references as system output. One can mention that both compression ratio and Levenshtein values are small for reference data. Indeed, the simpler the sentence the shorter it has to be, at the same time preserving the original content. However, one can see that in the first row of the Table 2 compression ratio is greater than 1.0 for the system output. This can be explained by the fact that a simpler sentence should not be always shorter than the original.

Finally, we provide some of the worst and best examples of the simplification results (Tables 3 and

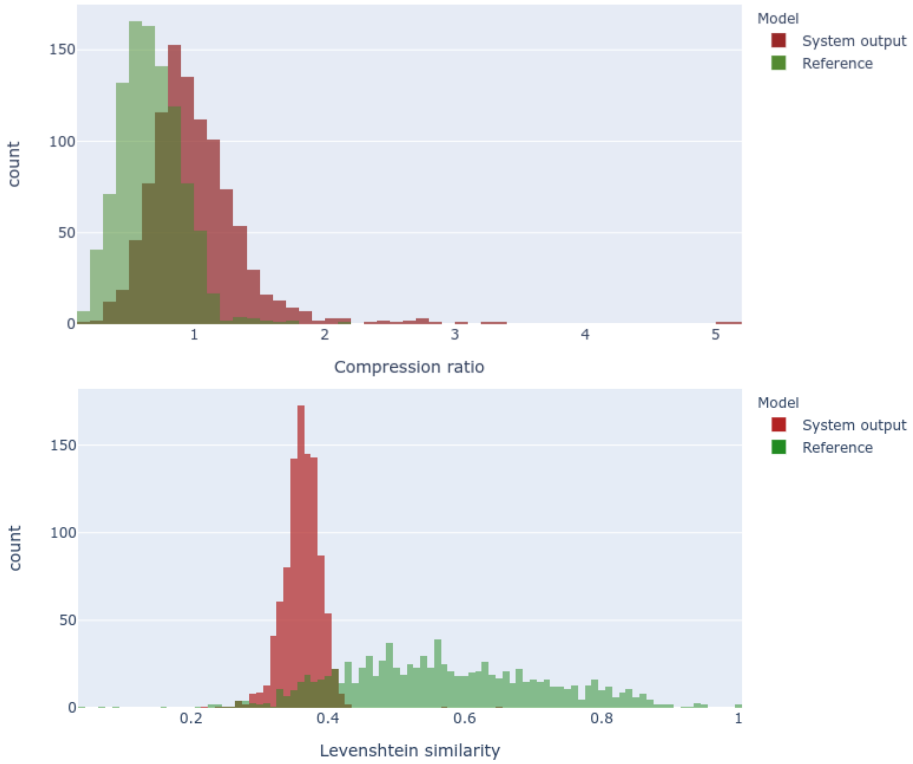


Figure 1: Compression ratio and Levenshtein similarity between system output and reference sentences on the dev part of public dataset (images were generated by the EASSE tool).

	SARI	Compression ratio	Levenshtein similarity
System output	32.47	1.01	0.37
Identity baseline	11.04	1.0	1.0
Truncate baseline	22.84	0.78	0.88
Reference	40.84	0.67	0.58

Table 2: Comparing back-translation method metrics to simple baselines (results were generated by the EASSE tool using the dev set).

4 respectively). One can see that the back-translation method sometimes can copy the source sentence, which we attribute to the performance of the underlying NMT model. In fact, we observed that such 'errors' appear when the sentence is not 'complex enough'; the simplification does not become a part of the translation process. The examples with high SARI scores show that the non-trivial transformation derives the simplified version of the sentence.

5 Conclusion

The proposed in this paper method can be applied to the Russian sentences simplification task. We show that the simple back-translation technique for sentence simplification can provide competitive results without fine-tuning or training. Such a result might be significant in green AI because the required computations for deep learning research have been doubling every few months [15], leading to significant carbon footprints. Besides the problem of air pollution, researchers, students, especially those from developing economies, may find it challenging to participate in deep learning research either due to the high computations cost or due to the absence of the datasets.

There is a limitation for applying our method outside the "cottonwool" conditions of the Dialogue

	Sentence	SARI
Original	Дания является одним из мировых лидеров в использовании возобновляемых источников энергии, в частности энергии ветра.	0.0
Simplified	Дания является одним из мировых лидеров в использовании возобновляемых источников энергии, в частности энергии ветра.	
Original	Разделение равнинных и горных районов между двумя государствами лишило бы многочисленных азербайджанских кочевников летних пастбищ.	9.8
Simplified	Разделение равнинных и горных районов между двумя государствами лишило бы многих азербайджанских кочевников летних пастбищ.	

Table 3: Examples of the worst simplifications according to SARI score.

	Sentence	SARI
Original	В вечерне-ночное время могут возникать ощущения нехватки воздуха, сердцебиение, потливость, озноб или приливы жара.	32.39
Simplified	В вечернее время может возникнуть чувство отсутствия воздуха, сердцебиения, потности, холода или жары.	
Original	1960 году была выпущена модель 172A. Изменения: хвостовое оперение и руль направления с обратной стреловидностью и крепления для поплавкового шасси.	35.74
Simplified	Модель 172A была выпущена в 1960 году.	

Table 4: Examples of the best simplification according to SARI score.

competition. The scalability of the applied technology to the problem of simplification depends on many factors. Most of them, such as dependency on the neural machine translator, the complexity of a sentence, memory requirements etc., are out of the scope of this study. However, some of the factors are easy to assess and overcome (for instance, switch to another sequence modelling toolkit can improve execution time). Although the scalability of the method is questionable, we claim that the approach, in general, can be investigated further by exploring other languages as well as other 'backbone' neural machine translators.

Acknowledgements

We thank Innopolis University for generously funding this research and anonymous reviewers for valuable comments.

References

[1] Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. EASSE: Easier automatic sentence simplification evaluation. 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): System Demonstrations*, pages 49–54, Hong Kong, China, November 2019. Association for Computational Linguistics.

[2] Vadim Gudkov, Olga Mitrofanova, and Elizaveta Filippskikh. Automatically ranked Russian paraphrase corpus for text generation. In *Proceedings of the Fourth Workshop on Neural Generation*

and Translation, pages 54–59, Online, July 2020. Association for Computational Linguistics.

[3] Ilya Gusev. Dataset for automatic summarization of russian news. In Andrey Filchenkov, Janne Kauttonen, and Lidia Pivovarova, editors, *Artificial Intelligence and Natural Language*, pages 122–134, Cham, 2020. Springer International Publishing.

[4] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[5] Tomoyuki Kajiwara and Mamoru Komachi. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[6] David Kauchak. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[7] J.P. Kincaid. *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis, 1975.

[8] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[9] Yinhan Liu, Jiatao Gu, Naman Goyal, X. Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[10] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. Multilingual Unsupervised Sentence Simplification. working paper or preprint, January 2021.

[11] Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. Controllable sentence simplification. In *LREC*, 2020.

[12] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[13] Ellie Pavlick and Chris Callison-Burch. Simple PPDB: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Berlin, Germany, August 2016. Association for Computational Linguistics.

[14] Andrey Sakhovskiy, Izhevskaya, Alexandra, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivana Smurov, and Ekaterina Artemova. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue"*, volume XX, pages xx–xx, 2021.

[15] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CoRR*, abs/1907.10597, 2019.

[16] Punardeep Sikka, Manmeet Singh, Allen Pink, and Vijay Mago. A survey on text simplification. *arXiv preprint arXiv:2008.08612*, 2020.

[17] Amane Sugiyama and Naoki Yoshinaga. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China, November 2019. Association for Computational Linguistics.

[18] E. Sulem, O. Abend, and A. Rappoport. Bleu is not suitable for the evaluation of text simplification. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2020.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[20] Tong Wang, Ping Chen, Kevin Amaral, and Jipeng Qiang. An experimental study of lstm encoder-decoder model for text simplification. *arXiv preprint arXiv:1609.03663*, 2016.

[21] Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. Text simplification using neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[22] Yu-An Wang and Yun-Nung Chen. What do position embeddings learn? an empirical study of pre-trained language model positional encoding, 2020.

[23] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

[24] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

[25] Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 584–594, 2017.

[26] S. Zhao, R. Meng, D. He, S. Andi, and P. Bambang. Integrating transformer and paraphrase rules for sentence simplification. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2020.

[27] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China, August 2010. Coling 2010 Organizing Committee.

Transfer Learning for Improving Results on Russian Sentiment Datasets

Anton Golubev
Bauman Moscow State
Technical University
Russia
antongolubev5@yandex.ru

Natalia Loukachevitch
Lomonosov Moscow State
University
Russia
louk_nat@mail.ru

Abstract

In this study, we test transfer learning approach on Russian sentiment benchmark datasets using additional train sample created with distant supervision technique. We compare several variants of combining additional data with benchmark train samples. The best results were achieved using three-step approach of sequential training on general, thematic and original train samples. For most datasets, the results were improved by more than 3% to the current state-of-the-art methods. The BERT-NLI model treating sentiment classification problem as a natural language inference task reached the human level of sentiment analysis on one of the datasets.

Keywords: Targeted sentiment analysis, distant supervision, transfer learning, BERT
DOI: 10.28995/2075-7182-2021-20-268-277

1 Introduction

Sentiment analysis or opinion mining is an important natural language processing task used to determine sentiment attitude of the text. One of its main business application is product monitoring consisting of studying customer feedback and needs. Nowadays most state-of-the-art results are obtained using deep learning models, which require training on specialized labeled data.

In recent years transfer learning has earned widespread popularity. This approach includes a pre-training step of learning general representations from a source task and an adaptation step of applying previously gained knowledge to a target task. In other words, deep learning model trained for a task is reused as the starting point for a model on a second task. Since there is a significant amount of text information nowadays, current state-of-the-art results can be possibly improved using transfer learning.

The most known Russian sentiment analysis datasets include ROMIP-2013 and SentiRuEval2015-2016 [4, 11, 12], which consist of annotated data on banks and telecom operators reviews from Twitter messages and news quotes. Current best results on these datasets were obtained using pre-trained RuBERT [20, 7] and conversational BERT model [23, 3] fine-tuned as architectures treating a sentiment classification task as a natural language inference (NLI) or question answering (QA) problem [7].

In this study, we introduce a method for automatic generation of annotated sample from a Russian news corpus using distant supervision technique. We compare different variants of combining additional data with original train samples and test the transfer learning approach based on several BERT models. For most datasets, the results were improved by more than 3% to the current state-of-the-art performance. On SentiRuEval-2015 Telecom Operators Dataset, the BERT-NLI model treating a sentiment classification problem as a natural language inference task, reached human level according to one of the metrics.

The contributions of this paper are presented below:

- we propose a new method of automatic generation of additional data for sentiment analysis tasks from raw texts using the distant supervision approach and a sentiment lexicon,
- we compare several variants of combining the additional dataset with original train samples and show that three-step approach of sequential training on general, thematic and benchmark train samples performs better,
- we renew the best results on five Russian sentiment analysis datasets using pre-trained BERT models combined with transfer learning approach,

- we show that BERT-NLI model treating sentiment classification problem as a natural language inference task reaches human level on one of the datasets.

This paper is structured as follows. In Section 2, we overview related methods applied to the considered task. Section 3 describes sentiment analysis datasets used in this paper. Section 4 represents a process of automatic generation of an annotated dataset using distant supervision approach. In section 5 and 6 we briefly cover main preprocessing steps and BERT models applied in the current study. Section 7 presents the full track of transfer learning study including comparison of different variants of constructing an additional datasets and several ways of their combining with benchmark train samples.

2 Related Work

Russian sentiment analysis datasets are based on different data sources [20], including reviews [19, 4], news stories [4], posts from social networks [16, 11, 17]. The best results on most available datasets are obtained using transfer learning approaches based on the BERT model [23], more specifically on RuBERT [3] and Russian variant of BERT [20, 7, 14, 1]. In [7], the authors tested several variants of RuBERT and different settings of its applications, and found that the best results on sentiment analysis tasks on several datasets were achieved using Conversational RuBERT trained on Russian social networks posts and comments. Among several architectures, the BERT-NLI model treating the sentiment classification problem as a natural language inference task usually has the highest results.

For automatic generation of annotated data for sentiment analysis task, researchers use so-called distant supervision approach, which exploits additional resources: users' tags or manual lexicons [6, 16]. For Twitter sentiment analysis, users' positive or negative emoticons or hashtags can be used [5, 16, 13]. Authors of [18] use the RuSentiFrames lexicon for creating a large automatically annotated dataset for recognition of sentiment relations between mentioned entities.

In contrast to previous work, in the current study we automatically create a dataset for targeted sentiment analysis, which extracts a sentiment attitude towards a specific entity. The use of an automatic dataset together with manually annotated data allows us to improve the state-of-the-art results.

Table 1: Benchmark sample sizes and sentiment class distributions (%).

Dataset	Train sample				Test sample			
	Vol.	Posit.	Negat.	Neutral	Vol.	Posit.	Negat.	Neutral
ROMIP-2013[3]	4260	26	44	30	5500	32	41	27
SRE-2015 Banks[4]	6232	7	36	57	4612	8	14	78
SRE-2015 Telecom[4]	5241	19	34	47	4173	10	23	67
SRE-2016 Banks[5]	10725	7	26	67	3418	9	23	68
SRE-2016 Telecom[5]	9209	15	28	57	2460	10	47	43

3 Russian sentiment benchmark datasets

In our study, we consider the following Russian datasets (benchmarks) annotated for previous Russian sentiment shared tasks: news quotes from the ROMIP-2013 evaluation [4] and Twitter datasets from SentiRuEval 2015-2016 evaluations [11, 12]. Table 1 presents main characteristics of datasets including train and test sizes and distributions by sentiment classes. It can be seen in Table 1 that the neutral class is prevailing in all Twitter datasets. For this reason, along with the standard metrics of $F_1\ macro$ and accuracy, $F_1^{+-}macro$ and $F_1^{+-}micro$ ignoring the neutral class were also calculated.

The collection of news quotes contains opinions in direct or indirect speech extracted from news articles [4]. The task of ROMIP-2013 evaluation was to distribute quotations between neutral, positive and negative classes depending on its sentiment. It can be seen in Table 1 that dataset is rather balanced.

[3] http://romip.ru/en/collections/sentiment-news-collection-2012.html
[4] https://drive.google.com/drive/folders/1bAxIDjVz_0UQn-iJwhnUwngjivS2kfM3
[5] https://drive.google.com/drive/folders/0BxlA8wH3PTUfV1F1UTBwVTJPd3c

Twitter datasets from SentiRuEval-2015-2016 evaluations were annotated for the task of reputation monitoring [15, 11], which means searching sentiment-oriented opinions about banks and telecom companies. In such a way this task can be regarded as an entity-oriented sentiment analysis problem. Insignificant part of samples contains two or more sentiment analysis objects, so these tweets are duplicated with corresponding attitude labels. The SentiRuEval-2016 training datasets are much larger in size as they contain training and test samples of 2015 evaluation [12]. As it can be seen in Table 1, Twitter datasets are poorly balanced. This explains the choice of metrics considering only positive and negative classes.

4 Automatic generation of annotated dataset

The main idea of automatic annotation of dataset for targeted sentiment analysis task is based on the use of a sentiment lexicon comprising negative and positive words and phrases with their sentiment scores. We utilize Russian sentiment lexicon RuSentiLex [10], which includes general sentiment words of Russian language, slang words from Twitter and words with positive or negative associations (connotations) from the news corpus. For ambigous words, having several senses with different sentiment orientations, RuSentiLex describes senses with references to the concepts of RuThes thesaurus [9]. The current version of RuSentiLex contains 16445 senses.

As a source for automatic dataset generation, we use a Russian news corpus, collected from various sources and representing different topics, which is important in fact that the benchmarks under analysis cover several topics. The corpus was collected long before the evaluations, so there are no possible overlaps between additional and benchmark data. The volume of the original corpus was about 4 Gb of raw text, which implies more than 10 million sentences.

The automatically annotated dataset includes general and thematic parts. For creation of the general part, we select monosemous positive and negative nouns from the RuSentiLex lexicon, which can be used as references to people or companies, which are sentiment targets in the benchmarks. We construct positive and negative word lists and suppose that if a word from the list occurs in a sentence, it has a context of the same sentiment. The list of positive and negative references to people or companies (seed words) includes 822 negative references and 108 positive ones. Examples of such words are presented below (translated from Russian):

- positive: *"champion, hero, good-looker"*, etc.;
- negative: *"outsider, swindler, liar, defrauder, deserter"*, etc.

Sentences may contain several seed words with different sentiments. In such cases, we duplicate sentences with labels in accordance with their attitudes. The examples of extracted sentences are as follows (all further examples are translated from Russian):

- positive: *"A MASK is one who, on a gratuitous basis, helps the development of science and art, provides them with material assistance from their own funds"*;
- negative: *"Such irresponsibility — non-payments — hits not only the MASK himself, but also throughout the house in which he lives"*.

To generate the thematic part of the automatic sample, we search for sentences that mention named entities depending on a task (banks or operators) using the named entity recognition model (NER) from DeepPavlov [3] co-occurred with sentiment words in the same sentences. We searched for sentences not only with organizations from benchmarks, but also with others companies from the relevant field. To ensure that a sentiment word refers to an entity, we restrict the distance between two words to be not more than four words.

We remove examples containing a particle *"not"* near sentiment word because it could change attitude of text in relation to target. Sentences with sentiment word located in quotation marks were also removed because they could distort the meaning of the sentence being a proper name.Examples of extracted thematic sentiment sentences are as follows:

- for banks (positive): *"MASK increased its net profit in November by 10.7%"*
- for mobile operators (negative): *"FAS suspects MASK of imposing paid services on subscribers."*

Since the benchmarks contain also the neutral sentiment class, we need to extract sentences without

sentiments. For this task, we choose among examples selected by NER those that do not contain any sentiment words from the lexicon. Examples of extracted neutral sentences for both general and thematic parts are presented below:

- for persons: *"MASK is already starting training with its new team."*
- for banks: *"On March 14, MASK announced that it was starting rebranding."*
- for mobile operators: *"MASK has offered its subscribers a new service."*

While creating an additional dataset, we take into account the distribution of sentiment words in the resulting sample, trying to bring it as close as possible to uniform. A source corpus contains enough examples with a negative sentiment to form a balanced dataset, which can not be said about words with the positive sentiment. We made automatically generated dataset publicly available[6].

Table 2: Results based on training on additional dataset only.

Dataset	Model	Accuracy	$F_1\ macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
ROMIP-2013	BERT-single	28.32	21.54	45.74	46.19
	BERT-pair-QA	28.04	21.32	45.35	45.78
	BERT-pair-NLI	27.76	20.89	45.12	45.68
SRE-2015 Banks	BERT-single	33.42	25.10	39.17	42.29
	BERT-pair-QA	33.19	25.56	38.98	42.31
	BERT-pair-NLI	32.56	24.87	38.63	41.87
SRE-2015 Telecom	BERT-single	26.11	19.12	33.56	34.21
	BERT-pair-QA	26.12	19.05	32.61	34.43
	BERT-pair-NLI	25.13	19.25	31.78	34.02
SRE-2016 Banks	BERT-single	28.91	22.14	36.45	38.88
	BERT-pair-QA	29.43	21.72	35.62	38.26
	BERT-pair-NLI	28.58	20.42	34.38	37.73
SRE-2016 Telecom	BERT-single	25.86	19.57	32.87	34.59
	BERT-pair-QA	25.27	18.76	32.09	33.65
	BERT-pair-NLI	24.14	18.23	31.06	33.28

5 Text preprocessing

To create an additional sample from the Russian news corpus, it was necessary to divide raw articles into separate sentences. For this task, we used rule-based sentence splitter from spaCy library [22], which is able to determine sentence boundaries automatically. This solution showed better quality in preliminary studies in comparison with NLTK variant [2] and simple splitter based on regular expressions.

In addition to conceptual steps of creating an automatic dataset described in previous chapter, a few cleaning measures were performed. In accordance with calculated quantiles of sentences from test samples, too short and long examples were removed from additional data. To remove duplicate sentences from different sources, we use the metrics of cosine similarity between pairs of tf-idf representations of examples. When the value of the specified boundary value was exceeded, one of the sentences was randomly removed. Conducting experiments with different thresholds and exploring resulting samples, we set value equal to 0.8.

After bringing the additional sample to the desired format, standard preprocessing track described in [7], including replacing similar text elements with appropriate tokens and removing special symbols was carried out for all datasets.

6 BERT architectures

In our study, we consider three variants of fine-tuning BERT models [23] for sentiment analysis. These architectures can be subdivided into the single-sentence approach using only initial text as an input

[6] https://github.com/antongolubev5/Auto-Dataset-For-Transfer-Learning

and the two-sentence approach [21, 7], which converts the sentiment analysis task into a sentence-pair classification task by appending an additional sentence to the initial text.

The sentence-single model represents a vanilla BERT with an additional single linear layer on the top. The unique token *[CLS]* is added for the classification task at the beginning of the sentence. The sentence-pair architecture adds an auxiliary sentence to the original input, inserting the *[SEP]* token between two sentences. The difference between two models is in addition of a linear layer: for the sentence-pair model it is added over the final hidden state of *[CLS]* token, while for the sentence-single variant it is added on the top of the entire last layer.

In our study, we use pre-trained Conversational RuBERT[7] from DeepPavlov framework [8] trained on Russian social networks posts and comments which showed better results in preliminary study.

Table 3: Results based on training on additional data mixed with benchmark train samples.

Dataset	Model	Accuracy	$F_1\ macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
ROMIP-2013	BERT-single	65.21	54.32	45.12	44.67
	BERT-pair-QA	65.53	54.68	45.73	45.14
	BERT-pair-NLI	65.45	54.93	45.52	44.89
SRE-2015 Banks	BERT-single	69.34	56.84	36.39	40.19
	BERT-pair-QA	70.21	57.25	36.83	40.79
	BERT-pair-NLI	69.54	57.06	36.65	40.31
SRE-2015 Telecom	BERT-single	66.43	53.19	33.41	37.71
	BERT-pair-QA	66.19	52.83	33.21	37.43
	BERT-pair-NLI	67.11	53.48	33.73	38.03
SRE-2016 Banks	BERT-single	67.71	54.76	33.61	37.85
	BERT-pair-QA	67.61	54.85	34.53	36.89
	BERT-pair-NLI	67.67	54.85	32.12	36.76
SRE-2016 Telecom	BERT-single	65.12	52.43	32.19	36.43
	BERT-pair-QA	64.76	52.06	32.28	36.12
	BERT-pair-NLI	65.21	52.27	32.49	36.51

For the targeted sentiment analysis task, there are labels for each object of attitude so they can be replaced by a special token *[MASK]*. Since general sentiment analysis problem has no certain attitude objects, token is assigned to the whole sentence and located at the beginning.

The sentence-pair model has two kind of architecture based on question answering (QA) and natural language inference (NLI) problems. The auxiliary sentences for each model are as follows:

- pair-NLI: *"The sentiment polarity of MASK is"*
- pair-QA: *"What do you think about MASK?"*

7 Experiments and results

We consider different options of constructing pre-training samples from the collected data and combining the resulting additional dataset with benchmark train samples. Different constructing variants comprise the following options:

- training on the additional general and neutral thematic data only and studying dependence of the results on sentiment class distribution;
- training on the additional general and neutral thematic data mixed with the benchmark training set;
- training on the full generated data (the data of previous steps are extended with sentiment-oriented thematic examples) mixed with the benchmark training set;
- two-step approach: independent sequential training on additional dataset at the first step and on the benchmark training set at the second step;
- study of the dependence of the results on additional dataset size;

[7] http://docs.deeppavlov.ai/en/master/features/models/bert.html

- three-step approach: independent sequential training in three stages using: the general data part from the additional dataset, the thematic examples from the additional dataset and the benchmark training sets.

All the results presented in the tables below are averaging over 3 experiments with different random initializations of models weights.

7.1 Mixing additional data with train samples

As a starting point for research, we train the models only on the automatically generated dataset (general and thematic neutral sentences). We compare two options of constructing the additional sample: uniform balancing between three sentiment classes and balancing in accordance with the average values of classes proportions for all datasets from Table 1.

For both options, the sample size was chosen equal to 15000. The results obtained with uniform balancing are 2-3 % higher and presented in Table 2. It can be seen, that performance is significantly lower than the current state-of-the-art results for all five benchmark datasets.

For the next step, we mix the automatically annotated data with the benchmark training sets. We keep the balance of sentiment classes from the previous experiment. The results are presented in Table 3. For accuracy and $F_1\ macro$ metrics, the results improve significantly but still did not reach state-of-the-art level. It could be probably explained by assumptions about different topics and styles of texts in additional and benchmark datasets and time dependence of automatically generated data (too many sentences about sports and New Year celebration).

Table 4: Results based on training on extended with sentiment thematic additional data mixed with the benchmark training sets.

Dataset	Model	Accuracy	$F_1\ macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
ROMIP-2013	BERT-single	66.78	62.44	71.49	70.61
	BERT-pair-QA	67.11	62.18	71.94	71.18
	BERT-pair-NLI	67.89	63.24	72.27	71.65
SRE-2015 Banks	BERT-single	70.54	66.18	67.31	66.59
	BERT-pair-QA	70.87	66.71	68.24	66.91
	BERT-pair-NLI	71.15	67.03	67.69	67.23
SRE-2015 Telecom	BERT-single	67.84	62.31	63.78	62.06
	BERT-pair-QA	68.35	62.44	64.21	62.51
	BERT-pair-NLI	68.89	62.71	65.02	63.12
SRE-2016 Banks	BERT-single	68.14	63.81	63.91	62.33
	BERT-pair-QA	68.81	64.42	65.43	64.16
	BERT-pair-NLI	69.21	65.02	65.76	65.59
SRE-2016 Telecom	BERT-single	67.31	62.15	63.28	61.68
	BERT-pair-QA	67.59	62.31	63.46	62.01
	BERT-pair-NLI	68.16	63.37	64.19	62.21

7.2 Extension of additional sample by thematic data

Analyzing low results of the previous experiment, we supposed it may be associated with topic differences between automatic and benchmark datasets, since at this stage an automatic sample was collected using personal descriptive words only. This way, we extend the additional dataset with sentiment thematic examples using the list of well-known organizations (banks and operators) and sentences obtained with NER from DeepPavlov, keeping sample size and sentiment class ratio unchanged.

The results are presented in Table 4. For all F_1 metrics, the performance seems much better than in the previous experiment (mixed general additional sample and training benchmark datasets), but still worse than current state-of-the-art results.

7.3 Two-step transfer learning approach

The two-step transfer learning consists in the sequential training on two samples and differs from the previous one in that we do not mix automatically generated data with benchmarks train sets. At the first step, the models are trained on the additional data, then model weights are frozen and training continues on the training data from the benchmarks.

During the same experiment, we study the dependence between the results and size of additional dataset. It was found that with increasing sample size, the results improve too. The boundary between extension of additional dataset and increasing the results was set at a sample size of 27000 (9000 per each sentiment class). Using the two-step approach allows us to overcome the current best results for almost all datasets. The results of described experiment and comparison with the state-of-the-art results [20, 7] are presented in Table 5.

Table 5: Results based on using the two-step approach.

Dataset	Model	Accuracy	$F_1\ macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
ROMIP-2013	BERT-single	79.95	71.16	85.39	85.61
	BERT-pair-QA	80.21	71.29	85.72	85.93
	BERT-pair-NLI	**80.56**	**71.68**	**86.14**	**86.19**
	Current SOTA	80.28	70.62	85.52	85.68
SRE-2015 Banks	BERT-single	86.06	79.11	64.87	66.73
	BERT-pair-QA	86.34	79.58	65.29	67.02
	BERT-pair-NLI	**87.62**	**80.72**	**68.44**	**71.39**
	Current SOTA	86.88	79.51	67.44	70.09
SRE-2015 Telecom	BERT-single	77.11	69.76	61.89	66.95
	BERT-pair-QA	**78.14**	**70.03**	**64.53**	**68.29**
	BERT-pair-NLI	77.96	69.68	64.52	68.21
	Current SOTA	76.63	68.54	63.47	67.51
SRE-2016 Banks	BERT-single	81.94	74.08	67.24	70.68
	BERT-pair-QA	**84.36**	**77.43**	**72.32**	**74.06**
	BERT-pair-NLI	84.19	75.63	68.52	70.89
	Current SOTA	82.28	74.06	69.53	71.76
SRE-2016 Telecom	BERT-single	75.82	69.78	65.04	74.22
	BERT-pair-QA	77.25	69.71	67.35	76.22
	BERT-pair-NLI	**77.59**	69.84	**68.11**	75.93
	Current SOTA	–	**70.68**	66.40	**76.71**

7.4 Three-step transfer learning approach

For the final experiment of the study, we divide the first step of the previous experiment into two: sequential training on the general and thematic data. At first, the models are trained on the general data, then the weights are frozen and the training continues on the thematic examples retrieved with the list of organizations and NER from DeepPavlov. After the second weights freezing, the last stage of learning on the original training samples begins. Taken together, this sequence represents the three-step transfer learning approach.

During this experiment, we also changed the additional sample by adding sentiment examples to thematic part of additional sample. The logic consisted in the selection among thematic sentences, those which contain sentiment words. Thus, the first step sample contains 18000 general examples and the second sample consists of 9000 thematic examples (both samples are equally balanced across sentiment classes).

The use of three-step approach combined with addition of sentiment thematic contexts to the sample, improved the results by a few more points. New state-of-the-art results as well as comparison with manual labelling for SentiRuEval-2015 telecom dataset are presented in Table 6. According to the organizers of SentiRuEval-2016 evaluation, one participant sent the results of manual annotation of the test

set [12]. As it can be seen, BERT-pair-NLI model reaches human sentiment analysis level by $F_1^{+-}micro$ metric.

Table 6: Results based on the three-step approach.

Dataset	Model	Accuracy	$F_1\ macro$	$F_1^{+-}macro$	$F_1^{+-}micro$
ROMIP-2013	BERT-single	80.27	71.78	85.82	86.07
	BERT-pair-QA	80.78	72.09	86.14	86.42
	BERT-pair-NLI	**82.33**	**72.69**	**86.77**	**87.04**
	Current SOTA	80.28	70.62	85.52	85.68
SRE-2015 Banks	BERT-single	87.65	80.79	65.74	67.46
	BERT-pair-QA	87.92	81.12	66.47	68.55
	BERT-pair-NLI	**88.14**	**81.63**	**68.76**	**72.28**
	Current SOTA	86.88	79.51	67.44	70.09
SRE-2015 Telecom	BERT-single	77.85	70.42	62.29	67.38
	BERT-pair-QA	**79.21**	70.94	65.68	69.11
	BERT-pair-NLI	79.12	**71.16**	**65.71**	**70.65**
	Current SOTA	76.63	68.54	63.47	67.51
	Manual [12]	–	–	70.30	70.90
SRE-2016 Banks	BERT-single	83.21	75.31	68.45	71.69
	BERT-pair-QA	**85.59**	**78.93**	**74.05**	**75.12**
	BERT-pair-NLI	85.43	76.85	70.23	72.07
	Current SOTA	82.28	74.06	69.53	71.76
SRE-2016 Telecom	BERT-single	76.79	70.64	66.16	75.27
	BERT-pair-QA	78.42	70.54	**68.65**	**77.45**
	BERT-pair-NLI	**78.62**	**71.18**	69.36	76.85
	Current SOTA	–	70.68	66.40	76.71

8 Conclusion

In this study, we presented a method for automatic generation of annotated sample from a Russian news corpus using distant supervision technique. We compared different options of combining additional data with benchmark train samples and improved current state-of-the-art results by more than 3% using BERT models together with the transfer learning approach. The best variant was three-step approach of sequential training on general, thematic and benchmark train samples with intermediate freezing of the model weights. On one of benchmarks, the BERT-NLI model treating a sentiment classification problem as a natural language inference task, reached human level according to one of the metrics.

Acknowledgments

The reported study was funded by RFBR according to the research project № 20-07-01059.

References

[1] Baymurzina DR, Kuznetsov DP, Burtsev MS. Language model embeddings improve sentiment analysis in Russian // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2019. — P. 53–62.

[2] Bird Steven Edward Loper, Klein Ewan. Natural Language Processing with Python. O'Reilly Media Inc. — 2009.

[3] Burtsev M. DeepPavlov: Open-Source Library for Dialogue Systems // Proceedings of ACL 2018, System Demonstrations. — 2018. — P. 122–127.

[4] Chetviorkin Ilia, Loukachevitch Natalia. Evaluating sentiment analysis systems in Russian // Proceedings of the 4th biennial international workshop on Balto-Slavic natural language processing. — 2013. — P. 12–17.

[5] Efficient Twitter sentiment classification using subjective distant supervision / Tapan Sahni, Chinmay Chandak, Naveen Reddy Chedeti, Manish Singh // 2017 9th International Conference on Communication Systems and Networks (COMSNETS) / IEEE. — 2017. — P. 548–553.

[6] Go Alec, Bhayani Richa, Huang Lei. Twitter sentiment classification using distant supervision // CS224N project report, Stanford. — 2009. — Vol. 1, no. 12. — P. 2009.

[7] Golubev Anton, Loukachevitch Natalia. Improving Results on Russian Sentiment Datasets // Proc. of the Artificial Intelligence and Natural Language (AINL 2020). — 2020. — P. 109–121.

[8] Kuratov Yu. Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". — 2019. — P. 122–127.

[9] Loukachevitch Natalia, Dobrov Boris V. RuThes linguistic ontology vs. Russian wordnets // Proceedings of the seventh global wordnet conference. — 2014. — P. 154–162.

[10] Loukachevitch Natalia, Levchik Anatolii. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). — 2016. — P. 1171–1176.

[11] Loukachevitch Natalia, Rubtsova Yuliya. Entity-oriented sentiment analysis of tweets: results and problems // International Conference on Text, Speech, and Dialogue / Springer. — 2015. — P. 551–559.

[12] Loukachevitch Natalia, Rubtsova Yuliya. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis // Proceedings of International Conference Dialog-2016//Proceedings of International Conference Dialog-2016. — 2016.

[13] Mohammad Saif, Salameh Mohammad, Kiritchenko Svetlana. Sentiment lexicons for Arabic social media // Proceedings of the tenth international conference on language resources and evaluation (LREC'16). — 2016. — P. 33–37.

[14] Moshkin Vadim, Konstantinov Andrey, Yarushkina Nadezhda. Application of the BERT Language Model for Sentiment Analysis of Social Network Posts // Russian Conference on Artificial Intelligence / Springer. — 2020. — P. 274–283.

[15] Overview of replab 2013: Evaluating online reputation monitoring systems / Enrique Amigó, Jorge Carrillo De Albornoz, Irina Chugur et al. // International conference of the cross-language evaluation forum for european languages / Springer. — 2013. — P. 333–352.

[16] Rubtsova Y. Constructing a corpus for sentiment classification training // Software and Systems. — 2015. — no. 109. — P. 72–78.

[17] Rusentiment: An enriched sentiment analysis dataset for social media in russian / Anna Rogers, Alexey Romanov, Anna Rumshisky et al. // Proceedings of the 27th International Conference on Computational Linguistics. — 2018. — P. 755–763.

[18] Rusnachenko Nicolay, Loukachevitch Natalia, Tutubalina Elena. Distant supervision for sentiment attitude extraction // Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — 2019. — P. 1022–1030.

[19] Smetanin Sergey, Komarov Mikhail. Sentiment analysis of product reviews in Russian using convolutional neural networks // 2019 IEEE 21st Conference on Business Informatics (CBI) / IEEE. — Vol. 1. — 2019. — P. 482–486.

[20] Smetanin Sergey, Komarov Mikhail. Deep transfer learning baselines for sentiment analysis in Russian // Information Processing & Management. — 2021. — Vol. 58, no. 3. — P. 102484.

[21] Sun Chi, Huang Luyao, Qiu Xipeng. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, V. 1. — 2019. — P. 380–385.

[22] spaCy: Industrial-strength Natural Language Processing in Python / Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd. — Zenodo, 2020.

[23] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.

A Corpus-Based Model of the English Phrasal Verb Construction: Attraction

Ekaterina Golubkova
Moscow State Linguistic University
Chaikina street, dom 6, kv. 150
125315 Moscow
The Russian Federation
katemg@yandex.ru

Alexander Trubochkin
Moscow State Linguistic University
Bibliotechnaya street, dom 16, kv. 12
141406 Moscow oblast, Khimki,
The Russian Federation
nalugu@mail.ru

Abstract

The article investigates the semantic of English phrasal verbs (PhVs) which are viewed as lexico-grammatical constructions. Triangulation of introspective, cognitive and corpus methods of analysis allows us to identify the semantic dimensions which feature the semantic pattern of the PhV-construction. The construction reveals the features of attraction involving new verbs provided the action or motion event is identical. Depending on the attraction strength level between the verb and the particle a new verb may be accepted to fill in the corresponding slot of the construction, which gives rise to a new phrasal verb. It allows us to categorise PhVs according to the attraction level and spot their PhV-patterns on corpus data.

Keywords: attraction; corpus data; phrasal verbs; construction
DOI: 10.28995/2075-7182-2021-20-278-288

Корпусно-когнитивное моделирование семантики фразово-глагольной конструкции: аттракция

Голубкова Екатерина
Московский государственный
лингвистический университет
125315, Россия, г. Москва,
ул. Чайкина, д. 6, кв. 150
katemg@yandex.ru

Трубочкин Александр
Московский государственный
лингвистический университет
141406, Россия, Московская область,
г. Химки, ул. Библиотечная, д. 16, кв. 12
nalugu@mail.ru

Аннотация

В статье рассматривается семантика фразовых глаголов в составе лексико-грамматических конструкций. Триангуляция методов интроспективного, корпусного и когнитивного анализов позволяет установить семантические характеристики фразово-глагольной конструкции. Конструкция обладает свойством аналогической аттракции, допускающей в орбиту конструкции новые лексические единицы. В зависимости от уровня аттракции между глаголом и частицей, определяемой через анализ корпусных данных, новый глагол может заполнять соответствующий слот фразово-глагольной конструкции, образуя новую единицу номинации в языке.

Ключевые слова: аттракция; корпусные данные; фразовые глаголы; конструкция

1 Introduction

According to the viewpoints of different researchers, the phrasal verb (PhV) is an indivisible linguistic unit with a certain structure. From this perspective, in line with a basic tenet of the theory of Construction Grammar, namely, that constructions are form–meaning pairings [3], so the meaning of construction

cannot be formed compositionally but is shaped by the interaction of semantics and grammar, we assume that the phrasal verb can be viewed as a construction, too.

The aim of the current research is to define characteristic features of the phrasal verb construction (PhV-construction) and to determine the leading factors due to which the semantics of the phrasal verb construction can change. To this effect, we argue that this semantic change is linked to a variable, which we call 'attraction strength' [1], which can be defined as the ability to collocate and, specifically for new verbs, to be accepted by the construction to fill in the corresponding slot specifying the integrity and the unambiguity of the construction represented by the phrasal verb.

Another task of this research is to investigate the semantic behavioural pattern of phrasal verbs that can establish interconnectedness between the elements of the phrasal verb construction by measuring and examining the attraction strength.

The statistical basis of the research was: The BNC [10] and The Intelligent Web-based Corpus iWEB [11]. The experimental base was the phrasal verb cluster 'Leaving' (45 phrasal verbs) with the particle *out* taken from Longman phrasal verb dictionary [8]. The methods used to conduct the following measurements are: the collexeme analysis [4], the polynomial approximation of the result data which is used to describe alternately ascending and descending values for the analysis of a sizeable dataset of an unstable value.

2 Attraction strength in phrasal verb constructions

2.1 Initial data

To start the analysis of the functioning of a phrasal verb construction, we will turn to one of the clusters in the segment of phrasal verbs accompanying the particle *out*, namely, the cluster 'Leaving' of 45 phrasal verbs displayed in Tables 1 and 2, and try to reveal some semantic dimensions of an action associated with the phrasal verb construction based on empirical data assigned to the amount of contribution of the agent to perform an action, where three degrees of intensity are singled out: low contribution = 1, average contribution = 2, high contribution = 3.

The intensity is the empirical quantitative parameter which specifies semantic dimensions of action based on the data retrieved from the BNC [10] and represents the semantics of phrasal verbs. Table 1 indicates a random distribution of the degree of intensity of the semantic dimensions of action (manner, strain, speed, duration, intention, morality, physicality, reversibility etc.) among the phrasal verbs under analysis. The classification of manner adverbs and the semantic dimensions is based on the offline introspection analysis [7] involving a native English speaker from the UK in the experiment. After Talmy we assume that the component of manner of action in phrasal verbs is likely to be expressed within the verb itself. We added a few semantic dimensions to the general concept of manner, relying on the poll taken with the native speaker of English, and placed them in the table header with a view to indicate the intensity of each semantic dimension corresponding to each test phrasal verb. Thus, Table 1 prototypes the semantics of the phrasal verbs in a digital manner, which we call 'the semantic matrix' of a phrasal verb cluster.

In order to uncover the possible regularity of change of contribution of the semantic dimensions depicted in Table 1, we research the behaviour pattern of the phrasal verbs using the collexeme method of analysis [2].

2.2 Attraction of verbs to the 'Verb+out' construction using collexeme analysis

In order to measure attraction, we apply the collexeme analysis to estimating the attraction of the verb (and the particle further) attracted by the slots of the construction. The collexeme analysis that deals with indivisible items such as lexemes appears to be applicable to our task because, from the viewpoint of Construction Grammar, constructions are already inseparable units, which enable us to substitute them for lexemes in the co-lexeme analysis. Moreover, the algorithm of co-lexeme analysis is not mathematically cumbersome and consists in probability calculation and comparison of the probability of success (positive outcome) of a certain word form of a certain lexeme in the corpus with the threshold value that is defined as the probability of success of the corresponding word form of all the lexemes of the same part of speech in the corpus. In terms of Construction Grammar, it comes to the calculation of

the probability of success[1] of a certain phrasal verb construction in the corpus compared to the threshold value that is defined as the probability of success of the Verb+*out* construction, in other words, the probability of success of all the analogous constructions in the corpus. Having compared these two values, we get the value of attraction. Thus, the collexeme analysis is chosen as the most convenient method for our research.

Phrasal verb		Semantic dimensions (aspects of action)									
Verb	Particle	Manner	Strain 1- low 2- avr, 3- high	Speed 1- low 2- avr. 3- high	Duration 1- low 2- avr. 3- high	Intention 1- unint. 2- hesitat. 3- intent.	Morality 1- immoral 2- suspect 3- moral	Physicality 1- nearly 0 2- limited 3- real	Reversibility 1- irrevers. 2- partially 3- reversable	Toolability 1- toolless 2- auxilary 3- toolfull	Agents 1- one 2- a few 3- a lot
allow	**out**	Controlled	1	3	1	2	2	3	3	1	1
back	**out**	Renegade	2	3	3	3	1	1	2	2	1
bail	**out**	Forceful	3	3	1	3	3	3	1	3	1
break	**out**	Challenging	3	3	1	3	1	1	3	2	1
breeze	**out**	Lightharted	1	3	1	3	2	3	3	1	1
bug	**out**	Disorderly	3	3	1	3	2	3	1	1	1
bust	**out**	Secretive	3	3	1	3	1	3	1	2	1
buy	**out**	Gentle force	2	3	3	3	2	1	1	2	1
coax	**out**	Careful	1	1	3	3	2	3	3	1	1
check	**out**	Orderly	1	3	1	3	3	2	3	3	1
clear	**out**	Forceful	2	3	1	3	2	3	3	2	1
clock	**out**	Orderly	1	3	1	3	3	2	3	3	1
come	**out**	Neutral	2	2	2	3	2	3	3	1	1
draw	**out**	Careful	1	1	3	3	2	3	3	1	1
duck	**out**	Secretive	3	3	1	3	1	1	1	1	1
encourage	**out**	Careful	1	1	3	3	2	3	3	1	1
fall	**out**	Accidental	1	3	1	1	2	3	3	1	1
fly	**out**	Forceful	3	3	1	2	2	3	3	1	1
get	**out**	Neutral	2	3	2	3	2	1	3	1	1
go	**out**	Neutral	2	3	2	3	2	3	3	1	1
let	**out**	Controlled	1	3	1	2	2	3	3	1	1
light	**out**	Disorderly	2	3	1	3	2	3	3	1	1
log	**out**	Orderly	1	3	1	3	2	1	3	3	1
move	**out**	Orderly	1	1	3	3	2	3	3	1	1
pile	**out**	Disorderly	3	2	2	2	2	3	3	1	3
pop	**out**	Sudden	2	3	1	1	2	3	3	1	3
pour	**out**	Controlled	2	2	2	3	2	3	3	1	3
pull	**out**	Controlled	3	1	2	3	2	1	3	1	1
punch	**out**	Desperate	3	1	2	3	2	2	1	1	1
put	**out**	Forceful	3	3	1	3	1	3	3	1	1
run	**out**	Desperate	3	3	1	3	1	3	2	1	1
sally	**out**	Agressive	3	3	1	3	2	3	1	1	3
see	**out**	Respectful	1	2	2	3	3	3	1	1	1
set	**out**	Orderly	1	3	1	3	2	3	3	1	1
ship	**out**	Specific	1	1	2	3	2	3	1	3	1
shoot	**out**	Sudden	3	3	1	2	2	3	3	1	1
show	**out**	Friendly	1	2	2	3	3	3	1	1	1
sign	**out**	Orderly	1	1	1	3	2	1	3	3	1
slip	**out**	Secretive	2	1	2	3	1	3	3	1	1
spill	**out**	Uncontrolled	3	2	3	1	2	3	1	1	3
start	**out**	Orderly	1	3	2	3	2	2	3	1	1
step	**out**	Orderly	2	3	2	1	2	3	3	1	1
storm	**out**	Agressive	3	1	2	3	1	3	1	1	1
strike	**out**	Decisive	2	3	1	3	2	2	3	3	1
want	**out**	Reluctant	1	1	2	3	2	1	3	1	1

Table 1: A random distribution of 45 phrasal verbs in cluster 'Leaving' and the values of intensity of their semantic dimensions (the semantic matrix of the phrasal verb cluster 'Leaving')

Table 2 indicates the results of the queries to the corpus necessary to calculate attraction strength of the phrasal verbs under analysis. The value of attraction in the 0-line of Table 2 indicates the attraction threshold P(threshold) = 0,008 of the Verb+*out* construction. In other words, the verbs with the value of P(a)[2] > 0,008 are attracted by the construction and if P(a) < 0,008 then the construction repels them.

[1] The Probability of success, known as one of the key decision factors in Probability Theory, is the ratio of success cases or, in terms of our research, desired occurrences of specific lexical items (in particular, verbs, particles or PhV-constructions) over all outcomes of the same kind derived from the corpus data.

[2] In the paper we call the attraction strength P(a), the attraction threshold – P(threshold). P is a capital to not be confused with the p-value in statistics which we conduct to assess the reliability of the findings (Section 2.3, cf. Table 4). We assigned attraction strength to the capital P since the calculation of attraction is strongly connected to the calculation of the probability of success.

Phrasal verb			Variable B		Variable C		Variable D
№	Verb	Particle	Occurrence of verbs	RegEx corpus query for verbs	Occurrence of the Verb+[Pron]+OUT constructions	RegEx corpus query for phrasal verbs	Attraction **P(a)** of the verb to the construction
0	all verbs		15735322	VERB+ deduct modals _vm	125895	VERB+ out_rp *add:* VERB+ _pp out_rp	**0,008**
1	**storm**	**out**	659	STORM_v	95	STORM_v out, STORM_v _pp out	0,1442
2	**pull**	**out**	12921	PULL_v	1747	PULL_v out, PULL_v _pp out	0,1352
3	**sally**	**out**	39	SALLY_v	5	SALLY_v out, SALLY_v _pp out	0,1282
4	**step**	**out**	5520	STEP_v	692	STEP_v out, STEP_v _pp out	0,1254
5	**bail**	**out**	355	BAIL_v	143	BAIL_v out, BAIL_v _pp out	0,1211
6	**set**	**out**	38829	SET_v	4608	SET_v out, SET_v _pp out	0,1187
7	**pour**	**out**	3448	POUR_v	391	POUR_v out, POUR_v _pp out	0,1134
8	**spill**	**out**	1335	SPILL_v	151	SPILL_v out, SPILL_v _pp out	0,1131
9	**pop**	**out**	1956	POP_v	154	POP_v out, POP_v _pp out	0,0787
10	**slip**	**out**	4667	SLIP_v	339	SLIP_v out, SLIP_v _pp out	0,0726
11	**duck**	**out**	581	DUCK_v	40	DUCK_v out, DUCK_v _pp out	0,0688
12	**break**	**out**	17394	BREAK_v	1108	BREAK_v out, BREAK_v _pp out	0,0637
13	**check**	**out**	9355	CHECK_v	592	CHECK_v out, CHECK_v _pp out	0,0633
14	**run**	**out**	38304	RUN_v	2139	RUN_v out, RUN_v _pp out	0,0558
15	**strike**	**out**	7059	STRIKE_v	333	STRIKE_v out, STRIKE_v _pp out	0,0472
16	**come**	**out**	143322	COME_v	6435	COME_v out, COME_v _pp out	0,0449
17	**back**	**out**	4150	BACK_v	177	BACK_v out, BACK_v _pp out	0,0427
18	**punch**	**out**	911	PUNCH_v	38	PUNCH_v out, PUNCH_v _pp out	0,0417
20	**fly**	**out**	8571	FLY_v	339	FLY_v out, FLY_v _pp out	0,0396
21	**ship**	**out**	1562	SHIP_v	60	SHIP_v out, SHIP_v _pp out	0,0384
19	**bust**	**out**	236	BUST_v	9	BUST_v out, BUST_v _pp out	0,0381
22	**clear**	**out**	6094	CLEAR_v	230	CLEAR_v out, CLEAR_v _pp out	0,0377
23	**go**	**out**	236313	GO_v	8493	GO_v out, GO_v _pp out	0,0359
24	**coax**	**out**	307	COAX_v	10	COAX_v out, COAX_v _pp out	0,0326
25	**shoot**	**out**	7203	SHOOT_v	234	SHOOT_v out, SHOOT_v _pp out	0,0325
26	**log**	**out**	483	LOG_v	14	LOG_v out, LOG_v _pp out	0,029
27	**get**	**out**	211006	GET_v	6010	GET_v out, GET_v _pp out	0,0285
28	**fall**	**out**	25843	FALL_v	714	FALL_v out, FALL_v _pp out	0,0276
29	**move**	**out**	37290	MOVE_v	971	MOVE_v out, MOVE_v _pp out	0,026
30	**draw**	**out**	21401	DRAW_v	519	DRAW_v out, DRAW_v _pp out	0,0243
31	**put**	**out**	67040	PUT_v	1616	PUT_v out, PUT_v _pp out	0,0241
32	**let**	**out**	34194	LET_v	785	LET_v out, LET_v _pp out	0,023
33	**clock**	**out**	349	CLOCK_v	6	CLOCK_v out, CLOCK_v _pp out	0,0172
34	**start**	**out**	39316	START_v	491	START_v out, START_v _pp out	0,0125
35	**pile**	**out**	1012	PILE_v	12	PILE_v out, PILE_v _pp out	0,0119
36	**bug**	**out**	198	BUG_v	1	BUG_v out, BUG_v _pp out	0,0101
37	**buy**	**out**	24741	BUY_v	232	BUY_v out, BUY_v _pp out	0,0094
38	**allow**	**out**	31422	ALLOW_v	107	ALLOW_v out, ALLOW_v _pp out	0,0034
39	**light**	**out**	3365	LIGHT_v	11	LIGHT_v out, LIGHT_v _pp out	0,0033
40	**want**	**out**	86579	WANT_v	179	WANT_v out, WANT_v _pp out	0,0021
41	**sign**	**out**	8782	SIGN_v	14	SIGN_v out, SIGN_v _pp out	0,0016
42	**see**	**out**	181678	SEE_v	204	SEE_v out, SEE_v _pp out	0,0011
43	**show**	**out**	57617	SHOW_v	38	SHOW_v out, SHOW_v _pp out	0,0007
44	**encourage**	**out**	44	ENCOURAGE_v	0	ENCOURAGE_v out, ENCOURAGE _pp out	0
45	**breeze**	**out**	11073	BREEZE_v	0	BREEZE_v out, BREEZE_v _pp out	0

Table 2: An ordered distribution of 45 phrasal verbs in cluster 'Leaving' according to their attraction strength to the PhV-construction – Variable D

The value of the variable B in the 0-line indicates the number of instances of all verbs in any form represented in the corpus except all the modals as they do not shape phrasal verbs. The value of the variable C in the same line indicates the occurrence of the Verb+*out* phrasal verb constructions with any form of all verbs found in the corpus. Other lines indicate the same values but regarding the number of instances of the particular verb and the variable D represents the probability of success which is calculated by the formula P(a) = C÷B. Comparing this result value of each line ($C_N \div B_N$) with the 0-line ($C_0 \div B_0$), we can get the attraction strength of each tested verb to the Verb+*out* construction. To represent the data, we grade the phrasal verbs from Table 1 according to their attraction strength[3] and put them in Table 2.

This distribution of the phrasal verbs to attraction strength reveals three distinct groups:

(a) Group 1 with high attraction strength (coloured green);
(b) Group 2 with moderate attraction strength (coloured white);
(c) Group 3 with low attraction strength (coloured red).

[3] Attraction strength is assigned to the comparison of values of the variable D for each verb with the attraction threshold displayed in the 0-line P(threshold) = 0,008, which allows us to grade phrasal verbs according to their attraction strength in a descending sequence.

It can be seen in Figure 1 below that the phrasal verbs with the particle *out* fall into 3 groups:

(a) Group 1 takes the value of attraction strength P(a) > 0,8;
(b) Group 2 takes the value of P(a) which falls in P(threshold) ≤ P(a) ≤ 0,8;
(c) Group 3 takes the value of P(a) < P(threshold), where P(threshold) is at 0,008 marked with the red line in Figure 1.

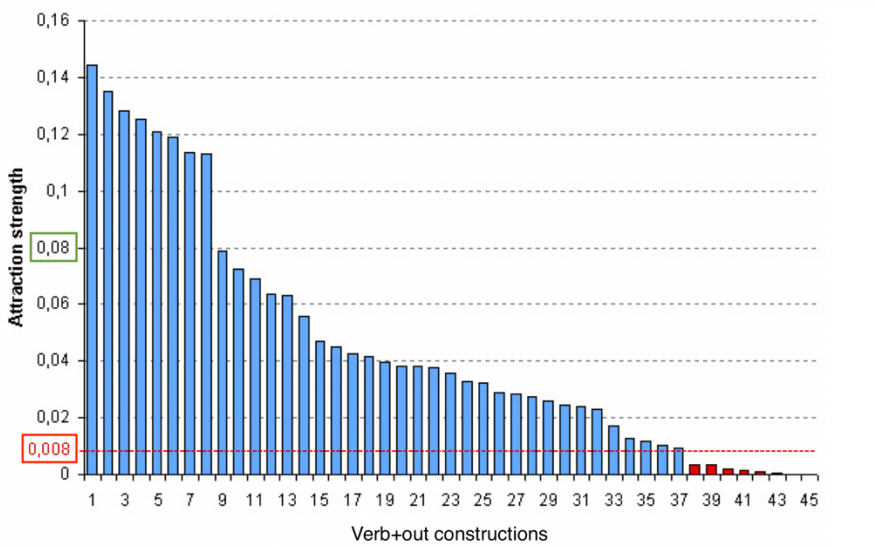


Figure 1: The distribution of the attraction strength of the verbs to the Verb+*out* construction

According to the attraction strength of the verbs to the Verb+*out* construction (cf. Figure 1 and Table 2) we regroup the phrasal verbs in Table 1 as follows, in Table 3.

The distribution of the verbs inside Group 2 (coloured white) shown in Table 3 suggests that the verbs with more intensity of strain (the value is 3) and manner tend towards Group 1 (coloured green) and the verbs with less intensity of strain (the value is 1) and manner tend towards Group 3 (coloured red). The choice of the intensity value was guided by the experiment on the basis of the behavioral S → R scheme [9], or the stimulus–reaction scheme, in which the native speaker of English was instructed to evaluate their reaction response for each semantic dimension to a given stimulus – as soon as a construction with a test phrasal verb was uttered by another participant in the experiment.

After the experiment all the collected data were analysed from the viewpoint of the offline introspection [7], assigned with an integer value from 1 to 3 and put in Table 1, which allowed us to arrange the data by the value of specific dimensions such as 'manner' or 'strain' and present them in Table 2.

The preliminary observation of the arranged data leads us to two assumptions:

(i) Phrasal verbs with the manner of action, such as aggressive, forceful, tend to belong to Group 1, and verbs with the opposite manner, such as friendly, careful, lighthearted, reluctant, respectful, gentle, tend to belong Group 3 in accordance with the attraction strength of the verb to the phrasal verb construction. Thus, the weaker attraction strength to the construction the verb has, the 'softer' the manner of the verb is, while the more attraction strength the verb has, the 'harder' its manner is.
(ii) Phrasal verbs with greater 'Strain' tend to the top of this category revealed by the distribution in Table 3 (Group 1) and phrasal verbs with weak 'Strain' tend to stay at the bottom (Group 3). Thus, the greater attraction strength to the construction the verb has, the greater strain of the action assigned to the verb is.

This inference can be observed in Figure 3 in comparison with Figure 2. The diagram in Figure 2 indicates the behaviour of the semantic dimension 'Strain' at the random distribution (cf. Table 1) of phrasal verbs where we observe no dependence of the semantic dimension on the distribution.

Phrasal verb		Semantic dimensions (aspects of action)									
Verb	Particle	Manner	Strain 1- low 2- avr, 3- high	Speed 1- low 2- avr. 3- high	Duration 1- low 2- avr. 3- high	Intention 1- unint. 2- hesitat 3- intent.	Morality 1- immoral 2- suspect 3- moral	Physicality 1- nearly 0 2- limited 3- real	Reversibility 1- irrevers. 2- partially 3- reversable	Toolability 1- toolless 2- auxilary 3- toolfull	Agents 1- one 2- a few 3- a lot
storm	out	Agressive	3	1	2	3	1	3	1	1	1
pull	out	Controlled	3	1	2	3	2	1	3	1	1
sally	out	Agressive	3	3	1	3	2	3	1	1	3
step	out	Orderly	2	3	2	1	2	3	3	1	1
bail	out	Forceful	3	3	1	3	3	3	1	3	1
set	out	Orderly	1	3	1	3	2	3	3	1	1
pour	out	Controlled	2	2	2	3	2	3	3	1	3
spill	out	Uncontrolled	3	2	3	1	2	3	1	1	3
pop	out	Sudden	2	3	1	1	2	3	3	1	3
slip	out	Secretive	2	1	2	3	1	3	3	1	1
duck	out	Secretive	3	3	1	3	1	1	1	1	1
break	out	Challenging	3	3	1	3	1	1	3	2	1
check	out	Orderly	1	3	1	3	3	2	3	3	1
run	out	Desperate	3	3	1	3	1	3	2	1	1
strike	out	Decisive	2	3	1	3	2	2	3	3	1
come	out	Neutral	2	2	2	3	2	3	3	1	1
back	out	Renegade	2	3	3	3	1	1	2	2	1
punch	out	Desperate	3	1	2	3	2	2	1	1	1
fly	out	Forceful	3	3	1	2	2	3	3	1	1
ship	out	Specific	1	1	2	3	2	3	1	3	1
bust	out	Secretive	3	3	1	3	1	3	1	2	1
clear	out	Forceful	2	3	1	3	2	3	3	2	1
go	out	Neutral	2	3	2	3	2	3	3	1	1
coax	out	Careful	1	1	3	3	2	3	3	1	1
shoot	out	Sudden	3	3	1	2	2	3	3	1	1
log	out	Orderly	1	3	1	3	2	1	3	3	1
get	out	Neutral	2	3	2	3	2	1	3	1	1
fall	out	Accidental	1	3	1	1	2	3	3	1	1
move	out	Orderly	1	1	3	3	2	3	3	1	1
draw	out	Careful	1	1	3	3	2	3	3	1	1
put	out	Forceful	3	3	1	3	1	3	3	1	1
let	out	Controlled	1	3	1	2	2	3	3	1	1
clock	out	Orderly	1	3	1	3	3	2	3	3	1
start	out	Orderly	1	3	2	3	2	2	3	1	1
pile	out	Disorderly	3	2	2	2	2	3	3	1	3
bug	out	Disorderly	3	3	1	3	2	3	1	1	1
buy	out	Gentle force	2	3	3	3	2	1	1	2	1
allow	out	Controlled	1	3	1	2	2	3	3	1	1
light	out	Disorderly	2	3	1	3	2	3	3	1	1
want	out	Reluctant	1	1	2	3	2	1	3	1	1
sign	out	Orderly	1	1	1	3	2	1	3	3	1
see	out	Respectful	1	2	2	3	3	3	1	1	1
show	out	Friendly	1	2	2	3	3	3	1	1	1
encourage	out	Careful	1	1	3	3	2	3	3	1	1
breeze	out	Lightharted	1	3	1	3	2	3	3	1	1

Table 3: An ordered distribution of the phrasal verbs according to their attraction strength and the intensity of their semantic dimensions

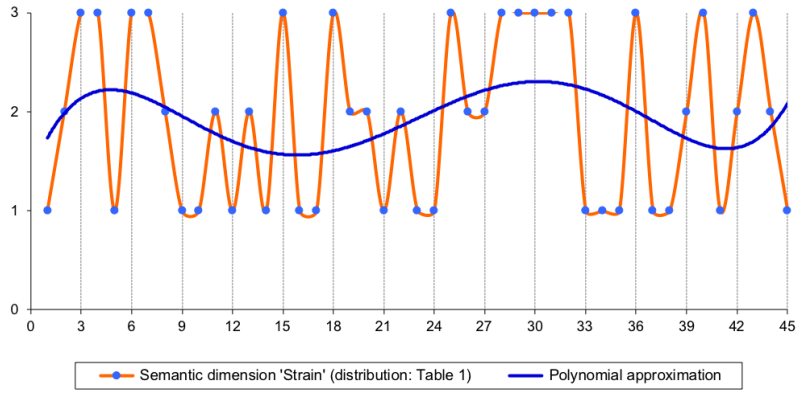


Figure 2: A random distribution of empirical data of the semantic dimension ‘Strain’ of 45 test phrasal verbs according to Table 1.

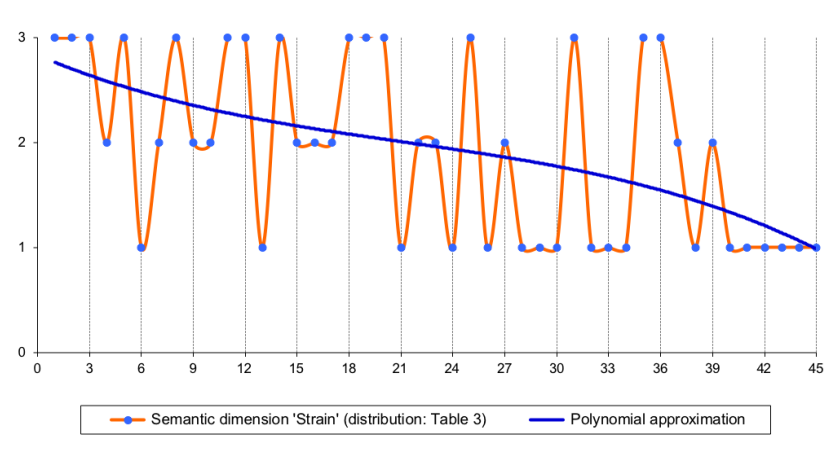


Figure 3: An ordered distribution of empirical data of the semantic dimension 'Strain' of 45 test phrasal verbs according to Table 3.

An overlaid trend line (coloured blue) shows **the behaviour pattern** of phrasal verbs regarding the semantic dimension of action 'Strain' which can be easily seen if we apply polynomial approximation known as the easiest conventional method to generalise empirical result data.

2.3 Attraction of the particle OUT to the Verb+*out* construction

The value of attraction of the particle can be measured by using collexeme analysis based on the corpus data. According to Gries's [2] method of defining attraction strength, the threshold value of particle attraction was calculated (0.2742) which further should be compared with the ratio of occurrences of the *out* in the corpus (0.7810) which is 3 times as high as the threshold value. It led us to conclusion that the particle is strongly attracted to the verbal form. This level of attraction, as we can see, is strong enough to let us consider most cases of 'Verb+*out*' as an integral unit. As a result, we get the set of values of attraction strength of the particle *out* to the corresponding verbal construction.

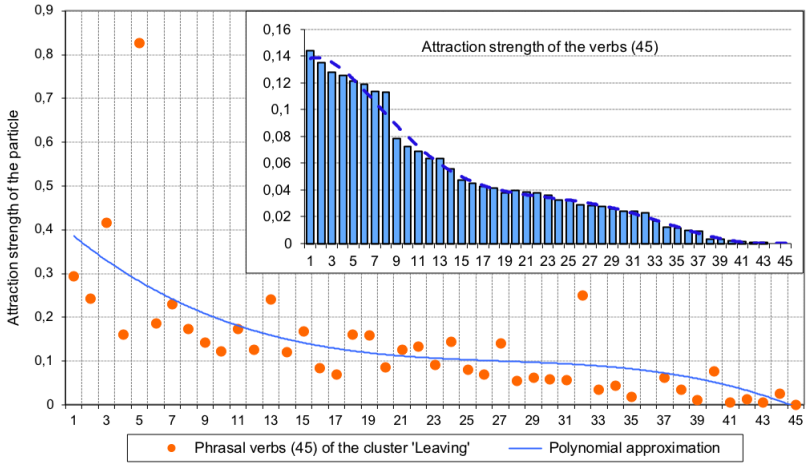


Figure 4: The particle attraction strength to the Verb+*out* construction

The diagram in Figure 4 shows the distribution of the particle *out* attraction strength to the construction 'Verb+*out*', where the polynomial approximate curve indicates the same trend as shown above (cf. Figure 1) which confirms the attraction force between the particle *out* and the verb. In its turn, it illustrates a steady correlation between the particle and the verb as if they function as an indivisible unit. The trends of mutual attraction between the verb and the particle also coincide with the trend of the semantic dimension 'Strain' of the corresponding phrasal verbs (cf. Figure 3). These concordant trends make it

possible to assume that the semantic dimensions ‘strain’, ‘manner’ and the attraction strength also become concordant.

In fact, having graded the result data according to the attraction strength of the particle *out*, in Figure 5 we show the correlation between the attraction of the particle *out* to each of the 45 tested phrasal verbs and the change of ‘Strain’ which is their semantic dimension of action. This correlation is also confirmed by the correlation matrix (cf. Table 4) in which Pearson correlation coefficient (PCC) takes the value of 0.464 for the verb and 0.422 for the particle in respect of the correlation between the attraction level and the change of the semantic dimension ‘strain’ of the tested phrasal verbs. The PCC values of 0.337 and 0.353 account for the correlation between the semantic change of manner and the level of attraction between the verb and the particle respectively, which indicates the positive leaner correlations in either case.

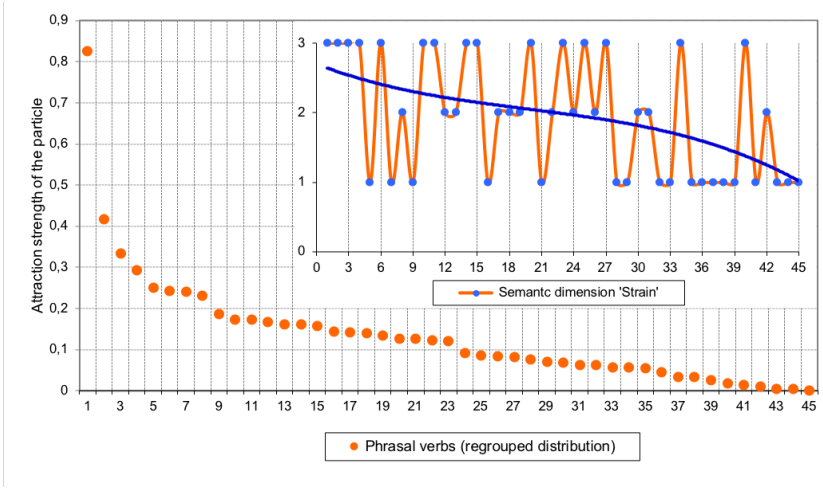


Figure 5: The correlation between the attraction strength of the particle to the PhV-construction and the semantic dimension ‘Strain’ of the tested phrasal verbs

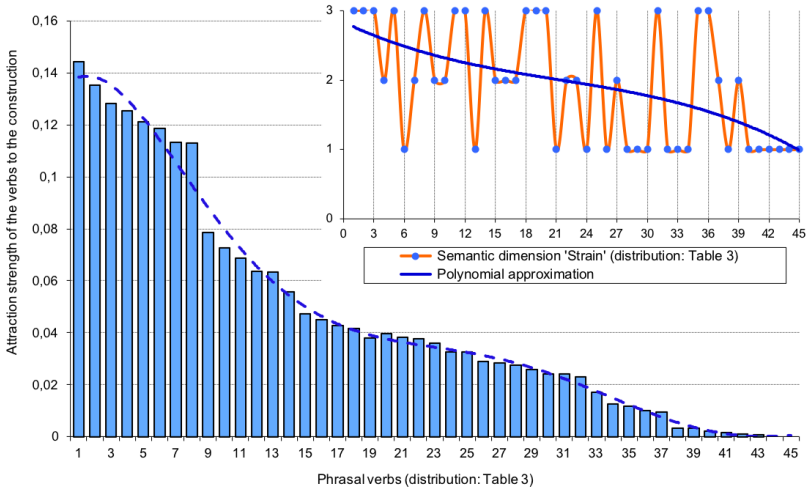


Figure 6: The correlation between the attraction strength of the tested phrasal verbs to the PhV-construction and their semantic dimension ‘Strain’

The statistical significance check p-values of $0.001 < 0.05$ (for the verb) and $0.004 < 0.05$ (for the particle) towards the correlation between the attraction level and the semantic change of the aspect ‘Strain’ along with the p-values of $0{,}024 < 0.05$ (for the verb) and $0.017 < 0.05$ (for the particle) towards the correlation between the attraction level and the semantic change of manner suggest that attraction features the change of certain semantic dimensions of phrasal verbs, in particular ‘strain’ and manner, where ‘strain’ stands for the amount of energy involved in performing an action.

Correlation Matrix

		Semantic aspect 'Strain'	Semantic aspect 'Manner'	Attraction of verb	Attraction of particle
Semantic aspect 'Strain'	Pearson's r	—			
	p-value	—			
Semantic aspect 'Manner'	Pearson's r	0.269	—		
	p-value	0.074	—		
Attraction of verb	Pearson's r	0.464 **	0.337 *	—	
	p-value	0.001	0.024	—	
Attraction of particle	Pearson's r	0.422 **	0.353 *	0.650 ***	—
	p-value	0.004	0.017	< .001	—

Note. * p < .05, ** p < .01, *** p < .001

Table 4: The correlation matrix of attraction and semantic variables of the tested phrasal verbs (processed by Jamovi statistical software platform [12])

Evidently, all the considered p-values are less than the conventional statistical significance threshold p = 0.05 and in case of the semantic dimension 'strain' the p-values are less than the 0.01-threshold, which suggests that the correlations are statistically significant and confirms the hypothesis.

Whereas the distribution of the value of the semantic dimension 'Strain' is also affected by the particle, which can be seen from the comparison of the built-in diagrams in Figures 5 and 6, the tendency remains the same keeping agreement with the data distributions of the verbs and the particle attraction strength. These are shown in the diagrams in Figures 1 and 4 where their interdependence can be easily traced, a fact that demonstrates a verb-particle behaviour dependence. This behaviour pattern is represented in Figure 7.

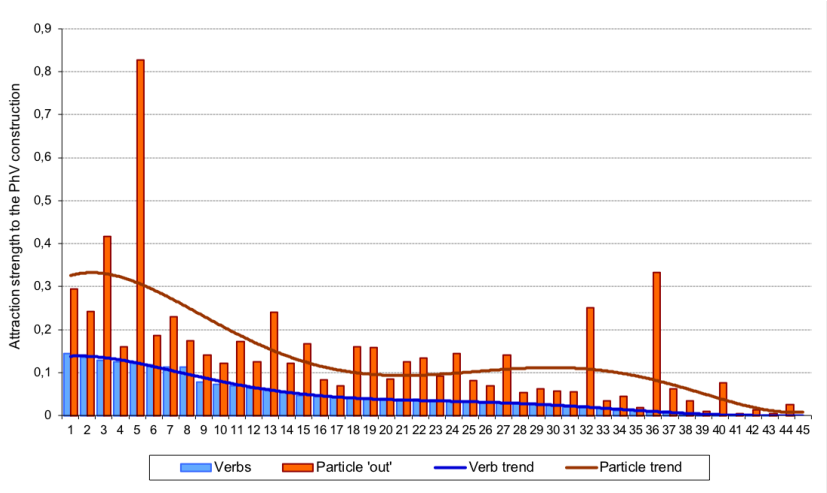


Figure 7: The correlation between the attraction strengths of the verb and the particle

The results suggest coordination between the particle, verb and some semantic dimensions (for example, manner, strain) that shapes an indivisible specific constructional unity allowing new verbs which exceed the attraction threshold set by the construction into the corresponding slot of the construction. These verbs provide for the specification of the meaning that corresponds to the meaning of the general construction which, in its turn, obtains this specification from the situation to which it is eventually linked. If the situation of communication has such specifications, that is, we deal with a specific form of the situation, it forces the construction into changing, attracting new verbs which are capable of conforming to the meaning of the situation in each specific case. This constructional unity can be called a phrasal verb construction (PhV construction). It seems reasonable to single out a specific phrasal verb construction that can retain its form and hold the general phrasal verb construction as an embedded structure which has to acquire a new form whenever the situation changes, for example (cf. Table 5).

Given the possibility of measuring attraction we can arrange phrasal verbs in new clusters of synonyms according to their level of attraction strength that defines the level of their stability in constructions. The higher level of attraction the verb has, the more stable the verb is. Then such constructions are also more stable, which also accounts for their indivisibility, and can be freely understood by the listener even if they are entirely new.

New PhVs based on analogy	General PhV construction	Semantic modification	Specific PhV construction
to coffee up	X bucks Y up X perks Y up	of manner: using coffee	I had to work the night shift so I **coffeed** myself **up** numerous times.
to tea down	X calms Y down	of manner: using tea	He seems so keyed up, we can try to **tea** him **down**.
to burn by	X comes by [prep] Y	of manner: quicker	Although she loves her, she rarely **burns by** at her mother's.
to spirit down	X brings Y down	of result: less animated	He was excited about new project. We had to **spirit** him **down** before the investors came.
to fall near	X comes near [prep] Y	of manner: unexpectedly to get to know	He offered me a senior post soon after we had **fallen near** at the congress.

Table 5: Attraction of new members of the PhV construction: semantic conformation of the general PhV construction to the meaning of the specific PhV construction

Constituting a PhV construction the particle and the verb hold mutual attraction to a Verb+Particle pattern forming a PhV lexical unit which steadily correlates with certain semantic dimensions disclosing the semantic behavioural pattern of the unit.

3 Conclusion

In the present paper we view phrasal verbs as lexico-grammatical constructions in line with the theory of Construction Grammar. Given this concept, the results of our investigation led us to believe that verb-particle attraction contributes significantly to shaping the set of semantic dimensions such as 'strain' and 'manner' of the phrasal verb, which could be expressed through the level of verb-particle attraction strength and subsequently digitalised. It allows us to represent the semantics of phrasal verbs through the semantic matrix, in which the values correspond to the verb-particle attraction levels. The results suggest the possibility of classifying phrasal verbs by verb-particle attraction levels, which play an important role in phrasal verb production.

Depending on the level of verb-particle attraction strength a new participant may be accepted to fill in the corresponding slot of the construction, which gives rise to a new phrasal verb. It allows us to categorise PhVs according to the attraction level and recognize their PhV-patterns.

Following the results of the comparison of the attraction indexes of both verbs and particles, it was demonstrated that the particle is much more stable than the verb in a phrasal verb construction, which also confirms the typology of English as a satellite-framed language [6] from the viewpoint of Corpus Linguistics. This fact enables us to conclude that the verb takes an open position in the construction, and can be replaced by a new verb which is attracted or 'invited' into the construction on terms of sufficient attraction strength exceeding the attraction threshold or otherwise repelled due to semantic restrictions. Thus, the new participants which may be accepted by the construction are verbs. As a consequence of this acceptance any new participant shapes a new phrasal verb. That is to say, attraction acts between linguistic constituents of the construction pulling in more and more new participants (verbs) and shaping more and more phrasal verbs according to the same PhV pattern.

The results also indicated the presence of coordination between verb-particle attraction and the semantic dimensions 'manner' and 'strain' involved in the description of the action or motion event [5], revealing the strength of attraction which admits new verbs into the construction triggering the corresponding semantic change of the meaning of the construction.

References

[1] Golubkova E.E., Trubochkin A.V. (2019), Phrasal Verbs from the Viewpoint of Construction Grammar in Modern English [Frazovye glagoly kak grammaticheskie konstrukcii (na materiale anglijskogo jazyka)], Cognitive Studies of Language. Integrative Processes in Cognitive Linguistics: Papers of International Congress on Cognitive Linguistics. May, 16–18, 2019 [Kognitivnye issledovanija jazyka. Integrativnye processy v kognitivnoj lingvistike: Materialy Mezhdunarodnogo kongressa po kognitivnoĭ lingvistike 16–18 maja 2019 goda], Vol. 37, pp. 604–608. Access mode: https://nnov.hse.ru/mirror/pubs/share/direct/266828046.

[2] Gries Stefan, Stefanowitsch Anatol. Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations' // International Journal of Corpus Linguistics. — 2004. — Vol. 9 (1), P. 97–129. Access mode: doi.org/10.1075/ijcl.9.1.06gri.

[3] Langacker Ronald. Construction Grammars: cognitive, radical and less so. // Paper presented at the international Cognitive Linguistics Conference. — Logroño, 2003.

[4] Rakhilina E.V. (2010), Construction Linguistics [Lingvistika konstrukcij], Moscow : Azbukovnik, pp. 35–39. Access mode: http://rakhilina.ru/files/rakh_lingconst.pdf.

[5] Talmy Leonard. Path to Realization: A Typology of Event Conflation // Proceedings of the Seventeenth Annual Meeting of the BLS. — 1991. — P. 480–519. Access mode: doi.org/10.3765/bls.v17i0.1620.

[6] Talmy Leonard. Toward a cognitive semantics. — Cambridge, MA : MIT Press, 2000. — Vol. 2.

[7] Talmy Leonard. Introspection as a Methodology in Linguistics // Proceedings of the Tenth International Cognitive Linguistic Conference. — Buffalo, USA, 2007. — P. 1–20. Access mode: http://www.acsu.buffalo.edu/~talmy/talmyweb/Handouts/introspection2.pdf.

[8] Taylor Andrew. Longman Phrasal Verb Dictionary Paper (Phrasal Verb Dictionary). — Harlow : Pearson Education Limited, 2000. — P. 35–36. Access mode: www.pearson.com/english/catalogue/dictionaries/browse/specialised/phrasal-verbs-dictionary.html.

[9] Watson John. Behaviorism (revised ed.). — Chicago, USA : University of Chicago Press, 1930 (1924). Access mode: OCLC https://www.worldcat.org/title/behaviorism/oclc/3124756, https://archive.org/details/behaviorism032636mbp/page/n259/mode/2up.

Corpora

[10] BNC World. The British National Corpus: 100 million words. — Oxford University Computing Services on behalf of the BNC Consortium, 2001. — version 2. Access mode: http://www.natcorp.ox.ac.uk/ [electronic resource].

[11] Davies, Mark. The iWEB Corpus: The 14 Billion Word Web Corpus. — Provo, UT : Brigham Young University, 2018. Access mode: https://www.english-corpora.org/iweb/ [electronic resource].

Software

[12] Jamovi. The Jamovi project. Open statistical platform. — Sydney, Australia, 2020. — version 1.2.17.0. Access mode: https://www.jamovi.org/ [software package].

Russian News Clustering and Headline Selection Shared Task

Ilya Gusev
Moscow Institute of Physics and Technology
Moscow, Russia
ilya.gusev@phystech.edu

Ivan Smurov
ABBYY,
Moscow, Russia
ivan.smurov@abbyy.com

Abstract

This paper presents the results of the Russian News Clustering and Headline Selection shared task. As a part of it, we propose the tasks of Russian news event detection, headline selection, and headline generation. These tasks are accompanied by datasets and baselines. The presented datasets for event detection and headline selection are the first public Russian datasets for their tasks. The headline generation dataset is based on clustering and provides multiple reference headlines for every cluster, unlike the previous datasets. Finally, the approaches proposed by the shared task participants are reported and analyzed.

Keywords: clustering, event detection, headline selection, headline generation, news, embeddings, Russian, dataset, NLP evaluation

DOI: 10.28995/2075-7182-2021-20-289-301

Дорожка по кластеризации и выбору заголовков для русских новостей

Гусев И. О.
МФТИ
Москва, Россия
ilya.gusev@phystech.edu

Смуров И. М.
ABBYY
Москва, Россия
ivan.smurov@abbyy.com

Аннотация

В статье представлены результаты соревнования по кластеризации и выбору заголовков для новостей на русском. В соревновании предлагаются задачи по обнаружению новостных событий, отбору и написанию заголовков для новостей. Вместе с задачами предоставляются наборы данных и базовые решения. Представленные наборы данных для обнаружения новостных событий и выбора заголовков — первые общедоступные наборы данных на русском языке для своих задач. Набор данных для написания заголовков использует кластеризацию и содержит набор эталонных заголовков для каждого кластера, в отличие от предшественников. Представлены и проанализированы подходы, предложенные участниками соревнования.

Ключевые слова: кластеризация, определение событий, выбор заголовков, написание заголовков, новости, эмбеддинги, русский язык, набор данных

1 Introduction

Automatic news feeds and aggregators are a common way to read, search and analyze news. Event clustering is one of the core features for many news aggregators, including Google News and Yandex News. The news event clustering task is to collect news from different news agencies about the same event. It is essential to present all events from different perspectives to create a comprehensive picture. Moreover, selecting the most suitable headline and other entities is possible only within such clusters. The task can also be helpful for news monitoring systems. Correct clustering should help to accurately calculate any statistics about events, companies, or people's mentions in the news.

As for natural language research, this task is a good benchmark for different text clusterization or classification models. For example, the models should identify various types of paraphrasing and recognize named entities to distinguish clusters correctly.

After event clustering, it is necessary to choose the most relevant headline for every cluster. Thus, it is the headline selection task. One can define relevance in many ways. We provide a list of criteria by which a title can be considered suitable. The main points of this list are informativeness and the absence of clickbait.

Instead of choosing one of the existing headlines, one can generate a completely novel headline based on news texts. The main reason to do this is to evade the situations where all presented headlines are inadequate.

To solve these three problems, we composed a shared task as a part of the Dialogue 2021 conference[1]. These three problems are considered as independent tasks and evaluated separately. We provided datasets for each of them and hosted Kaggle-like competitions on the CodaLab platform to determine the best models.

The primary source of data for all tasks is the Telegram Data Clustering contest[2]. Telegram conducted it in 2020. The task was to build a news aggregator over a provided document collection. The subtasks were language detection, category detection, and event clustering. Only HTML documents without any additional annotations were given.

The contributions of our paper are as follows: we present datasets for the Russian news event clustering, headline selection, and headline generation tasks along with baselines. The first two datasets are the first of their kind for the Russian language. The last one is the first dataset for headline generation for Russian with multiple reference headlines. Furthermore, we present and analyze some of the works of the shared task participants.

2 Related work

Many works on event clustering exist. This task is also known as "documental event detection" or just "event detection"[37, 19]. We will use these names as synonyms in this paper.

The Topic Detection and Tracking (TDT) initiative [37] was the first significant effort for this task. The event detection task was a part of TDT. There were only 25 events in the dataset. The CMU approach used TF-IDF embeddings, hierarchical agglomerative clustering with average linking, and a time window with incremental IDF for online detection.

Azzopardi et al.[3] used TF-IDF document representations and incremental k-means clustering with cosine similarity between these representations. There were no human annotations involved, and the comparison was with Google News automatic clustering.

Miranda et al.[28] focused on cross-lingual clustering. They used TF-IDF document representation with separate vectors for different document sections and incremental centroid clustering. There were also two versions of these representations: monolingual and cross-lingual. They introduced a multilingual dataset adapted from Rupnik et al.[30] containing articles in English, Spanish and German. It was manually annotated with monolingual and cross-lingual event labels. Linger et al.[25] utilized the same multilingual dataset, but used the multilingual DistilBERT[13] and the Sentence-BERT[32] triplet network structure for multilingual document representation.

There are some works on news clustering that focus on thematic news clustering instead of event clustering (with categories like "sports" or "technologies") [34].

There are few papers on Russian news event clustering. Dobrov et al.[14] described event clustering on a ROMIP-2006 news collection. Three different document representations were utilized: TF-IDF for texts, TF-IDF for titles, and a conceptual index. Several clustering methods were used, including hierarchical clustering, DBSCAN[12], and centroid clusterings. They utilized custom software for creating a manual reference clustering.

Voropaev et al.[41] used the same document collection we do, Telegram Data Clustering contest document collection, but without a large-scale manual markup. They took the solution of one of the Telegram contest participants as a reference clustering. They utilized TF-IDF, BERT[6, 22], LASER[1] for document representation, and hierarchical clustering or DBSCAN[12].

[1] http://www.dialog-21.ru/evaluation/

[2] https://contest.com/docs/data_clustering2

As for headline selection, we did not find any papers with a similar task definition. However, there are several works on clickbait detection in headlines[11, 35] which is a part of our task.

The headline generation is a well-covered task in previous papers. Statistical models for news headline generation in Banko et al.[5] are among the first approaches to this task. Takase et al.[29] were the first to use encoder-decoder neural models for headline generation. In 2019, there was a Dialogue Evaluation shared task for Russian news headline generation[27]. The main drawback of the 2019 shared task was the fact that it was about single-document headline generation. It means that there was only one actual headline for every article, so automatic evaluation methods were not effective in this setup.

3 Clustering

3.1 Data and metrics

For the first two tasks, we took news documents from the Telegram Data Clustering contest covering three days of May 2020. After using all baseline clustering algorithms, we sampled a set of document pairs from their output and included a small number of random pairs. Next, we annotated every pair with Yandex Toloka, a Russian crowdsourcing platform. The task was to determine whether two documents describe the same event. The annotators were provided with a comprehensive guide and a simple heuristic: "if two titles of two documents are interchangeable, these documents are from the same cluster". Five people annotated each pair. Annotators were required to pass training, exam, and their work was continuously evaluated through the control pairs ("honeypots"). We included only pairs with an agreement of 4/5 or 5/5 in the final markup.

	May 25	May 27	May 29
#Documents	19380	20080	19096
#Pairs	14838	8493	8476
#Positives	7301	3918	3896
Fraction of positives	49.2%	46.1%	46.0%

Table 1: Clustering dataset statistics

The final statistics for every day are shown in Table 1. There are 661 unique hosts in the training document collection and at least 500 unique news agencies. The top 5 news agencies by the number of documents are shown in Table 2. The distribution of news agencies by the number of documents is in Figure 1.

Host	#Documents
tass.ru	645
lenta.ru	301
www.mk.ru	223
www.rosbalt.ru	199
ura.news	188

Table 2: Top 5 hosts by the number of documents, May 25

According to annotation guidelines, a pair of documents refer to the same cluster when they have the same: time of the event; numbers, such as the stock price of a company or the number of victims; locations. A pair of documents are not from the same cluster when they contain: inconsistent facts, such as the time or place of the event or significantly distinguished number of victims; description of an event in one of the documents, and a commentary on this event in another document. These criteria are similar to the definition of an event in TDT[37].

We do not publish documents themselves, only two URLs and an annotation result, to evade legal issues. However, Telegram shares archives with all documents, so it is easy to join documents and

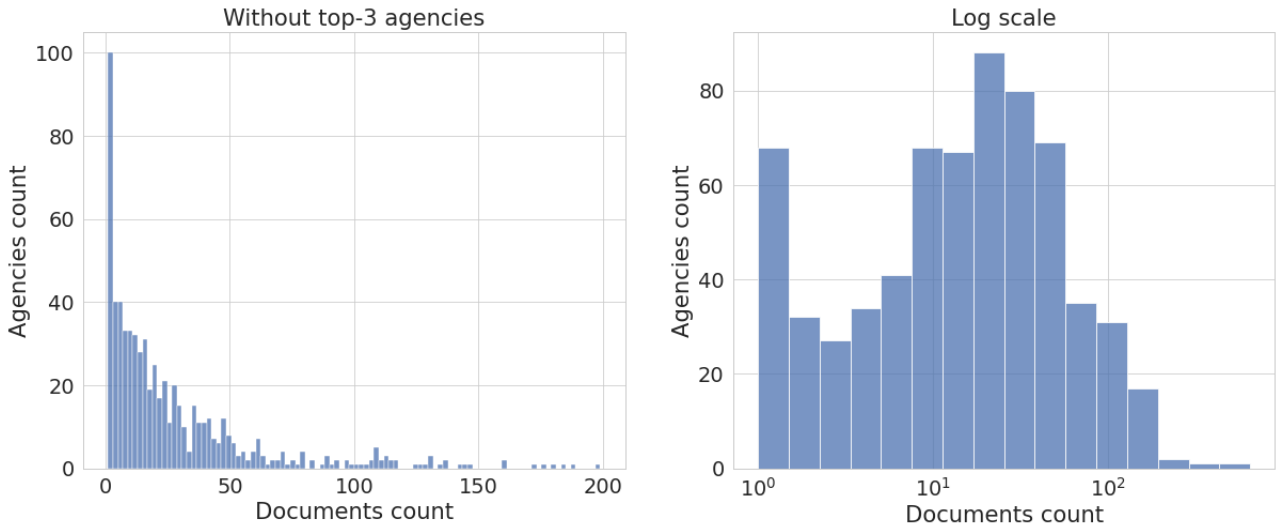


Figure 1: News agencies distribution by the number of documents, May 25

annotations. An example of the joining process can be found in the baseline script[3]. No additional scraping is needed.

In the shared task rules, we forbade the use of any news documents from the test set for training or pretraining, including pretraining word vectors on these documents. In the real world, texts are not available for the model in advance, so the tested models must not use the test-time texts in any way. However, it was permissible to use texts with earlier dates from the Telegram contest for pretraining.

It is possible to collect non-pairwise annotations for clustering with crowdsourcing methods, but it is hard to organize proper quality control of the annotation process. Non-pairwise annotations we know of for event detection were done by experts, not via crowdsourcing.

As the primary metric for this task, we chose F1-score (corresponds to Dice similarity coefficient for clustering). The proper clustering metrics like Adjusted Rand Index[31] are not directly applicable to our dataset as we have no reference clustering for all documents. Several dozen possible pair-counting based clustering metrics[18] are available, and it is an active research question what metric is better[17]. As for our choice, we decided that the metric should be interpretable from the classification perspective, and our classes are imbalanced, so we chose F1.

3.2 Baselines

As a baseline clustering algorithm, we utilized hierarchical agglomerative clustering with average linking.

We used different pretrained embeddings with cosine similarity to compute the distance matrix. The simplest ones are FastText[15] text embeddings, TF-IDF with truncated SVD for dimensionality reduction (latent semantic analysis, LSA[21]), and Universal Sentence Encoder (USE[38]). FastText text embeddings are computed as a concatenation of average, maximum, and minimum of FastText word embeddings. The examples of clusters based on USE embeddings are shown in Figure 2.

Text2Title embeddings are FastText embeddings with an additional linear matrix trained to determine whether a headline and a text are from the same document. HNSW[26] index on the original FastText embeddings was used to mine hard negatives for the triplet loss[20]. The model architecture is depicted on Figure 3. The main advantage of this model is its speed, as it consists of only one linear layer on top of the FastText embedding.

[3]https://github.com/dialogue-evaluation/Russian-News-Clustering-and-Headline-Generation/blob/main/baselines.ipynb

Figure 2: Dendrograms for two clusters based on USE embeddings, cosine similarity, and average linking. These clusters contain several errors.

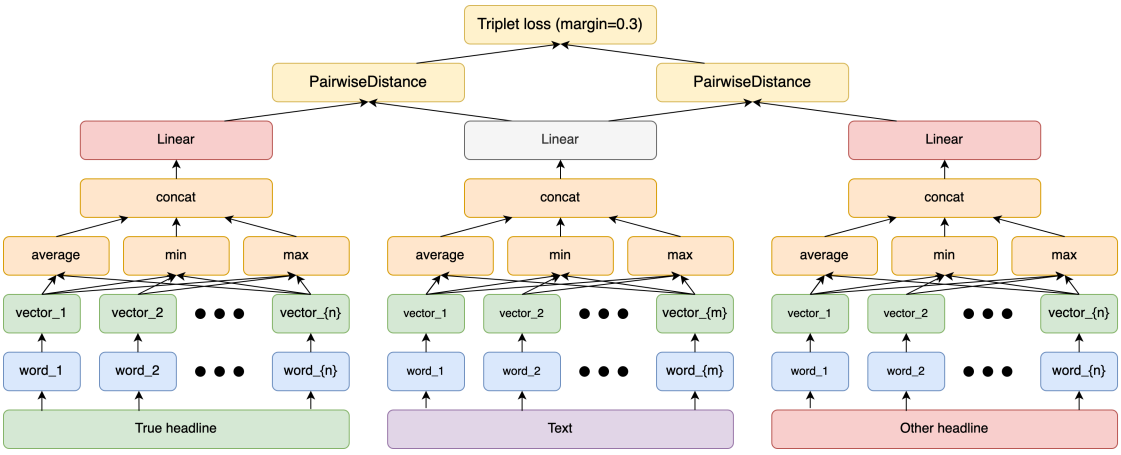


Figure 3: Text2Title model architecture

TTGenBottleneck is a BERT seq2seq model based on RuBERT[22] with practically disabled cross-attention between encoder and decoder. It was trained to predict a headline for a document. "Bottleneck" is the encoder embedding of the first token: the only embedding decoder can attend. We design it[4] to contain all information from the text needed to generate a headline.

We emphasize that none of the baseline models were fine-tuned to the news event detection task in any way.

We present the results of all baselines in Table 3. Text2Title has better scores than plain FastText, as additional pretraining helps to create more representative embeddings. TTGenBottleneck is the best baseline model.

Model	Validation	Public LB	Private LB
TTGenBottleneck	93.4	93.9	93.7
USE	89.3	89.4	87.8
Text2Title	86.5	86.4	84.8
LSA	83.1	82.7	80.5
FastText	80.9	81.9	80.1

Table 3: F1-scores for positive pairs in % for baseline models

3.3 Results

The task can be solved both as a classification task or as a clustering task. The baselines were for the clustering only, but many participants preferred to use classification models.

The final results are in Table 4. There are several participants without scores in the private leaderboard. We made a technical mistake that disabled automatic score evaluation of the private part, and the participants had to resubmit their answers. Not all of them managed to do that.

Rank	Codalab login	Public LB	Private LB
1	**maelstorm** [5]	**96.9**	**96.0**
2	naergvae [4]	96.7	96.0
3	g2tmn [16] [6]	96.5	95.7
4	Kouki [33]	95.5	95.5
5	alexey.artsukevich	95.8	95.3
6	smekur [33]	94.6	93.9
7	nikyudin	93.8	93.0
8	landges	91.6	90.6
9	kapant	90.7	89.9
10	bond005	90.2	89.2
11	anonym	90.6	89.1
12	mashkka_t [40]	85.3	71.5
13	vatolinalex [33]	95.2	47.6
-	blanchefort	94.1	
-	imroggen	90.3	
-	Abiks	89.4	
-	dinabpr [40]	84.4	

Table 4: Final results of the clustering track, F1-scores for positive pairs in %

[4] https://huggingface.co/IlyaGusev/gen_title_tg_bottleneck_encoder
[5] Leonid Pugachev and Alim Adelshin, DeepPavlov
[6] https://github.com/oldaandozerskaya/DE2021_news_similarity

The best solution was a classification model, an ensemble of 4 bert-base-multilingual models trained with stochastic weight averaging (SWA)[2]. The second-placed model was also classification-based and consisted of a single RuBERT[22] in a standard pair classification setting. Finally, the third-placed model was a hard-voting classification ensemble of RuBERT models.

The best clustering-based model (4th place overall) was based on SBERT[32]: a siamese BERT model over RuBERT. The authors[33] improved the model by using Global Multihead Pooling[10] and contrastive loss. It should be noted that this model scored only 0.005 F1 less than the best classification model on the private test. Additionally, this is the only model that did not experience score deduction on the private test set compared to the public set.

Unsurprisingly all top-placed models utilized pretrained masked language models. In particular, the top three classification-based models and the best clustering-based model (4th place overall) all used BERT[6]. The best solution not utilizing language model embeddings was able to score 0.930 F1 (7th place overall) and had the following architecture: CatBoost[9] over FastText[15] and USE[38] as well as some handcrafted features (hyperlinks and named entities intersections).

Another system of note is GPT-3-based[23] zero-shot model. The idea is as follows: the two headlines' perplexity is computed and compared to thresholds. While the final score is rather unimpressive, 0.7 F1, the completely unsupervised nature of training makes this result worth mentioning.

4 Headline selection

4.1 Data and metrics

The documents for this task are from the same collection as the documents for the clustering task. The task was, for each given pair of headlines, to predict which headline is better. There are four possible options: "left", "right", "draw", and "bad". The "bad" option is for the case if two headlines are from different news events. Annotation conditions are the same as for the clustering task. According to annotation guidelines, one headline is better than the other if some of the following conditions are met: it contains more information than the other; it does not hide any details; it does not contain undefined entities; it has no grammatical errors; it is not emotional; it is not too wordy. Some of these criteria are subjective, so we use only examples with high agreement to reduce subjectivity.

The final statistics for every day are shown in Table 5.

	May 25	May 27	May 29
#Pairs	5091	3147	3103
#Left won	2185	1254	1216
#Right won	2167	1269	1207
#Draw	362	184	161
#Bad	377	440	518

Table 5: Headline selection dataset statistics

The primary metric for this task is weighted accuracy. The dataset contains a lot of "bad" examples. They were added to the dataset to prevent data leaks for the clustering task. We ignore them in the denominator and numerator. The systems can predict any label for them. For the other three labels, scores in the numerator are calculated as stated in Table 6. We chose such weights because incorrect predictions for the "draw" pairs should not be penalized as hard as "left/right" errors. We do not provide F1-scores, as the dataset is balanced, and the only information these scores add to accuracy is how well "draw" pairs are detected.

4.2 Baseline

We used a ranking gradient boosting model (CatBoost[9]) with USE embeddings as features and PairLogit loss as a baseline for this task. This CatBoost mode is designed to take pairs as input, so it fits the

	Left	Right	Draw
Left	1.0	0.0	0.5
Right	0.0	1.0	0.5
Draw	0.5	0.5	1.0

Table 6: Predictions weights in the accuracy metric

task perfectly. We present the result accuracy of this model in Table 7.

$$PairLogitLoss(pairs, res) = -\sum_{p,n \in pairs} log(\frac{1}{1 + e^{-(res_p - res_n)}})$$

Model	Validation	Public LB	Private LB
USE + Catboost	81.0 $\pm$ 1.5	81.2 $\pm$ 0.5	81.1 $\pm$ 0.4

Table 7: Weighted accuracy for the baseline model, 5 runs, in %

4.3 Results

Most of the successful submissions are based on the same schema as the baseline. The results are presented in Table 8. There is only one submission that surpasses the baseline significantly. It utilizes an ensemble of ranking models trained on different embeddings, including USE, SBERT[32], RuBERT[22], XLM-R[39], and mT5[44]. The most effective single model was mT5.

Rank	Codalab login	Public LB	Private LB
1	**sopilnyak** [42]	**86.0**	**85.4**
2	landges	81.3	82.0
3	nikyudin	83.2	81.6
4	LOLKEK	80.8	81.4
5	maelstorm	81.8	79.8
6	a.korolev	65.8	66.2

Table 8: Final results of the headline selection track, weighted accuracy in %

5 Headline generation

5.1 Data and metrics

In this track, the task was to generate a headline for a news cluster. It should be similar to any of the headlines of cluster documents. Other formulations to this task are possible. For example, there is a much more complicated task formulation where one should generate a headline similar to the best headline of the cluster. However, we chose the weak formulation as it does not require any annotation. We were not able to use the documents and annotations of the first two tracks because of possible data leaks.

To make a dataset entirely unknown for participants, we scraped news documents from a test version of the Telegram news aggregator[7] from March 9 to March 12, 2021. We used the TTGenBottleneck clustering method with a distance threshold skewed into precision to the detriment of recall, as it forms very restricted clusters.

We made available for participants only clustered document texts without any headlines or additional meta information. There are 6726 documents and 1035 clusters in the dataset.

[7] https://1398.topnews.com/ru/

This formulation can be considered more robust than the single document (and single news agency [27]) headline generation as it does not force the model to generate a headline in a particular style to get good automatic metrics.

We use traditional automatic metrics for headline generation, ROUGE[24], and BLEU[8]. We calculate metrics between a predicted headline and every actual headline of a cluster and use the maximum score as a prediction score.

5.2 Baselines

There are two baselines for this track. The first is Random Lead-1, where we choose the first sentence of a random document from the cluster as a baseline. The second baseline is a generative sequence to sequence[36] RuBERT[22] trained on news text-title pairs from the Telegram contest[8]. We generate a headline for every text in the cluster and choose randomly from them. We present the metrics for the baselines in Table 9.

Model	ROUGE	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Seq2seq RuBERT	44.9 ± 0.3	74.5 ± 0.5	52.5 ± 0.4	32.8 ± 0.4	49.5 ± 0.5
Random Lead-1	30.2 ± 0.5	47.8 ± 0.4	36.8 ± 0.6	20.4 ± 0.4	33.6 ± 0.6

Table 9: Results of headline generation baseline models, 5 runs, in %

5.3 Results

There were only two participants in this track, mainly because of hard time restrictions. Their scores can be seen in Table 10. The baseline remained unbeaten.

The solution by `Rybolos` was based on fine-tuning the ruGPT-3 Large model. We suppose their metrics are low compared to baselines because of possible technical error or only a tiny part of the dataset used for training.

Codalab login	ROUGE	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
LOLKEK	38.7	69.5	46.3	26.4	43.3
Rybolos	29.2	59.6	36.5	17.6	33.5

Table 10: Final results of the headline generation track, in %

6 Conclusion

6.1 Reproducibility

All materials of the shared task are available at the official repository[9], including all data, annotation guidelines, baselines, and links to the CodaLab competitions. The competitions themselves will be permanently open in order to make comparison easier for new researchers.

6.2 Organization notes

The timeline was as follows:

- February 8: Clustering task started on Codalab.
- February 26: Headline selection task started on Codalab.
- March 13: Headline generation task started on Codalab.
- March 22: Final deadline for all competitions.
- March 28: Final deadline for paper submission.

[8] https://huggingface.co/IlyaGusev/rubert_telegram_headlines

[9] https://github.com/dialogue-evaluation/Russian-News-Clustering-and-Headline-Generation

The main reason for a short time window for submissions for the last task was that it was unclear where to get previously unseen data from roughly the same domain. We should have solved this question even before the shared task was announced. Unfortunately, the news clustering task was the only one that we fully prepared before the start.

The other major problem was with the private leaderboard for the clustering task, as the initial private part of the test dataset was a copy of a public part by our technical mistake, so it was impossible to measure metrics without resubmitting answers.

6.3 General conclusions

In the event detection task, most of the successful models were classification-based BERT models. However, it turns out clustering embeddings can be almost as effective when trained with correct pooling and loss function. Moreover, they generalize better, they are easier to deploy, and they are more computationally effective. It is arguably the most important takeaway of the shared task.

Unsurprisingly, big multilingual models and ensembles showed the best metrics in the headline selection task. Nevertheless, the task participants did not diverge much from the baseline solution, so we hope more sophisticated schemas and models will be developed in the future.

There was only one week to develop a valid generative model in the headline generation task. Only two participants managed to present a complete model in these challenging time restrictions, so we consider this track results inconclusive. However, the dataset itself can be helpful in future research.

6.4 Future research

There are several possible directions for future research:

1. Finding models with good accuracy/speed trade-offs is a very promising research path. Almost all of the models used by participants were extremely parameter-heavy and slow. Distillation of these models into lightweight ones is needed in order to enable their use in production systems.
2. Different clustering methods should be inspected. Solutions for the clustering task were agglomerative-centered — almost no one used BIRCH clustering[43], and only a few people used DBSCAN[12] and its modifications.
3. The Telegram Data Clustering news documents are multilingual, so datasets and models should be multilingual too. We created only a Russian dataset because we used a Russian crowdsourcing platform and had a constrained money budget. There are no principal reasons why these datasets should not be multilingual.
4. The existing clustering dataset is focused on 24-hour time windows. Clustering on more extensive periods can be significantly harder. Moreover, online event detection (as it originally was in TDT) is also possible and requires windowed or incremental clustering.
5. Synthetic checklists[7] can be used to evaluate the quality of event detection and headline selection. It is easy to create pairs of documents differing only in numbers, entities, or time of an event.
6. As for the headline generation task, the clustering enables proper conditioning. It is possible to train a model to write headlines in the style of the particular news agency without a thematic bias, as agencies will be equally presented in clusters.
7. It is possible to revisit other TDT tasks in light of recent advances in building document representations with big pretrained models like BERT.

Acknowledgements

We would like to thank the participants of all three tracks, especially Tatiana Shavrina, Ivan Bondarenko, and Nikita Yudin for helpful comments and valuable suggestions.

References

[1] Artetxe Mikel, Schwenk Holger. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond // Trans. Assoc. Comput. Linguistics. — 2019. — Vol. 7. — P. 597–610. — Access mode: https://transacl.org/ojs/index.php/tacl/article/view/1742.

[2] Averaging weights leads to wider optima and better generalization / Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov et al. // arXiv preprint arXiv:1803.05407. — 2018.

[3] Azzopardi Joel, Staff Christopher. Incremental Clustering of News Reports // Algorithms. — 2012. — Vol. 5, no. 3. — P. 364–378. — Access mode: https://doi.org/10.3390/a5030364.

[4] BERT for Russian news clustering / Khaustov, Gorlova, Kalmykov, Kabaev // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[5] Banko Michele, Mittal Vibhu O, Witbrock Michael J. Headline generation based on statistical translation // Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. — 2000. — P. 318–325.

[6] Bert: Pre-training of deep bidirectional transformers for language understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.

[7] Beyond Accuracy: Behavioral Testing of NLP Models with CheckList / Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 4902–4912. — Access mode: https://www.aclweb.org/anthology/2020.acl-main.442.

[8] Bleu: a Method for Automatic Evaluation of Machine Translation / Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. — Philadelphia, Pennsylvania, USA : Association for Computational Linguistics, 2002. — Jul. — P. 311–318. — Access mode: https://www.aclweb.org/anthology/P02-1040.

[9] CatBoost: unbiased boosting with categorical features / Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev et al. // arXiv preprint arXiv:1706.09516. — 2017.

[10] Chen Qian, Ling Zhen-Hua, Zhu Xiaodan. Enhancing Sentence Embedding with Generalized Pooling // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA : Association for Computational Linguistics, 2018. — Aug. — P. 1815–1826. — Access mode: https://www.aclweb.org/anthology/C18-1154.

[11] Clickbait detection / Martin Potthast, Sebastian Köpsel, Benno Stein, Matthias Hagen // European Conference on Information Retrieval / Springer. — 2016. — P. 810–817.

[12] A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). — AAAI Press, 1996. — P. 226–231.

[13] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter / Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf // arXiv preprint arXiv:1910.01108. — 2019.

[14] Dobrov Boris, Pavlov Andrey. Basic line for news clusterization methods evaluation // Proceedings of the 5-th Russian Conference RCDL-2010. — 2010.

[15] Enriching word vectors with subword information / Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — P. 135–146.

[16] Glazkova Anna. Towards News Aggregation in Russian: a BERT-based Approach to News Article Similarity Detection // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[17] Gösgens Martijn, Tikhonov Alexey, Prokhorenkova Liudmila. Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures // arXiv preprint arXiv:1911.04773. — 2019.

[18] Ground Truth Bias in External Cluster Validity Indices / Yang Lei, James C. Bezdek, Simone Romano et al. // Pattern Recogn. — 2017. — May. — Vol. 65, no. C. — P. 58–70. — Access mode:

https://doi.org/10.1016/j.patcog.2016.12.003.

[19] A History and Theory of Textual Event Detection and Recognition / Yanping Chen, Zehua Ding, Qinghua Zheng et al. // IEEE Access. — 2020. — Vol. 8. — P. 201371–201392.

[20] Hoffer Elad, Ailon Nir. Deep metric learning using Triplet network. // ICLR (Workshop) / Ed. by Yoshua Bengio, Yann LeCun. — 2015. — Access mode: http://dblp.uni-trier.de/db/conf/iclr/iclr2015w.html#HofferA14.

[21] Indexing by latent semantic analysis. / S. Deerwester, S.T. Dumais, G.W. Furnas et al. // Journal of the American Society for Information Science 41. — 1990. — P. 391–407.

[22] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // CoRR. — 2019. — Vol. abs/1905.07213. — 1905.07213.

[23] Language Models are Few-Shot Learners / Tom Brown, Benjamin Mann, Nick Ryder et al. // Advances in Neural Information Processing Systems. — Vol. 33. — Curran Associates, Inc., 2020. — P. 1877–1901. — Access mode: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[24] Lin Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. — Barcelona, Spain : Association for Computational Linguistics, 2004. — Jul. — P. 74–81. — Access mode: https://www.aclweb.org/anthology/W04-1013.

[25] Linger Mathis, Hajaiej Mhamed. Batch Clustering for Multilingual News Streaming // Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts co-located with 42nd European Conference on Information Retrieval, Text2Story@ECIR 2020, Lisbon, Portugal, April 14th, 2020 [online only]. — Vol. 2593 of CEUR Workshop Proceedings. — CEUR-WS.org, 2020. — P. 55–61. — Access mode: http://ceur-ws.org/Vol-2593/paper7.pdf.

[26] Malkov Yu A., Yashunin D. A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2020. — Vol. 42, no. 4. — P. 824–836.

[27] Malykh V.A. Kalaidin P.S. Headline Generation Shared Task on Dialogue'2019 // Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference "Dialogue". — 2019.

[28] Multilingual Clustering of Streaming News / Sebastião Miranda, Artūrs Znotiņš, Shay B. Cohen, Guntis Barzdins // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — Brussels, Belgium : Association for Computational Linguistics, 2018. — Oct.-Nov. — P. 4535–4544. — Access mode: https://www.aclweb.org/anthology/D18-1483.

[29] Neural headline generation on abstract meaning representation / Sho Takase, Jun Suzuki, Naoaki Okazaki et al. // Proceedings of the 2016 conference on empirical methods in natural language processing. — 2016. — P. 1054–1059.

[30] News across languages-cross-lingual document similarity and event tracking / Jan Rupnik, Andrej Muhic, Gregor Leban et al. // Journal of Artificial Intelligence Research. — 2016. — Vol. 55. — P. 283–316.

[31] Rand William M. Objective criteria for the evaluation of clustering methods // Journal of the American Statistical association. — 1971. — Vol. 66, no. 336. — P. 846–850.

[32] Reimers Nils, Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. — Association for Computational Linguistics, 2019. — P. 3980–3990. — Access mode: https://doi.org/10.18653/v1/D19-1410.

[33] Smirnova, Vatolin, Shkarin. Russian News Similarity Detection with SBERT: pre-training and fine-tuning // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[34] Stankevicius Lukas, Lukosevicius Mantas. Testing Pre-trained Transformer Models for Lithuanian News Clustering // Proceedings of the Information Society and University Studies 2020, Kaunas, Lithuania, April 23, 2020. — Vol. 2698 of CEUR Workshop Proceedings. — CEUR-WS.org, 2020. — P. 46–53. — Access mode: http://ceur-ws.org/Vol-2698/p08.pdf.

[35] Stop clickbait: Detecting and preventing clickbaits in online news media / Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, Niloy Ganguly // 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM) / IEEE. — 2016. — P. 9–16.

[36] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to Sequence Learning with Neural Networks // Advances in Neural Information Processing Systems. — Vol. 27. — Curran Associates, Inc., 2014. — Access mode: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

[37] Topic Detection and Tracking Pilot Study: Final Report / J. Allan, J. Carbonell, G. Doddington et al. // Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. — Lansdowne, VA, USA, 1998. — Feb. — P. 194–218. — 007.

[38] Universal sentence encoder / Daniel Cer, Yinfei Yang, Sheng-yi Kong et al. // arXiv preprint arXiv:1803.11175. — 2018.

[39] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // CoRR. — 2019. — Vol. abs/1911.02116. — 1911.02116.

[40] Using Generative Pretrained Transformer-3 Models for Russian News Clustering and Title Generation tasks / Tikhonova, Pisarevskaya, Shliazhko, Shavrina // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”. — Vol. XX. — 2021. — P. xx–xx.

[41] Voropaev Pavel, Sopilnyak Olga. Comparison of news clustring methods. — 2020.

[42] Voropaev Pavel, Sopilnyak Olga. Transformer-based Embeddings for Russian News Clustering and Headline Selection // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”. — Vol. XX. — 2021. — P. xx–xx.

[43] Zhang Tian, Ramakrishnan Raghu, Livny Miron. BIRCH: an efficient data clustering method for very large databases // ACM sigmod record. — 1996. — Vol. 25, no. 2. — P. 103–114.

[44] mT5: A massively multilingual pre-trained text-to-text transformer / Linting Xue, Noah Constant, Adam Roberts et al. // CoRR. — 2020. — Vol. abs/2010.11934. — 2010.11934.

Unreasonable Effectiveness of Rule-Based Heuristics in Solving Russian SuperGLUE Tasks

Iazykova Tatyana
HSE University
Moscow, Russia
tvyazykova@edu.hse.ru

Bystrova Olga
HSE University / Sberbank
Moscow, Russia
ovbystrova@edu.hse.ru

Kapelyushnik Denis
HSE University
Moscow, Russia
dmkapelyushnik@edu.hse.ru

Kutuzov Andrey
University of Oslo
Oslo, Norway
andreku@ifi.uio.no

Abstract

Leaderboards like SuperGLUE are seen as important incentives for active development of NLP, since they provide standard benchmarks for fair comparison of modern language models. They have driven the world's best engineering teams as well as their resources to collaborate and solve a set of tasks for general language understanding. Their performance scores are often claimed to be close to or even higher than the human performance. These results encouraged more thorough analysis of whether the benchmark datasets featured any statistical cues that machine learning based language models can exploit. For English datasets, it was shown that they often contain annotation artifacts. This allows solving certain tasks with very simple rules and achieving competitive rankings.

In this paper, a similar analysis was done for the Russian SuperGLUE (RSG), a recently published benchmark set and leaderboard for Russian natural language understanding. We show that its test datasets are vulnerable to shallow heuristics. Often approaches based on simple rules outperform or come close to the results of the notorious pre-trained language models like GPT-3 or BERT. It is likely (as the simplest explanation) that a significant part of the SOTA models performance in the RSG leaderboard is due to exploiting these shallow heuristics and that has nothing in common with real language understanding. We provide a set of recommendations on how to improve these datasets, making the RSG leaderboard even more representative of the real progress in Russian NLU.

Keywords: leaderboards, benchmark, heuristics, rule-based, language models, natural language understanding
DOI: 10.28995/2075-7182-2021-20-302-317

Эффективность правиловых эвристик в решении тестовых сетов Russian SuperGLUE

Языкова Татьяна
НИУ ВШЭ
Москва, Россия
tvyazykova@edu.hse.ru

Быстрова Ольга
НИУ ВШЭ / Сбербанк
Москва, Россия
ovbystrova@edu.hse.ru

Капелюшник Денис
НИУ ВШЭ
Москва, Россия
dmkapelyushnik@edu.hse.ru

Кутузов Андрей
Университет Осло
Осло, Норвегия
andreku@ifi.uio.no

Аннотация

SuperGLUE и подобные ему наборы тестовых заданий для оценки решения задач по обработке естественного языка способствуют активному развитию этой области знания, так как они предлагают методологию тестирования современных моделей и подходов. Им удается привлекать к разработке автоматических решений задач на понимание языка лучшие мировые инженерно-технические группы с их ресурсами. Количественные результаты, полученные на некоторых тестовых сетах для английского языка, сравнимы с человеческими или превышают их. Всё это привело к более пристальному изучению наборов тестовых заданий на предмет наличия в них

каких-либо неучтённых статистических особенностей, на которые могут опираться языковые модели. Так, исследования, проведенные на подобных наборах тестовых заданий для английского языка, выявили наличие артефактов разметки в представленном языковом материале. Использование таких артефактов в построении простых правил позволяет решать некоторые из тестовых заданий на уровне, сравнимом с современными языковыми моделями.

В данной статье приводится похожий анализ для Russian SuperGLUE, системы оценки и рейтинга моделей для решения задач на понимание естественного языка для русского. Уязвимость данного тестового набора задач для простых эвристик была также подтверждена, а наши решения на основе правил часто превышали результаты таких небезызвестных языковых моделей, как GPT-3 и BERT. Мы предполагаем (как наиболее простое объяснение), что языковые модели, находящиеся в верхней части рейтинга Russian SuperGLUE, также в значительной степени полагаются на простые эвристики, а значит, не представляется возможным говорить о реальном понимании языка. Основываясь на результатах нашего исследования, мы составили набор рекомендаций по улучшению Russian SuperGLUE, которые позволят обеспечить более справедливую оценку русскоязычных систем, нацеленных на моделирование 'понимания' естественного языка.

Ключевые слова: рейтинг, тестовые сеты, эвристики, правила, языковые модели, понимание естественного языка

1 Introduction

These days many researchers are coming to a dreadful realisation that we are not that much advanced in natural language understanding (NLU) as we used to think. Huge Transformer-based models are crowning the SuperGLUE leaderboard [34], yet one should not trust these shining examples so fast. It has been shown in [23] that actually these models are exploiting statistical patterns related to the lack of diversity in data or class imbalances to demonstrate amazing performance without looking deeper and truly emulating natural language understanding. The danger of having such statistical cues is not in their mere presence. The core of the problem is that they are inherent to particular datasets only and therefore are hardly applicable to the language itself. It means that the systems do not really 'understand' natural language: instead, they are utilising statistical cues that are typical to these specific datasets, so the whole process comes down to simple pattern matching. Modern language models are trained on the amount of data no native speaker will hardly ever see [20]. But are they really as superior as we believe them to be, and is it even necessary to do genuine NLU to solve the test sets at this level of performance?

The issue became even more relevant now that the SuperGLUE benchmark, that was initially created for English, was adopted for Russian in the form of Russian SuperGLUE (RSG) benchmark and the corresponding leaderboard [33]. In this paper, we study the possibility to achieve results comparable to ones in the leaderboard without using any machine learning algorithms. We manually examined the datasets in order to find statistical regularities. As a result, we came up with a list of simple rule-based heuristics (for instance, label instances as 'entailment' if they contain the word 'был' '*was*'). We do not have direct proofs that machine learning based models also make use of these shallow heuristics in the case of RSG. But we do know that this was confirmed to be true for the English SuperGLUE [32], and we know that deep neural nets are extremely efficient in capturing regularities useful for their objective function. Following the Occam's Razor, we argue that finding and exploiting shallow statistical cues (not necessarily the ones we found manually) is much more plausible explanation for the observed performance of pre-trained language models than the assumption that they 'understand' Russian discourse.

Moreover, we evaluated a set of trivial baselines, such as random choice, majority class and random balanced choice. The goal was to compare state-of-the-art (SOTA) results against those and to see whether cutting-edge deep learning architectures (GPT-3, BERT, etc) significantly outperform them. As we found out, this is not always the case.

1.1 Contributions

The contributions of this work can be formulated as follows:

1. We introduced a set of simple rule-based heuristics applicable to various datasets of Russian SuperGLUE benchmark[1], and evaluated their performance on the test data.
2. We evaluated the performance of even more trivial baselines (random choice, majority class, etc) on the Russian SuperGLUE tasks, to establish a lower boundary for language models' performance.

[1] https://github.com/tatiana-iazykova/2020_HACK_RUSSIANSUPERGLUE

3. A number of suggestions, spotted annotation errors and generally problematic or controversial cases are given for the authors of the Russian SuperGlUE benchmark, for further improvement.

2 Previous work

Leaderboards provide the NLP community with tools to evaluate language models. This competition ensures a fair ground for comparison as the models are required to solve the same tasks on a single independently curated set of data. For example, the GLUE leaderboard [11] was initially designed for English and consists of several diverse natural language understanding tasks and a diagnostic dataset with openly available labels to evaluate models.

By March 2021, the situation with the GLUE dataset is the following: 14 different models hold a higher ranking than the human performance which is equal to 87.1 [25]). The knowledge about language is considered as key to solve the GLUE or any NLU tasks, yet when the SOTA approach [7] (as of now) exceeded human performance by 3.8, the creators of this model hypothesised that it was not necessary for those specific datasets. With 14 other models outperforming the human level as well, it has soon become clear that the benchmark itself is no longer able to provide a challenging evaluation system. As a result, the authors of GLUE designed SuperGLUE [34] for a more representative analysis of the current progress in NLU. To track this progress for other languages, other researchers created language-specific benchmarks similar to GLUE and SuperGLUE, e.g. Russian SuperGLUE [33] explored in this paper or CLUE [4] for the Chinese language.

Although the SuperGLUE benchmark is more recent, its current SOTA score of 90.3 [6] also managed to surpass the human performance [34] by 0.5. As these competitions attract the world's best engineering teams with almost unlimited resources, models like T5 [9], GPT-3 [17], BERT [3] and its optimized versions like RoBERTa [29] usually hold top rankings and yet their performance scores differ by a mere fraction. These models prove their reputation by achieving scores that are very close to or even higher than human benchmark, and this is where some room for criticism appears.

Such complex models require considerable resources, raising questions about their general utility [8]. Indeed, for the majority of us the size and efficiency of a model is as important as the performance scores, and some trade-off has to be allowed. Through such discussions, e. g. [31], the NLP community attempts to increase the transparency of benchmarks. Fortunately, the leaderboards are open for changes and new functionality. For example, the MOROCCO project has been recently launched to evaluate Russian SuperGLUE models in two additional dimensions: inference speed and GPU RAM usage[2].

Although these issues are important, another question — probably a deeper one — is raised by how exactly large-scale language models are 'solving' certain NLU tasks. For example, BERT has skyrocketed the performance in many NLP tasks for English, yet if we take a closer look into its 'language skills', we might be disappointed [32]. It appears that BERT never misses an opportunity to use shallow heuristics while solving tasks on natural language inference [23, 13, 14], reading comprehension [12, 36, 2, 18], argument reasoning comprehension [26] and text classification [14].

The above-mentioned analysis is mostly English-centred, and we are truly grateful to the creators of the Russian SuperGLUE [33], since it is now possible to have a fair ground for comparing Russian NLU models. It is the first standardized set of diverse NLU benchmarks for Russian.

Some of the instances for its datasets were translated from the corresponding tasks in the SuperGLUE, while the others were collected by the RSG authors from scratch [10].

In this paper, we explore all the datasets thoroughly to test their vulnerability to shallow heuristics. The results are compared to other approaches represented in the Russian SuperGLUE leaderboard. It should be noted that the RSG has been created very recently, and the human performance of 0.811 is still at the top of the leaderboard. As of early May 2021, the highest score of 0.679 was achieved by an ensemble of Transformer models.

[2] https://russiansuperglue.com/performance/

3 Methodology

The Russian SuperGLUE benchmark consists of 9 datasets or tasks, that follow the GLUE and SuperGLUE methodology. Each task is designed to evaluate if a model or an approach can solve problems with the help of logic, common sense and reasoning. Data is split into training, validation and test samples. The true labels of the test set are not openly available and to evaluate a system on the test set, it is necessary to submit the predictions to the leaderboard. Currently there are two versions of Russian SuperGLUE present, namely 1.0 and 1.1; our research was based on the latest 1.1 version.

Our general approach was to identify shallow heuristics and design rule-based functions that would surpass the results achieved by the trivial baselines (majority class, random choice and random balanced choice) and potentially approach SOTA scores. Being native Russian speakers, we invested our efforts into manual exploration of each dataset. Additionally, ELI5[3], a tool to debug machine learning classifiers and explain their predictions, was applied to some of the tasks. It was used to check if any tokens are more specific to one of the classes in the dataset. Moreover, whenever the lemmatisation was needed, `pymorphy2` morphological analyzer [15] was used.

As the datasets differ significantly, there was no intention to identify a single heuristic to solve them all: we analyzed them separately. Heuristics found in the training sets were applied to the validation sets to get an idea of their performance. All of the heuristics that were proved to work on training and validation sets were combined into functions with a set of if-else statements. To determine the order of these statements, we tested different sequences empirically and chose the ones with the higher performance scores.

Finally, these rule-based functions were applied to the relevant test sets. To handle examples that did not trigger any of the heuristics, three aforementioned baseline methods were used to predict the label. All the predictions were grouped by their baseline function and submitted to the leaderboard to receive scores for each dataset individually as well as the total score per submission. The results are shown in the Table 8 in the section 4. Below we first describe task-specific heuristics in more detail.

3.1 Linguistic Diagnostic for Russian (LiDiRus)

Inspired by [35], the authors of the original SuperGLUE benchmark included a small curated test dataset called AX-b for the analysis of the models' overall performance. It was 'provided not as a benchmark, but as a tool for error analysis, qualitative model comparison, and development of adversarial examples' [11]. LiDiRus is a Russian version of this dataset, where each sentence was translated from English into Russian with the help of 'professional translators and linguists to ensure that the desired linguistic phenomena remain' [33].

We identified a set of heuristics for this dataset. They are grouped in Table 1, which also demonstrates how many samples in the validation set were covered by each heuristic and the percentage of their correct predictions. Only basic split on white-space is applied for pre-processing sentences for all heuristics but one. 'All lemmas in sentences 1 and 2 overlap' required lemmatisation first.

As the dataset does not assume any training and validation samples, the corresponding parts of the Textual Entailment Recognition for Russian (TERRa) dataset from the same RSG benchmark were used to make predictions to calculate the class distribution for the majority class and random weighted baseline functions if the utterances did not trigger the use of any heuristics. TERRa's class distribution differs from LiDiRus[4] but maintains the same dataset organisation.

The performance of the aforementioned heuristics (as well as heuristics for other RSG tasks) is consolidated into Table 8 which can be found in section 4. It provides SOTA scores for each task as of May 2021, performance scores of the baseline functions, as well as the results for heuristics-based approach supported by one of the three baseline functions. The evaluation metric used for LiDiRus is Matthews correlation coefficient [22]. The authors of the original benchmark for English suggested this metric, as

[3] `https://github.com/eli5-org/eli5`

[4] The labels are distributed in the following proportions: 58.4% not_entailment, 41.6% entailment for LiDiRus vs. 49.15% not_entailment, 50.85% entailment for TERRa

	Heuristic	Target label	Coverage	Correct
1	Number of tokens in sentence 1 differs from sentence 2 by more than 10	not_entailment	24.3%	65.2%
2	Sentences 1 and 2 differ by two commas	not_entailment	27.3%	64.1%
3	Sentences 1 and 2 differ by two words	not_entailment	16%	66.6%
4	The presence of 'и', 'не', 'что', 'никогда', 'вовсе', 'это' ('and', 'not', 'that', 'never', 'at all', 'this') in only one of the two sentences	not_entailment	29.3%	66.3%
5	Vocabularies of two sentences overlap by 100% (lemmatised data)	entailment	4%	64.4%
6	'Чтобы', 'будет', 'от', 'он' ('in order to', 'will', 'from', 'he') occur in both sentences	entailment	11.6%	57%

Table 1: LiDiRus: identified heuristics with their coverage of the validation set and the percentage of correct predictions

it can be applied to unbalanced binary classification problems and its values range from -1 to 1, with 0 being the performance of uninformed guessing [11].

As it is a diagnostic dataset, the SOTA approach is hardly applicable to it, though the fact that there is a small performance gap between heuristics and other models deserves to be mentioned. It supports the hypothesis that shallow heuristics might play a significant part in the results of the approaches which apply pre-trained language models to solve NLP tasks.

3.2 Russian Commitment Bank (RCB)

Russian Commitment Bank is a Natural Language Inference task dataset that consists of naturally occurring discourses where the task is to predict the relation of one phrase (hypothesis) to the given text (premise), where the options are entailment, contradiction and neutral.

The training data is distributed unequally in this dataset (46.3% — neutral, 35.4% — entailment, 18.3% — contradiction for train data; 52.7% — neutral, 33.6% — entailment, 13.6% — contradiction for validation data). This imbalance can potentially lead to a substantial bias towards a certain class for the large pre-trained language models. The model can simply predict the majority class and still achieve a rather good result, though it by any means would not be natural language understanding.

The number of instances in the training data is 438, in the validation — 220 and in the test — 438. Two metrics are used to evaluate the model's performance on solving this task: Accuracy and F1, as is the case with the corresponding Commitment Bank task in the original SuperGLUE [11]. According to the authors of the SuperGLUE, the imbalanced nature of the dataset (relatively fewer neutral examples in the English version and significantly more neutral instances in the Russian SuperGLUE) was the reason for them using two metrics, where they used macro-F1 for multi-class problems.

One of the heuristics (you can see its performance in Table 2) used for solving this dataset utilised the correlation between the label and the number of words in the hypothesis or the premise.

Instances with 5-7 words in the hypothesis would more likely have the 'neutral' label (median for it is 5) and instances with less than 5 words in the hypothesis would more likely have the 'contradiction' label (median for it is 4).

Instances with more than 30 words in them would likely belong to the 'entailment' category (median for the 'entailment' is 27).

As we can see from the Table 2, the heuristics do not cover all the data, leaving some answers to be predicted with the help of three baselines (majority class, random choice, random balanced choice). The results achieved with the help of the heuristics were comparable with results of large pre-trained language models in the RSG leaderboard, which are given in the section 4 of this paper.

	Heuristic	Target label	Coverage	Correct
1	The hypothesis is a sub-string of the premise	entailment	26%	40%
2	75% intersection of the hypothesis and premise's vocabularies (lemmatised data)	entailment	5%	45%
3	The presence of 'признать' ('admit') (lemmatised data)	entailment	6%	36%
4	The presence of 'подозревать, считать, говорить, думать, надеяться, понять, уверять' ('suspect, consider, say, think, hope, assure, realise') (lemmatised data)	neutral	6%	36%
5	Hypothesis > 5 words	contradiction	41%	23%
6	$4 <$ hypothesis < 8 words	neutral	34%	70%
7	More than 30 words in the premise	entailment	35%	39%

Table 2: RCB: identified heuristics with their coverage of the validation set and the percentage of correct predictions

3.3 Choice of Plausible Alternatives for Russian language (PARus)

To evaluate progress in open-domain common sense casual reasoning, the authors of Russian SuperGLUE provided the Choice of Plausible Alternatives for Russian language (PARus) dataset. It is based on the English COPA [30]. A typical task in PARus consists of a premise and two alternatives, where the goal is to select the alternative that has a causal or effect relation with the premise.

There are 400 samples in the train dataset and 100 in the validation set. Since there is no semantics behind the labels, the difference between label distribution in the training and validation data should be considered irrelevant. Also because of this lack of label meaning, it was challenging to find linguistic heuristics to solve this task. All textual data was lemmatised to get better results. The heuristics used for tackling this task are shown in Table 3.

The heuristics check whether one of the choices has more shared lemmas with the premise than the others, and if so, then this choice should be taken as an answer. If the vocabulary overlap was the same for all choices, one of the baseline functions was applied. If one of choices had more words than the other, then this choice was taken as an answer.

	Heuristic	Coverage	Correct
1	If one of choices has more shared lemmas with the premise than the others, it is taken as an answer (lemmatised data)	22%	64%
2	If one of choices has more words than the others, then this choice should be taken as an answer (lemmatised data)	59%	52%
3	The combination of these two heuristics (lemmatised data)	66%	52%

Table 3: PARus: identified heuristics with their coverage of the validation set and the percentage of correct predictions

As we can see from Table 3, these heuristics cover less than 70% of the data, therefore many answers still depend on one of three baselines. Overall results are presented in the Table 8 in section 4. The maximal accuracy score was 0.516. To achieve SOTA performance, we probably need more complex algorithms. It proves that this task fulfills its goal and we do need to learn some open-domain common sense casual reasoning to solve such tasks.

3.4 Russian Multi-Sentence Reading Comprehension (MuSeRC)

The MuSeRC dataset is collected for the reading comprehension task. It contains more than 900 paragraphs across 5 different domains: elementary school texts, news, fiction stories, fairy tales, and summaries of TV series and books [10]. Samples were collected based on the following criteria:

1. the passage length is less than 1.5K characters;
2. the passage contains named entities;
3. if the passage contains only one named entity, then it must have one or more co-reference relations.

Furthermore, the authors of the dataset ensured correct sentence splitting and used these sentences in a crowd-sourcing effort at the Yandex.Toloka platform. In it, humans were asked to generate questions, a set of answers for each of them and to check that answering a question requires consulting with more than one sentence in the text. The answer can be either True or False, so all the answers are either correct or incorrect with no in-between. The number of correct answers varies and each question/answer pair is treated individually[5].

A set of heuristics identified for this dataset is grouped in Table 4.

	Heuristic	Target label	Coverage	Correct
1	All lemmas from the answer occur in the text (lemmatised data)	True	39.2%	58.8%
2	The answer is longer than 11 tokens	True	10.3%	72.3%
3	More than 6 overlapping lemmas between the answer and the text (lemmatised data)	True	18.9%	73.9%
4	No overlapping lemmas between the answer and the text (lemmatised data)	False	9.9%	89.1%
5	The answer is shorter than 4 tokens	False	46.4%	64.9%
6	One overlapping lemma between the answer and the text (lemmatised data)	False	18.6%	69.3%

Table 4: MuSeRC: identified heuristics with their coverage of the validation set and the percentage of correct predictions

While predicting, the if-else statements dealt with the 'True' label first, as it is less frequent in the data. The function yields the intended label as long as at least one of the heuristics gets triggered. After that, the opposite set of heuristics is applied.

The overall performance is given in Table 8 which can be found in section 4. To provide the evaluation metrics, the dataset authors roughly followed the evaluation procedure by [28, 21]. Since each answer-option can be assessed independently, F1-averaged (F1a) is applied to evaluate binary decisions over all the answer options in the dataset. It is a harmonic mean of precision and recall per question. Exact Match (EM) is the exact match per each instance, i.e. each set of predictions should be the same as of the answers [10].

We were not able to reach neither the SOTA score nor the human performance, although the obtained results are on par with some of those produced by large pre-trained language models. In fact, at the time of submission, our heuristics-based approach combined with the majority class baseline function achieved higher performance scores for this task than Multilingual Bert and RuGPT3Small[6].

3.5 Textual Entailment Recognition for Russian (TERRa)

Textual Entailment Recognition is another dataset dedicated to the Natural Language Inference task. This task requires to recognise, given two text fragments, whether the meaning of one text is entailed (can be inferred) from the other text [33]. This task is similar to the RCB, yet in TERRa there are only

[5]The average number of questions is approximately 20. The labels are distributed in the following proportions: 55% false and 45% true for the training set vs. 55.6% false and 44.4% true for the validation set.

[6]https://huggingface.co/sberbank-ai/rugpt3small_based_on_gpt2

two categories (entailment/not_entailment) instead of three. The number of instances in the training data is 2 616, in the validation — 307 and in the test — 3 198.

Similar to the RCB, one of the heuristics (Table 5, heuristics 6 and 7) used in solving this dataset utilised the interplay between the label and the number of words. Unlike with the RCB, in this dataset it was possible to find such relations only between the label and the number of words in the premise. Instances with less than 29 words would more likely have the label 'not_entailment' (median number of words for not_entailment is 29) whereas if the number of words was more than 32, then the label is likely'entailment' (median number of words for entailment is 32).

Another heuristic (8 in Table 5) for TERRa dealt with the presence of specific words, namely 'только' ('only') and 'мужчина' ('man') in the hypothesis. It was possible to find a rather noticeable correlation between their presence and the label.

	Heuristic	**Target label**	**Coverage**	**Correct**
1	The hypothesis is a sub-string of the premise	entailment	1%	50%
2	Vocabularies of the hypothesis and the premise overlap by 33% (lemmatised data)	not_entailment	11%	69%
3	Vocabularies of the hypothesis and the premise overlap by 75% (lemmatised data)	entailment	9%	52%
4	Vocabularies of the hypothesis and the premise overlap by 66% (lemmatised data)	entailment	9%	56%
5	Vocabularies of the hypothesis and the premise overlap by 100% (lemmatised data)	entailment	14%	65%
6	Less that 29 words in the premise	not_entailment	45%	58%
7	More than 32 words in the premise	entailment	45%	60%
8	The presence of 'только', 'мужчина' ('only', 'man') (lemmatised data)	not_entailment	21%	66%

Table 5: TERRa: identified heuristics with their coverage of the validation set and the percentage of correct predictions

As we can see from the Table 5, the heuristics do not cover all the data, leaving some answers to be predicted with the help of the three trivial baselines. However, the results achieved with the help of the heuristics were comparable with the results of large pre-trained language models in the RSG leaderboard and even outperformed the ones by RuGPT3Medium and RuGPT3Small.

3.6 Russian Words in Context (RUSSE)

Depending on its context, a word can have multiple, potentially unrelated, senses. For example, the Russian word 'лук' ('onion'/'bow') can mean either vegetable or weapon depending on its surrounding words. The 'word in context' task can be described as a binary classification problem, and the goal is to predict whether a given word has the same meaning in both given sentences. The Russian SuperGLUE task borrows original data from the Russe Word Sense Induction and Disambiguation shared task [27].

To find out whether the decision can be made based on simple rules, we checked whether the target word appears in the same form in both sentences. In addition, we calculated the proportion of shared tokens to all tokens in both sentences and the difference in their lengths.

The heuristics cover about 50% of the data and make correct predictions in about 65% cases. According to Table 8 in Section 4, we managed to achieve 0.595 accuracy score.

3.7 The Winograd Schema Challenge for Russian (RWSD)

The original purpose of the Winograd Schema Challenge (WSC) was to serve as an alternative Turing test to evaluate an automatic system's capacity for common sense inference [19]. The challenge evaluates the models' ability to identify the antecedent of the pronoun, which might be critical, for example, for translation purposes [5]. The performance scores on the WSC for English quickly progressed from a

	Heuristic	Target label	Coverage	Correct
1	Target word in the same form	True	14%	58%
2	Tokens overlap by more than 10%	True	4%	50%
3	Number of tokens in sentence 1 differs from sentence 2 by more than 6	False	49 %	65 %

Table 6: RUSSE: identified heuristics with their coverage of the validation set and the percentage of correct predictions

simple guess to near-human level [1] after neural language models trained on massive corpora were applied to solve this challenge.

The RWSD dataset is a Russian translation of the pronoun disambiguation task used in the SuperGLUE benchmark [24]. RWSD maintains the same structure providing a pair or a batch of sentences that differ by one or two words:

Example 1: 'Кубок не помещается в коричневый чемодан, потому что он слишком большой.' ('The trophy doesn't fit into the brown suitcase because it is too large.')
Example 2: 'Кубок не помещается в коричневый чемодан, потому что он слишком маленький.' ('The trophy doesn't fit into the brown suitcase because it is too small.')

There is an ambiguity in these sentences, namely 'он' ('it') might refer to either 'кубок' ('the trophy') or 'чемодан' ('suitcase'). Each sentence is followed by an antecedent and a pronoun for disambiguation, which can be successfully resolved if a model assigns a 'false' label to the first example for the pair of 'чемодан'('suitcase') and 'он' ('it'), and if 'true' is assigned to the second example for the same pair.

The model cannot rely on the word order or the structure of a sentence, as the task is organised so that they cannot be used for the disambiguation process [24]. Each sentence might be either 'true' or 'false' depending on a suggested pair of antecedents and pronouns. For example, one has to pay attention to a special word, i.e. 'большой' ('large') or 'маленький' ('small') in the aforementioned sentences.

There is a clearly unequal distribution of classes in the RWSD dataset. The labels for the training and validation sets are distributed as follows: 51% 'false' and 49% 'true' labels for the former and 55.4% 'false' and 44.6% 'true' labels for the latter. However, the 'false' values appear 67% of the times in the *test set*, which is very different from the datasets provided for training and validation.

We were not able to identify any heuristic that would surpass the performance score of predictions made by the majority class baseline (see Table 8 for reference), but this misfortune carries one of our most important findings.

Apparently, the very same approach to choose the most common value was used by many sophisticated models listed in the Russian SuperGLUE leaderboard by the time of our submission. The SOTA score, which is the score achieved by Multilingual T5, several BERT variations (trained on multilingual data and on Russian corpora only), RuGPT3Medium and RuGPT3Small, is 0.669: the same as achieved by our majority class baseline function. While solving the task, these models allegedly opted to predict using the majority class rather than try to actually solve the Winograd Schema Challenge. Such models as Golden Transformer, RuGPT3XL few-shot and RuGPT3Large apparently made an attempt to really predict something, but their results are sub-optimal:0.545, 0.649 and 0.636 respectively, which is in fact below the 0.662 tf-idf baseline provided by the RSG creators. The problem is similar to that with Winograd Schema Challenge (WSC) in the original SuperGLUE [11].

3.8 Yes/no Question Answering Dataset for Russian (DaNetQA)

DaNetQA is a question answering dataset for yes/no questions. Each example is a triplet of (passage, question, answer), with the title of the page as optional context [33]. The answers are encoded in a True/False formal similar to the corresponding SuperGLUE 'BoolQ' dataset. As with the Russian Commitment Bank task, here we can also notice the unequal distribution of labels (Train: True — 60.7%,

	Heuristic	Target label	Coverage	Correct
1	The question starts with ‘был’ (‘was/were’)	True	45%	58%
2	The question starts with ‘есть’ (‘is/are’)	True	13%	81%
3	The question starts with ‘входит ли’ (‘does it belong to’)	False	37%	100%
4	The question starts with ‘едят ли’ (‘do they eat’)	False	2%	53%
5	The question starts with ‘правда ли’ (‘is it true’)	False	18%	89%
6	More than 5 words in the question	False	46%	71%
7	More than 90 words in the passage	False	48%	53%

Table 7: DaNetQA: identified heuristics with their coverage of the validation set and the percentage of correct predictions

False — 39.3%) and the mismatch of this relation among training and validation data (Vaidation: True — 50.2%, False — 49.8%). The number of instances in the training data is 1 749, in the validation — 821 and 805 in the test set.

Like with the RCB and TERRa, one of the heuristics used in solving this dataset utilised the relations between the label and the number of words in questions (Table 7, heuristic 6) or passage (heuristic 7).

Instances with more than 5 words would more likely have the label ‘False’ (median number of words for False is 6 in the training data).

Heuristic 3 exploits correlation between the beginning of the question and the label: if the question starts with ‘входит ли’ (‘does it belong to’), the label in the validation dataset is False 100% of the time.

As we can see from the Table 7, the heuristics do not cover all the data, leaving some answers to be predicted with the help of the three trivial baselines. However, the results achieved with the help of the heuristics were comparable with the results of large pre-trained language models in the RSG leaderboard.

3.9 Russian Reading Comprehension with Commonsense Reasoning (RuCOS)

Russian reading comprehension with Commonsense reasoning (RuCoS) is a large-scale reading comprehension dataset which requires common sense reasoning. Unlike MuSeRC, the main data domain for RuCoS is news articles and there is more data for this task. Also in this task, a system is given a list of named entities from which it should choose the right answer (while in the MuSeRC, the answers do not have to be named entities at all). RuCoS consists of queries automatically generated from news articles; the answer to each query is a text span from a summarizing passage of the corresponding article.

This task is based on the English ReCoRD benchmark [28]. All text examples were collected from Russian media. The texts were then filtered by the IPM frequencies of the contained words and, finally, manually reviewed.

The heuristics applied to this task dealt with the presence of name entities in the question. The algorithm simply predicted all the entities present in the question. A modification of this approach was to sort the remaining entities based on the frequency of their appearance in the text. We tried several threshold values for this rule and finalized our choice on the following rule: all entities whose stems appeared less than three times were removed from our predictions.

Both heuristics were applied every time we made predictions. Thus, their coverage is 100%. We managed to outperform the tf-idf baseline with both F1 score end EM metric around 0.26, but the SOTA results are on par with human performance score, which is 0.93/ for F1 and 0.89 for EM.

	Metrics	Hum.	SOTA	maj.	rand.	r.(b)	H maj.	H rand.	H r.(b)
LiDiRus	M. Corr	0.626	**0.231**	0.000	0.024	0.000	0.147	0.149	**0.182**
RCB	Avg. F1	0.680	**0.452**	0.217	0.332	0.319	0.400	**0.401**	**0.401**
	Acc.	0.702	**0.546**	**0.484**	0.347	0.374	0.438	0.436	0.438
PARus	Acc.	0.982	**0.908**	0.498	0.474	0.480	0.478	**0.508**	0.470
MuSeRC	F1a	0.806	**0.941**	0.000	0.477	0.450	**0.671**	0.669	0.669
	EM	0.420	**0.819**	0.000	0.078	0.071	0.237	0.195	**0.202**
TERRa	Acc.	0.920	**0.871**	0.513	0.503	0.483	**0.549**	0.547	0.548
RUSSE	Acc.	0.805	**0.729**	0.587	0.501	0.528	**0.595**	0.497	0.543
RWSD	Acc.	0.840	**0.669**	**0.669**	0.487	0.597	**0.669**	0.565	0.604
DaNetQA	Acc.	0.915	**0.917**	0.503	0.494	0.520	**0.642**	0.629	0.629
RuCoS	F1	0.930	**0.920**	0.250	0.250	0.250	**0.260**	0.260	0.260
	EM	0.890	**0.924**	0.247	0.247	0.247	**0.257**	0.257	0.257
Total		0.811	**0.679**	0.374	0.372	0.385	**0.468**	0.445	0.454

Table 8: Performance scores at the time of submission. 'Maj.' is the majority class baseline function, 'rand.' — random choice, 'r.(b)' — random balanced choice, H — the heuristics-based approach.

4 Discussion

Table 8 shows the best results after applying the heuristics described above to the Russian SuperGLUE test sets. The heuristic based approach (H) was combined with one of the trivial baseline functions. The majority value and weights for baseline functions were obtained from combined training and validation sets. For every task, we chose heuristic (or combination of several heuristics) that led towards the best score. Even if for some tasks we are still far away from beating SOTA performance, simple baselines and heuristics can achieve relatively good results. For RWSD task one can achieve SOTA performance just by using the majority class baseline.

Our heuristics approach works well for the RCB task with the difference between H maj. model and SOTA being about 5%. For TERRa, RUSSE and DaNetQA we are far from SOTA results but, still, our results are on the same level with RuBERT [16] and GPT models from the leaderboard.

Yet for several tasks, the heuristics approach did not work as well. RuCoS, MuSeRC and PARus proved that it is not enough to use dataset-specific statistical cues to solve them, so for these three tasks it seems that the large pre-trained language models really pick up some peculiarities of Russian language.

Since our approaches can be divided into two groups (trivial baselines and rule-based heuristics), we will look closer at them separately.

4.1 Trivial baselines

As it was mentioned before, first we chose three baseline methods to solve all tasks in Russian SuperGLUE: majority class, random choice and random weighted choice. When comparing these baselines to the other methods, we should keep in mind that they do not rely on any linguistic knowledge at all.

All three baselines show very interesting results. For the majority class baseline, the best result is the one for the RWSD task. It should be emphasized again that not only one can achieve the SOTA performance with the majority baseline here, but, at the moment of submission, half of the leaderboard models probably use this approach as their solver method, since they all have the same performance score. Simple random choice worked good on the RCB and RWSD as well. Random balanced choice outperforms the majority class approach on the DaNetQA, RWSD, TERRa, PARus, and RUSSE.

Across all RSG tasks, the balanced random choice baseline achieves the average score of 0.385. Language models obviously outperform this score, but the difference is marginal: only half of the systems in the leaderboard achieve a score higher than 0.5, and the best ensemble of transformers reaches 0.679

(the human performance is 0.811). For some benchmarks (for example, RuCOS), the random balanced baseline *outperforms* BERT and GPT-3 models. In one specific case of the RWSD benchmark, no model managed to outperform the *simple majority class baseline*.

From this, we conclude that the RSG leaderboard scores should be taken with a grain of salt and compared to the trivial baselines. For example, the 0.669 accuracy of the SOTA models on the RWSD dataset is not a sign of their 'human-like comprehension abilities': it is just that these models (or their authors) could not do any better than simply predict the same label for all the instances in the test set. For other tasks, the picture is only slightly better: in most cases, the leaderboard participants managed to improve the random balanced baseline only by a small margin. Another important finding is that the class balances in the RSG test sets are similar to those in the validation and training sets: this allows one to achieve boosted accuracies by simply replicating these distributions in the test answers. This is true for all the tasks evaluated by accuracy (six of the RSG tasks). If the class labels in these six datasets were perfectly balanced, the expected average accuracy of the random baseline would be 0.472. In the real RSG, this value is 0.538. This is certainly an undesired property for a test set in general; in this case it additionally makes it difficult to assess to what extent the large-scale language models' NLU performance for Russian is actually an artifact of this data leakage.

4.2 Rule-based heuristics

Rule-based heuristics tend to improve trivial baselines in cases of TERRa, RUSSE, RCB (considering Avg. F1 score). Here we categorize the described rules. Note that most of them are language-agnostic and can be tested on benchmarks for other languages as well.

1. Using text length (e. g. 'More than 30 words in the premise'): these rules are useful for RCB, PARus, MuSeRC, TERRa, RUSSE, DaNetQA.
2. Using binary lexical features (e. g. 'Presence of 'чтобы', 'будет', 'от', 'он'): these rules are useful for LiDiRus, RCB, TERRa, DaNetQA.
3. Using word forms or lemmas overlap (e. g. 'Sentences 1 and 2 use the same set of lemmas'): these rules are useful for LiDiRus, RCB, PARus, MuSeRC, TERRa, RUSSE.
4. Other task-specific heuristics.

The existence of such statistical cues is not a problem in itself: after all, this is what machine learning is after. What we see as problematic is the fact that the large over-parameterized models seem to mostly rely on them (judging by their performance scores which are not radically higher, and sometimes even lower than the scores of our rule-based approach). This means they do not employ valid inference strategies, and do not demonstrate anything close to much-praised 'natural language comprehension'. We again emphasize that our heuristics are extremely simplistic and often boil down to counting the number of words in the sentence or to finding the lexical overlap between the question and the answer (sometimes after lemmatisation). There is no doubt that billion-parameter language models can find much more statistical cues in the training data than the authors of this paper were able to come up with. But these regularities will only work on the test instances drawn from the same general population (annotated or generated according to the same guidelines). This is *pattern matching*, not *language understanding*.

5 Conclusion

The recently introduced Russian SuperGLUE (RSG) set of natural language understanding benchmarks [33] has already attracted well-deserved attention from the Russian NLP practitioners. The RSG leaderboard is filled with the impressive performance scores produced by sophisticated language models trained with bleeding-edge deep learning architectures (BERT, GPT-3, etc) on titanic corpora of Russian. But are these scores really so impressive? In this paper, we studied what performance can be achieved for the RSG benchmarks *without training any language models*.

First, we established the performance boundaries of the trivial baselines: random choice, majority class choice and balanced random class choice (probabilities weighted by the distribution of class labels in the training data). We found that in some cases, these baselines outperform large-scale language models. Second, we moved on to find out whether the RSG datasets contain other statistical regularities.

In has been shown in prior work for English and other languages that deep learning models are very prone to collecting low-hanging fruits and tracing shallow semantic and structural phenomena which help minimizing the loss on a particular dataset, instead of actually learning real linguistic generalizations.

To this end, we manually compiled a set of very simple custom rule-based heuristics for each RSG dataset (for example, '**set the label '*contradiction*' if the word** не **'not' is present in the hypothesis**', etc).

It turned out that up to 50% and more of instances (depending on a particular dataset) might be covered by this or that heuristic. Moreover, applying these rules to actually solve the RSG (with fallback to the majority class baseline if no rule is applicable) constitutes a system which achieves a very competitive RSG average score of 0.468. This outperforms RuGPT3-Small, on par with RuGPT3-Medium, and is very close to the BERT performance.

We conclude that most RSG datasets abound in statistical regularities which can easily be found at training time and employed at test time, without expensive and complicated language model pre-training. The reasons are arguably the same as with the English test sets[7]: compilation of benchmarks by crowdsourcing and the natural desire of crowd-workers to fulfill the job in the easiest way possible.

To sum up, we recommend the RSG maintainers to 1) modify the test sets to minimize the data leakage from label distributions; 2) diversify the datasets so as to eliminate at least the most striking statistical cues (it shouldn't be possible to find the correct answer by simply counting words); 3) provide official majority class and random weighted baselines. We believe this will make the Russian SuperGLUE leaderboard even more informative of the real state of the art in Russian natural language processing.

In the future, it will be useful to develop a Russian equivalent of the HANS benchmark [23]: a test set containing adversarial examples, or even simply examples drawn from sources substantially different from those in the RSG. It will allow to evaluate the generalization abilities of large pre-trained language models for Russian. It would also be interesting to study the correlations between the predictions of our heuristics and the predictions of the language models in the RSG leader-board, in order to find out whether they actually exploit similar rules.

Finally, in the course of working on this paper, we collected a large trove of annotation errors and generally problematic or controversial cases in the RSG datasets. We have shared these findings with the RSG maintainers, in the hope of its future improvement.

References

[1] An Analysis of Dataset Overlap on Winograd-Style Tasks / Ali Emami, Kaheer Suleman, Adam Trischler, Jackie Chi Kit Cheung // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 5855–5865. — Access mode: https://www.aclweb.org/anthology/2020.coling-main.515.

[2] Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets / Saku Sugawara, Pontus Stenetorp, Kentaro Inui, Akiko Aizawa // Proceedings of the AAAI Conference on Artificial Intelligence. — 2020. — Apr. — Vol. 34, no. 05. — P. 8918–8927. — Access mode: https://ojs.aaai.org/index.php/AAAI/article/view/6422.

[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // 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). — Minneapolis, Minnesota : Association for Computational Linguistics, 2019. — Jun. — P. 4171–4186. — Access mode: https://www.aclweb.org/anthology/N19-1423.

[4] CLUE: A Chinese Language Understanding Evaluation Benchmark / Liang Xu, Hai Hu, Xuanwei Zhang et al. // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona,

[7] In fact, many RSG test sets are translated from English.

Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 4762–4772. — Access mode: https://www.aclweb.org/anthology/2020.coling-main.419.

[5] Davis Ernest. Winograd Schemas and Machine Translation // CoRR. — 2016. — Vol. abs/1608.01884. — 1608.01884.

[6] He Pengcheng, Liu Xiaodong, Gao Jianfeng, Chen Weizhu. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. — 2021. — 2006.03654.

[7] ERNIE: Enhanced Language Representation with Informative Entities / Zhengyan Zhang, Xu Han, Zhiyuan Liu et al. // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 1441–1451. — Access mode: https://www.aclweb.org/anthology/P19-1139.

[8] Ethayarajh Kawin, Jurafsky Dan. Utility is in the Eye of the User: A Critique of NLP Leaderboards // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online : Association for Computational Linguistics, 2020. — Nov. — P. 4846–4853. — Access mode: https://www.aclweb.org/anthology/2020.emnlp-main.393.

[9] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer / Colin Raffel, Noam Shazeer, Adam Roberts et al. // Journal of Machine Learning Research. — 2020. — Vol. 21, no. 140. — P. 1–67. — Access mode: http://jmlr.org/papers/v21/20-074.html.

[10] Fenogenova Alena, Mikhailov Vladislav, Shevelev Denis. Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 6481–6497. — Access mode: https://www.aclweb.org/anthology/2020.coling-main.570.

[11] GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding / Alex Wang, Amanpreet Singh, Julian Michael et al. // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. — Brussels, Belgium : Association for Computational Linguistics, 2018. — Nov. — P. 353–355. — Access mode: https://www.aclweb.org/anthology/W18-5446.

[12] Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks / Anna Rogers, O. Kovaleva, Matthew Downey, Anna Rumshisky // AAAI. — 2020.

[13] HellaSwag: Can a Machine Really Finish Your Sentence? / Rowan Zellers, Ari Holtzman, Yonatan Bisk et al. // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4791–4800. — Access mode: https://www.aclweb.org/anthology/P19-1472.

[14] Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment / Di Jin, Zhijing Jin, Joey Tianyi Zhou, Peter Szolovits // Proceedings of the AAAI Conference on Artificial Intelligence. — 2020. — Apr. — Vol. 34, no. 05. — P. 8018–8025. — Access mode: https://ojs.aaai.org/index.php/AAAI/article/view/6311.

[15] Korobov Mikhail. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts / Ed. by Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko et al. — Springer International Publishing, 2015. — Vol. 542 of Communications in Computer and Information Science. — P. 320–332. — online; accessed: http://dx.doi.org/10.1007/978-3-319-26123-2_31.

[16] Kuratov Yuri, Arkhipov Mikhail. Adaptation of deep bidirectional multilingual transformers for Russian language // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". — 2019. — Access mode: http://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf.

[17] Language Models are Few-Shot Learners / Tom Brown, Benjamin Mann, Nick Ryder et al. // Advances in Neural Information Processing Systems / Ed. by H. Larochelle,

M. Ranzato, R. Hadsell et al. — Vol. 33. — Curran Associates, Inc., 2020. — P. 1877–1901. — Access mode: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[18] Learning and Evaluating General Linguistic Intelligence / Dani Yogatama, Cyprien de Masson d'Autume, J. Connor et al. // ArXiv. — 2019. — Vol. abs/1901.11373.

[19] Levesque Hector J., Davis Ernest, Morgenstern Leora. The Winograd Schema Challenge // Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning. — KR'12. — Rome, Italy : AAAI Press, 2012. — P. 552–561.

[20] Linzen Tal. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 5210–5217. — Access mode: https://www.aclweb.org/anthology/2020.acl-main.465.

[21] Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences / Daniel Khashabi, S. Chaturvedi, Michael Roth et al. // NAACL-HLT. — 2018.

[22] Matthews B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme // Biochimica et Biophysica Acta (BBA) - Protein Structure. — 1975. — Vol. 405, no. 2. — P. 442–451. — Access mode: https://www.sciencedirect.com/science/article/pii/0005279575901099.

[23] McCoy Tom, Pavlick Ellie, Linzen Tal. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 3428–3448. — Access mode: https://www.aclweb.org/anthology/P19-1334.

[24] Morgenstern Leora, Davis Ernest, Ortiz Charles L. Planning, Executing, and Evaluating the Winograd Schema Challenge // AI Magazine. — 2016. — Apr. — Vol. 37, no. 1. — P. 50–54. — Access mode: https://ojs.aaai.org/index.php/aimagazine/article/view/2639.

[25] Nangia Nikita, Bowman Samuel R. Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4566–4575. — Access mode: https://www.aclweb.org/anthology/P19-1449.

[26] Niven Timothy, Kao Hung-Yu. Probing Neural Network Comprehension of Natural Language Arguments // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4658–4664. — Access mode: https://www.aclweb.org/anthology/P19-1459.

[27] RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language / Alexander Panchenko, Anastasia Lopukhina, Dmitry Ustalov et al. // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". — Moscow, Russia : RSUH, 2018. — P. 547–564. — Access mode: http://www.dialog-21.ru/media/4539/panchenkoaplusetal.pdf.

[28] ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension / Sheng Zhang, Xiaodong Liu, Jingjing Liu et al. // CoRR. — 2018. — Vol. abs/1810.12885. — 1810.12885.

[29] RoBERTa: A Robustly Optimized BERT Pretraining Approach / Yinhan Liu, Myle Ott, Naman Goyal et al. // arXiv preprint arXiv:1907.11692. — 2019.

[30] Roemmele Melissa, Bejan Cosmin Adrian, Gordon Andrew S. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. // AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning. — 2011. — P. 90–95.

[31] Rogers Anna. How the Transformers broke NLP leaderboards. — 2019. — Jun. — Access mode: https://hackingsemantics.xyz/2019/leaderboards/.

[32] Rogers Anna, Kovaleva Olga, Rumshisky Anna. A Primer in BERTology: What We Know About How BERT Works // Transactions of the Association for Computational Linguistics. — 2020. — Vol. 8. — P. 842–866. — Access mode: https://www.aclweb.org/anthology/2020.tacl-1.54.

[33] RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark / Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton et al. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online : Association for Computational Linguistics, 2020. — Nov. — P. 4717–4726. — Access mode: https://www.aclweb.org/anthology/2020.emnlp-main.381.

[34] SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems / Alex Wang, Yada Pruksachatkun, Nikita Nangia et al. // Advances in Neural Information Processing Systems / Ed. by H. Wallach, H. Larochelle, A. Beygelzimer et al. — Vol. 32. — Curran Associates, Inc., 2019. — Access mode: https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[35] Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task / Allyson Ettinger, Sudha Rao, Hal Daumé III, Emily M. Bender // Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems. — Copenhagen, Denmark : Association for Computational Linguistics, 2017. — Sep. — P. 1–10. — Access mode: https://www.aclweb.org/anthology/W17-5401.

[36] What does BERT Learn from Multiple-Choice Reading Comprehension Datasets? / Chenglei Si, Shuohang Wang, Min-Yen Kan, Jing Jiang // ArXiv. — 2019. — Vol. abs/1910.12391.

On Developing a Web Resource to Study Argumentation in Popular Science Discourse

Ilina Daria
Novosibirsk State University, Institute of Philology Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
dviljina@gmail.com

Kononenko Irina
A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
irina_k@cn.ru

Sidorova Elena
A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
lsidorova@iis.nsk.su

Abstract

This paper discusses the experience of developing a web resource intended to study argumentation in popular science discourse. Such type of argumentation is, on the one hand, the main mean of achieving a communicative goal and, on the other hand, often not expressed in explicit form. The web resource is built around a corpus of 2256 articles, distributed over 13 subcorpora. The annotation model, which is based on the ontology of argumentation and D. Walton's argumentation schemes for presumptive reasoning, underlies the argument annotation of the corpus. The distinctive features of the argument annotation model are the introduction of weighting characteristics into text markup through assessing the persuasiveness of the argumentation, as well as highlighting argumentative indicators visually. The paper considers a scenario of argument annotation of texts, which allows constructing an argumentative graph based on the typical reasoning schemes. The scenario includes a number of procedures that enable the annotator to check the quality of the text markup and assess the persuasiveness of the argumentation. The authors have annotated 162 texts, using the developed web resource, and as a result, identified the most frequent schemes of argumentation (Example Inference, Cause to Effect Inference, Expert Opinion Inference), as well as described some specific indicators of frequent schemes. Based on the above-mentioned outcomes, the authors listed the indicators of the most frequent schemes of argumentation and made some recommendations for annotators about identifying the main thesis.

Keywords: popular science discourse; text corpus; argument annotation of text; argumentation indicator; annotation scenario

DOI: 10.28995/2075-7182-2021-20-318-327

О создании интернет-ресурса для исследования аргументации в научно-популярном дискурсе

Ильина Дарья
Новосибирский государственный университет, Институт филологии СО РАН, Новосибирск, Россия
dviljina@gmail.com

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

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

Аннотация

В статье представлен опыт создания лингвистического интернет-ресурса, предназначенного для исследования аргументации в научно-популярном дискурсе, в котором аргументация, с одной стороны, является основным средством достижения коммуникативной цели, а с другой, — часто не выражена в явном виде. Создан корпус объемом 2256 статей, распределенных по 13 подкорпусам. В основу аргументативной разметки корпуса текстов положена модель аннотирования на базе онтологии аргументации и перечня типовых схем рассуждения теории Д. Уолтона. Отличительной особенностью рассматриваемой модели является введение

в разметку весовых характеристик для оценки убедительности аргументации, а также явное выделение индикаторов. Рассматривается сценарий аргументативной разметки текста, который обеспечивает построение графа аргументации с использованием знаний о типовых схемах аргументации. Сценарий включает ряд процедур, позволяющих аннотатору проверить качество разметки и оценить убедительность выявленной аргументации. В результате разметки 162 текстов выявлены наиболее частотные схемы аргументации, такие как «От примера», «От причины к следствию», «От эксперта», и описаны их специфические индикаторы. Опыт разметки текстов позволил создать список индикаторов наиболее частотных схем и сформулировать некоторые рекомендации для разметки аргументации, в частности, относительно выявления главного тезиса.

Ключевые слова: научно-популярный дискурс; корпус текстов; аргументативная разметка текста; индикатор аргументации; сценарий разметки

1 Introduction

Over the past two decades, the Internet serves as the main platform for sharing ideas, gathering knowledge, disputes, and debates. Justifying opinions or statements is a field of argumentation theory that studies the use of arguments in discourse from a philosophical, linguistic, cognitive, and computational perspective. The analysis of argumentation, in particular, includes the transformation of unstructured text into “chains” or graphs of related structured arguments, which allows not only to evaluate individual statements, but also identify the relations between them, taking into account the focus on supporting or refuting the main thesis discussed by the author of the publication. Modern theoretical studies of argumentation, in fact, are connected with the practice of argumentation, and in this case, the study of this practice on a mass scale should be a priority for this area [9]. Along with theoretical studies, in recent years, scholars actively work to automate the extraction of arguments from texts [15].

The proper development of these studies requires the creation of argumentation corpora, where text fragments are marked up with the components of argumentative structures and relations between them. At the moment, there exist several annotated corpora of argumentative texts, most of them are monologue texts in English. The most famous resource with argument annotation is AIFdb, the former Araucaria corpus [22], which contains news articles, records of parliamentary and political debates. Another resources in the German language are the corpus of the University of Darmstadt, which includes subcorpora of student persuasive essays [24], news texts, and scientific articles; the Potsdam Corpus, which contains a small collection of microtexts, later translated into English [19]. There are projects in some other languages – Italian, Greek, Chinese, but as far as is known, not in the Russian language.

In most projects, argument annotation includes text segmentation, highlighting units of argumentation, marking up roles (premise, conclusion) and relations (support / attack) without specifying the structure of arguments. The exceptions are corpora created using the OVA system (Online Visualization of Argument – the successor of Araucaria) [4], where the developers implemented an extended annotation of the argumentative structure related to specific argumentation schemes (based on the Argumentation Schemes by D. Walton) [25]. Argumentation schemes formalize certain reasoning patterns used for persuasion, that is why it is so important to study statistics and different contexts of using a particular scheme within the corpus. This is confirmed both by the rapid growth of the AIFdb corpus [6], which has already absorbed some argumentation markup systems [2–3], and by the increasing interest in the problems of automatic argument extraction, where annotated data is required. However, as noted in [16], existing annotated corpora that were used to automatically classify argument schemes have several shortages such as limited validation, restricted size, or poor representation of a broad range of scheme types.

The proposed work was carried out as part of a research project aimed at creating a corpus of Russian-language popular science texts with extended argument annotation. The texts were annotated manually based on the argumentation model developed by the project participants.

Section 2 defines features of popular science discourse in the aspect of argumentation and presents information about developing corpus accompanied with some statistics. The following sections provide an introduction to the argument annotation model (section 3) and also describe the scenario of effective work for the annotator (section 4). Section 5 presents the outcomes of preparing argument annotation for 162 texts, based on which the authors outline patterns and offer some recommendations for the annotator of argumentation.

2 Features of Popular Science Discourse

Popular science texts are mainly intended to present the results of research to the general audience and to prove their validity. They are of undoubted value for the theory of argumentation, since they represent two views on the world – the scientific and the naive ones. B.V. Kasevich described the main difference between the naive view that is always seeking continuity (usually imaginary) and the scientific view that is seeking completeness (fundamentally unattainable) [10].

The author may be a researcher, a bearer of a scientific view, or a journalist, a bearer of a naive view. Regardless, the main goal of the author is to break the continuity of the reader's naive worldview and replace it (a fragment of the worldview associated with the topic of the text) with a scientific or at least more scientific than it was before reading the text.

The audience of a popular science text is broad, non-specific, including people of different levels of education, with different knowledge about the subject of the text, but anyway, their viewpoint is naive, non-professional. Moreover, it is important to notice that popular science texts have optional, non-obligational nature because the audience may choose to read or not read them. This challenges the author to make the text interesting for the reader, to draw his attention to the subject being described.

Non-specificity of the reader and his optional assignment of the text content force the author to simplify or even mask the argumentation, in addition to the necessity of enhancing the attractiveness of the text. Too much explicit and formalized argumentation, being cumbersome and “boring”, didactic, will rather turn the reader away than be accepted by him.

Thus, the scholar who annotates popular science articles has to restore many implicit statements. The authors of this article have found out that about 10% of the annotated statements are implicit. Implicit content recovery is important for identifying not only separate arguments (schemes of argumentation consisting of premises and a conclusion), but also structures of argumentation, i.e. argumentation graphs, which support one main thesis, the main idea (for more details see [7]).

The process of selecting texts for developing annotated corpus in Russian language was operated automatically from open sources, such as “Science and Life” (nkj.ru), “STRF” (strf.ru), “Postnauka” (postnauka.ru), etc. The compilers have accompanied the articles with information about the author(s), date of publication, and subject (if it was indicated in the original source) and have grouped them into the source-based subcorpora. Using automated tools, the compilers collected about 2.3 thousand popular science articles, and arranged 13 subcorpora.

3 Argument Annotation Model

The argumentative structure explicates the processes of reasoning and persuasion that underlie the text, highlights the components of the argumentation field and the relations between them (controversial thesis, arguments for or against). To describe the arguments and argumentative structures the authors apply an extended version of the argumentation ontology – Description Logic ontology [21]. The distinctive features of this ontology [27] are a branched system of classes for representing typical schemes of reasoning and tools for modeling and analyzing the persuasiveness of argumentation. The descriptions are based on the AIF format [5], according to which arguments are represented in the form of a directed graph with two types of nodes: information nodes (statement vertices) and scheme nodes (argument vertices).

«Example_Inference»		
Statement role	**Statement type**	**Statement description**
TypicalObject_Premise	TypicalObject_Statement	***a*** is typical of things that have ***F*** and may or may not have ***G***
CaseProperty_Premise	CaseProperty_Statement	In this case, the individual ***a*** has property ***F*** and also property ***G***
Conclusion	GeneralProperty_Statement	Generally, if ***x*** has property ***F*** then (usually, probably, typically) ***x*** also has property ***G***

Figure 1: The Example Inference scheme of argumentation

Fig. 1 shows an example of an argumentation scheme. It combines the statements found in the text (two premises and a conclusion) into a single structure. The same statement can be included in different structures, thereby linking the "minimal" units of reasoning (arguments identified in the text) into a single chain and in the general into a graph of argumentation.

One of the key aspects of argumentation is a conflict between arguments. While typical arguments and their relations are aimed at supporting a certain statement-thesis, conflict – criticizing or disproving a thesis. In the argument annotation model conflict is represented by a scheme that defines a relation either between two statements or between a statement and an argument supporting a thesis.

The model of argument annotation of text, in accordance with the ontology, can be represented as the following system:

<T, S, Arg, C, R_e, Ind, W_S>, where *T* is a text, *S* – a set of annotated statements, *Arg* – a set of arguments that are instances of argumentation schemes, *C* – a set of conflicts, R_e – a set of relations between statements specifying "conditional" equivalence, *Ind* – a set of argumentation indicators, W_S – assessments of the author's belief in the truth of the statements. The argument in this model specifies an n-ary relation over statements with a special position assigned to the conclusion, and the conflict is a binary directional relation.

Fig. 2 shows an example of text annotation built in accordance with the proposed model.

Figure 2: Argument annotation of a text

Argumentation in a marked-up text is presented by (a) a set of annotated statements and indicators and (b) a graph representation of a set of arguments corresponding to relations between statements. The graph representation together with the textual one gives a complete overview of arguments and text fragments covered by them. In Fig. 2, statements (rectangular vertices) correspond to instances of type statement classes, and arguments (vertices with rounded edges) correspond to scheme instances.

Distinctive features of the proposed argument annotation model are introduction of weighting characteristics into the markup to assess the persuasiveness of the argumentation, and also the explicit identification of indicators that not only point out the presence of arguments and their types, but can also affect the general assessment of the persuasiveness of the argumentation.

4 Argument Annotation Scenario

The annotation scenario includes the main stage when the annotator selects statements and constructs the argumentation graph and the stage when the annotator analyzes the argumentation.

The main stage includes the following steps:

1. investigate argumentation indicators found automatically;
2. segment the text into argumentative discourse units (ADUs), i.e. sentences, clauses or minimal text spans that have propositional content including nominalized propositions and prepositional phrases with the meaning of cause, effect, concession, contrast;

3. select text fragments related to argumentation and create on their basis statement nodes for the argumentation graph (when forming the description of the statement node, the annotator can modify the initial text fragment to avoid ambiguities, resolve anaphora or restore ellipsis);
4. identify implicit statements related to argumentation and create graph nodes that correspond to no text fragment;
5. define the roles for each statement (conclusion or premise) and build argument nodes connecting statement nodes into a single graph structure; relations between statement and argument nodes are directed (from premise to argument or from argument to conclusion);
6. determine a scheme for an argument node using a multidimensional hierarchical classification of reasoning schemes; for any scheme, a semiformal description of each element of its structure is given (see Fig. 1);
7. detail the structure of each argument based on the corresponding argumentation scheme; at this stage, the fields in the structure of the arguments are filled with the appropriate statements;
8. identify "conditionally" equivalent statements, i.e. statements that have the same propositional content, but differ in the degree of detail (in dictum) or in modal component (influencing persuasiveness).

To proceed to the next stage – the analysis and validation of the annotation – the resulting graph must be carefully checked, since even a small change in its structure can lead to significant discrepancies in the final assessment. Automatic graph checking includes loop search and connectivity analysis. It is necessary to notice that the graph may contain cycles resulting from conflicts, but looping of supporting argumentation chains is not allowed.

There are several procedures of validation check that has to be performed by the annotator:

- analysis of argumentation indicators that might not be included in the markup,
- comparative analysis of typical annotation elements with their implementations in other annotated texts,
- comparative study of argumentative relations (specifically, analysis of correlation with rhetorical annotation [13, 17]),
- analysis of disconnected subgraphs and identification of causes.

Development of the methodology is aimed at building a training base of the argument mining parser for texts in Russian. To validate such corpus, a formal check is absolutely needed, including the analysis of the graph for coherence and absence of cycles, and content check, which may consist of assessing the inter-annotator agreement. Computing inter-annotator agreement on a manually annotated corpus is crucial to evaluating the reliability of annotation. One of the previous attempts to overcome the problem of low inter-annotator agreement arising from the complexity of the underlying argumentation ontology has been to pre-select from existing larger scheme typologies (see [18]). However, note that annotators can not reach absolute agreement. In contrast to artificial formal-logical methods of proving a thesis, argumentation in popular science discourse is often based on the so-called "starting point of arguments". They are "the preferable, comprising values, hierarchies, and lines of argument" that appear to be more convincing for groups of individuals" [14, 20]. It is fundamentally impossible to classify these preferences in such a way that classes do not overlap and do not include each other. At this moment no study in this direction has been performed due to the limited scope of the annotation trial.

The stage of argumentation analysis consists in assessing the degree of persuasiveness of the statements and annotating them with weight characteristics (ranging from 0 to 1, where 1 corresponds to the maximum persuasiveness, and 0 – to the minimum one). The weighted assessment depends on the specified audience: general, scientific, or adolescent. After setting the initial weights, the weights for the entire graph (all arguments and theses) can be calculated to assess the degree of persuasiveness of the argumentation for a particular audience [26]. Obtained result may be compared with the opinion of the annotator to make sure that the graph is built correctly and to identify inaccuracies in the markup at the structural level.

5 Features of Argument Annotation

When annotating the argumentation, the above stages of text annotation are implemented.

5.1 Argumentation Indicators

To draw the attention of annotators to arguments presented in texts explicitly and to assist with highlighting the boundaries of ADUs and choosing appropriate argumentation schemes, the corpus is provided with the system of preliminary text processing. This procedure simplifies detecting specific hints in the text, such as various kinds of verbal clichés. These clichés indicate the presence of an argument in the text [8].

The automatic search for indicators is operating based on the pattern constructions that describe the classes of language expressions with regard to possible lexico-semantic classes, grammatical forms, punctuation, and compatibility in multi-word strings [12]. The experts suggest indicator patterns according to the analysis of means of expression of argumentation. Patterns may be expanded taking into account the variants through the constructing samples with variables and iterative search methods [1].

At the moment, one marked-up text contains an average of 18 arguments, among which 0.5 are conflicting. Table 1 shows most frequent argumentation schemes and their indicators.

Schemes	**Count**	**Examples of indicators** (given informally)
Example	257	например ‘for example’; в частности ‘particularly’; привести <Verb, Pers=1>… пример ‘let’s give …example’; показывать<Verb, Tense=pres>… как/что ‘shows…how/that’; мочь _percep [видеть/ наблюдать], что ‘(can) see / observe… that’
Cause to Effect	249	поэтому/потому (что)/так как ‘because (of the fact that)’; дело в том, что ‘the fact is that’; привести <Verb, Tense= pres/past>…к ‘result in’; объясняться / объяснить…тем, что ‘account for the fact that’; связан…с тем, что ‘connected with the fact that’; это объяснимо ‘it is explainable’; причина этого ‘the reason is’
Expert Opinion	203	по мнению _expert [ученый/эксперт] <Noun, Case=gen> ‘according to _expert’; _expert <Noun, Case=nom>_speech [утверждать/писать] / _intel-act [доказать/обнаружить] / _eval [соглашаться]<Verb>…, что ‘scientists claim/consider/agree that’; согласно/по _speech-prod [слово]<Noun, Case=dat> / _mental-prod [представление/гипотеза]<Noun, Case=dat> _expert <Noun, Case=gen> ‘according to the ideas of _expert’; _speech-prod [работа/статья] …_expert … _intel-act [показать/продемонстрировать], что ‘the work(s) / paper(s) / article(s) …_expert… show that’; подробнее (об этом) см. ‘for more details see’
Logical Conflict	110	неверно, что ‘it is not true that’; несмотря на …, ‘in spite of’; с одной стороны… с другой (же)… ‘on the one hand…on the other’
Practical Reasoning	91	для … нужно / требуется / необходимо ‘to do/for …, it is required / needed’; <Verb, mood=imperative >
Analogy	54	похожий… на ‘is similar to’; похожая ситуация… наблюдаться/сложиться ‘a similar state of affairs…(developed / observed)’
Sign	47	означать, что ‘it means that’; указывать на то, что ‘it indicates that’
Position to Know	35	по/согласно _observ-data [наблюдение/данные] ‘according to the data’; подтверждать/подтверждаться… _observ-data <Noun, Case=nom, instr> ‘confirmed by observations’; результаты (эксперимента) показывают ‘the results (of the experiment) demonstrate that’

Table 1: Indicators of argumentation schemes

Among the most frequent schemes of argumentation, *Example Inference* and *Expert Opinion Inference* turned out to be well formalized.

Specific indicators of the *Example Inference* (see Fig. 1) are lexemes belonging to the family of words with the root "example". The place of the indicator regarding the text fragments of the argument helps identify the role of the corresponding statement in the structure of the argument (Conclusion, Premise):

(1) *<Conclusion> Пример: 'Example:' <CaseProperty_Premise>*
<Conclusion> Привед<у/ем>… пример<а,ы,ов> 'let's give<numeral/quantifier/article> example' <CaseProperty_Premise>
<Conclusion>. Например, 'For example,' <CaseProperty_Premise>

Another less specific (weak), but frequent indicators of the scheme:

(2) <Conclusion>. *Так, 'Thus' <CaseProperty_Premise>*

This indicator can also be a part of a complex subordinating conjunction *мак, что* '*so that*' or *мак, как* '*in a way that*' or an adverb followed by a comma that marks the segment boundary (see Fig. 3).

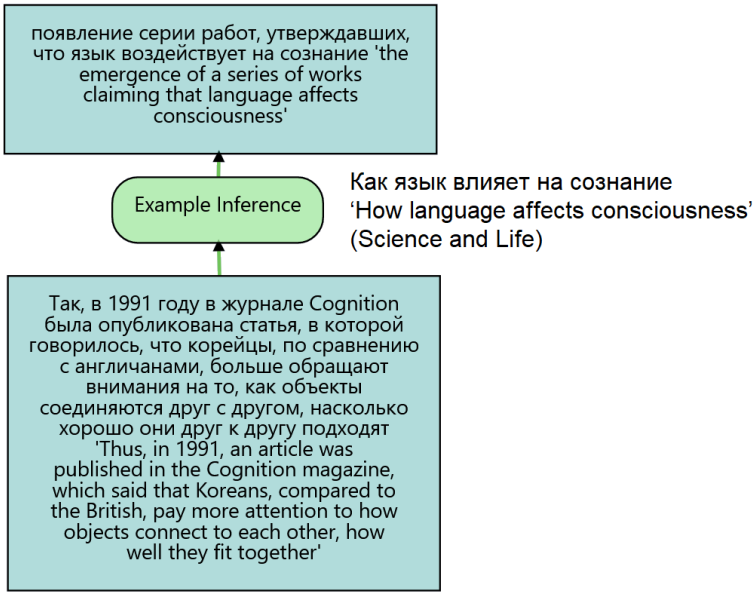


Figure 3: Example of the indicator of the Example_Inference

(3) *<Conclusion>: <CaseProperty_Premise>*
<Conclusion>. <CaseProperty_Premise>

The last two indicators are the weakest of the listed above: the colon can also express causal, explanatory, authorization and other relationships, and the full stop – any border between affirmative non-exclamatory sentences.

One more frequent scheme of argumentation with specific indicators is Expert Opinion Inference (see [11]).

To find out what proportion of argument schemes can be detected using indicators, an experiment was carried out with a system of 65 lexico-syntactic patterns that represent 4 classes of constructions of "expert opinion" indicators. The precision of the expertly created patterns reached 74.18% on the training collection and 65.73% on the test one. The precision measure on the training set demonstrates the frequency of an indicator constructions in argumentation (as opposed to non-argumentative narration). False positive results are caused by lexical, morphological and graphical homonymy and structural ambiguity.

5.2 Establish the Main Thesis

Annotated popular science articles contain one main idea (main thesis) or, less often, several ones. Accordingly, the annotation has one or several argumentative structures. The latter is observed mainly in texts where the authors do not limit themselves to a narrow topic, and do seek to tell about the entire field of research.[1]

In texts with one argumentative top, the statement of the main thesis, if it is explicit, is introduced either in the title, or in the lead (abstract under the heading), or in the first paragraph (most often in the first sentence or its dictum part), or in the last paragraph.

When extracting the main thesis from the title, it is necessary to take into account two features of this element of the text. Firstly, it is often a nominative sentence, i.e. to extract the main thesis, the annotator has to transform it into a verbal sentence (as a rule, representing a proposition of existence or functioning): for example, "Power of vowels" → "There is a power of vowels" → (eliminating figurative component) → "Vowels influence human actions". Secondly, the heading, being the most important means of drawing the reader's attention to the text, often reflects the main thesis not quite accurately, not specifically enough or too figuratively (see Table 2). This is due to the fact that the purpose of the author when composing the title is to point out an aspect of the topic that will be interesting or attractive to as many readers as possible (for example, a practical one), or to point out a more general topic.

Title	Main idea
Хочешь выиграть – подумай об этом на иностранном языке 'To win, think in a foreign language'	Думая на иностранном языке, люди принимают более взвешенные решения 'Thinking in foreign language helps people make rational decisions' (the 1st sentence)
Как язык влияет на сознание 'How language affects consciousness'	другой [иностранный] язык в буквальном смысле расширяет наше сознание и заставляет иначе взглянуть на мир 'a foreign language literally expands our consciousness and makes us look at the world differently' (the dictum part of the 1st sentence of the last paragraph)
Власть гласных 'The power of vowels'	Минимальные составляющие компоненты слов действительно способны изменять наше восприятие не только всего слова, но и объекта, который оно обозначает 'The minimal components of words are really capable of changing our perception of not only the whole word, but also the object that it stands for' (the dictum part of the last sentence)

Table 2: Examples of expressing the main idea in the title and in other parts of the text

Conclusion

To support the argument annotation of texts and the studies of argumentation, the project team (including the authors) has developed web platform (https://geos.iis.nsk.su/arg) [23] which provides the user with a set of specialized tools: text markup tools, graph editor, search services for finding arguments in annotated corpora, a linguistic module that performs preprocessing of texts and highlighting indicators, and a computational module that conducts an assessment of the persuasiveness of the argumentation depending on the initial weights of the statements specified by the annotator.

The presented above annotating technique covers the traditional division of both argument components into premises and conclusions as well as argumentative relations into support and attack. Moreover, being based on D. Walton's theory, the technique allows for a large subset of argumentation schemes (44 inference schemes and 23 conflict relations). Most prominent features of annotation procedure are as follows:

[1] E.g., Levontina I. Russkiy Natsionalnyy [The Russian National]. Elements.
URL: https://elementy.ru/nauchno-populyarnaya_biblioteka/432329/Russkiy_natsionalnyy (accessed 10.05.2021).

- reliance on the ontological model of argument annotation,
- consideration of argumentation indicators for the detection of arguments,
- comparative analysis and identification of correlations between argumentative and rhetorical structures [13],
- use of mathematical modeling methods to control the annotation process.

Based on the experience obtained during the process of text annotation, some recommendations for annotators have been prepared: rely on the list of indicators, identify the main thesis, differentiate argumentation and explanation. Outcomes of this research will underlie a detailed instruction on argument annotation. Since it is especially difficult to match arguments to reasoning schemes, the starting point for these instructions will be critical questions related to the main aspects of scheme classification.

Further attention is required to study stable combinations of argumentation schemes, examine the influence of various argumentative structures on the weights of propositions, and reveal linguistic indicators of these phenomena.

References

[1] Akhmadeeva I., Kononenko I., Salomatina N., Sidorova E. Indicator Patterns as Features for Argument Mining // International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). — Novosibirsk, Russia, 2019. — P. 0886-0891.

[2] Bex F, Lawrence J, Snaith M, Reed C. Implementing the argument web. — Int. J. of Communications of the ACM. — Vol. 56(10), 2013. — P. 66–73.

[3] Bex F, Reed C. Dialogue templates for automatic argument processing // Computational Models of Argument: Proc. of the 4th Int. Conf. COMMA (Vienna), IOS Press, 2012. — P. 366-377.

[4] Bex F., Snaith M., Lawrence J., Reed C. ArguBlogging: An application for the argument web. — Int. J. of Web Semantics: Science, Services and Agents on the World Wide Web, 2014. — Vol. 25. — P. 9-15.

[5] Chesñevar C.I., McGinnis J., Modgil S., Rahwan I., Reed C., Simari G., South M., Vreeswijk G., Willmott S. Towards an argument interchange format. — The knowledge engineering review, 21(4), 2006. — P. 293-316.

[6] Corpus AIFdb. Access mode: http://corpora.aifdb.org/.

[7] Eemeren F.H. van. The State of the Art in Argumentation Theory [Sovremennoe sostoyanie teorii argumentatsii], Crucial Concepts in Argumentation Theory [Vazhneishie kontseptsii teorii argumentatsii]. — St. Petersburg, Faculty of Philology of St. Petersburg State University, 2006. — P. 25-26.

[8] Eemeren F.H. van., Houtlosser P., Snoeck Henkemans F. Argumentative Indicators in Discourse: A Pragma-Dialectical Study, Argumentation Library. — Vol.12. Springer, Dordrecht, 2007. https://doi.org/10.1007/978-1-4020-6244-5

[9] Hinton M. Corpus Linguistics Methods in the Study of (Meta)Argumentation, Argumentation, 2020. https://doi.org/10.1007/s10503-020-09533-z

[10] Kasevich V.B. Buddhism. Worldview. Language [Buddizm. Kartina mira. Yazyk]. — Saint Petersburg, Saint Petersburg State University Publ., 2004.

[11] Kim I.E., Ilina D.V. Language expression of the argumentative framework “From popular opinion vs. from expert opinion” in the text of popular science article [Iazykovoe vyrazhenie argumentativnoi struktury «Ad populum» – «Ot eksperta» v tekste nauchno-populiarnoi stat`i]. — Vestnik NSU, Series: History and Philology, 2019. — Vol. 18. — N. 9. — P. 27–35.

[12] Kononenko I., Sidorova E. Development of the Lexicon of Argumentation Indicators // In: Kuznetsov S., Panov A. (eds) Artificial Intelligence. RCAI, Communications in Computer and Information Science. — Vol 1093, Springer, Cham., 2019. — P. 154-168.

[13] Kononenko I.S., Sidorova E.A., Akhmadeeva I.R. Comparative analysis of rhetorical and argumentative structures in the study of popular science discourse // Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf. “Dialogue”, 19(26), 2020. — P. 432-444.

[14] Kosarev A.V. On Ch. Perelman’s Rhetoric Theory of Argumentation [O ritoricheskoi` teorii argumentatcii KH. Perel`mana]. — Siberian Journal of Philosophy [Sibirskii` filosofskii` zhurnal], 2019. — Vol. 17(2). — P. 174–188.

[15] Lawrence J., Reed C. Argument mining: A survey. — Computational Linguistics, 45(4), 2019. — P. 765-818.

[16] Lawrence J., Visser J., Reed C. An Online Annotation Assistant for Argument Schemes // Proceedings of the 13th Linguistic Annotation Workshop eds. A. Friedrich, D. Zeyrek, and J. Hoek, Association for Computational Linguistics, 2019. — P. 100–107.

[17] Musi E., Alhindi T., Stede M., Kriese L., Muresan S., Rocci A. A Multi-layer Annotated Corpus of Argumentative Text: From Argument Schemes to Discourse Relations // Language Resources and Evaluation (LREC’2018): Proc. of the 11th Int. Conf. (Miyazaki, Japan), 2018. — P. 1629-1636.

[18] Musi E., Ghosh D., Muresan S. Towards feasible guidelines for the annotation of argument schemes // Proceedings of the third workshop on argument mining (ArgMining-2016), 2016. — P. 82-93.
[19] Peldszus A., Stede M. An annotated corpus of argumentative microtexts // Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation. — London, College Publications, 2016. — Vol. 2. — P. 801–816.
[20] Perelman Ch., Olbrechts-Tyteca L. The New Rhetoric: A Treatise on Argumentation. — Transl. by J. Wilkinson, P. Weaver, 2019. — Notre Dame, IN, Uni. of Notre Dame Press, 1971, 576 p.
[21] Rahwan I, Banihashemi B, Reed C, Walton D, Abdallah S. Representing and classifying arguments on the semantic web. — The Knowledge Engineering Review, 26(4), 2011. — P. 487-511.
[22] Reed C., Rowe G. Araucaria: Software for argument analysis, diagramming and representation, International Journal on Artificial Intelligence Tools, 2004. — Vol. 13(4). — P. 961–979.
[23] Sidorova E.A., Akhmadeeva I.R., Zagorulko Yu.A., Sery A.S., Shestakov V.K. Platform for the study of argumentation in popular science discourse [Platforma dlia issledovaniia argumentatcii v nauchno-populiarnom diskurse]. — Ontology of designing, 2020. — Vol.10, 4(38). — P. 489-502. DOI: 10.18287/2223-9537-2020-10-4-489-502
[24] Stab C., Gurevych I. Identifying Argumentative Discourse Structures in Persuasive Essay // Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014. — P. 46–56.
[25] Walton D., Reed C., Macagno F. Argumentation schemes. — Cambridge University Press, 2008.
[26] Zagorulko Yu., Domanov O., Sery A., Sidorova E., Borovikova O. Analysis of the Persuasiveness of Argumentation in Popular Science Texts // S.O.Kuznetsov et al. (Eds.): RCAI 2020, LNAI 12412, 2020. — P.351-367. https://doi.org/10.1007/978-3-030-59535-7_26
[27] Zagorulko Yu.A., Garanina N.O., Borovikova O.I., Domanov O.A. Argumentation modeling in popular science discourse using ontologies [Modelirovanie argumentatcii v nauchno-populiarnom diskurse s ispol`zovaniem ontologii`]. — Ontology of designing, 9(4), 2019. — P. 496-509.

Defining discourse relations: Supracorpora database of connectives

Inkova O.
Institute of Informatics Problems, FRC
CSC RAS / Moscow, Russia
University of Geneva, Geneva,
Switzerland
Olga.Inkova@unige.ch

Abstract

The research is focused on definitions of discourse relations, a topic that is currently little-studied. The paper gives a brief overview of existing solutions for discourse relations definitions: Rhetorical Structure Theory (RST), Segmented Discourse Representation Theory (SDRT), Penn Discourse Treebank (PDTB), and Cognitive approach to Coherence Relations. The author shows criteria used to define a discourse relation, or, in case of a narrower definition, a logical-semantic relation, in these approaches and outlines the shortcomings of the described definitions. The author also describes the principles used to build the classification and the definitions of logical-semantic relations (LSR) in the Supracorpora Database of connectives (SDB). The classification is based on four basic semantic operations upon which rests every LSR's definition: implication, location on the chronological scale, comparison, correlation between specific and general or an element and a set. The classification consistently distinguishes the levels at which the LSR can be established: propositional, illocutionary, and metalinguistic. Each LSR is defined on the basis of these two criteria. Thus, for example, for the LSR of alternative based on the comparison operation, one has the choice between the LSR of propositional, illocutionary and metalinguistic alternative (*We will go to the mountains or to the sea* vs. *Put the gun away, or are you scared?* vs. *The symbol of the year or, simply speaking, cutie-pie*). In case of LSRs based on implication or comparison, the polarity criterion is added, distinguishing whether the LSR is established between p and q or their negative correlates $\neg p$ and $\neg q$ are also to be taken into account in order to obtain a correct interpretation (cf. well-known descriptions of how the Russian conjunction *no* 'but' functions). In addition, semantic and pragmatic characteristics of the context are also considered in the classification. For example, in the case of the LSR of specification and generalization, the semantic correlation between p and q (together with their intensional and extensional interpretations) is taken heed of. Several definitions of LSR and corresponding examples are provided. Thus, the LSR of extensional specification is defined as follows: based on the operation of correlation between the general and the particular; established at the propositional level; X contains a generalized notion or state of things p; Y contains a more particular q-notion, limiting p-extensional. And the LSR of intensional specification is defined as follows: based on the operation of correlation between the general and the particular; established at the metalinguistic level; X contains a generalized concept or state of things p; Y contains a more particular q-notion, limiting p-intensional. The definitions used in the SDB definitions make it possible to evaluate, on the basis of the proposed criteria, the semantic closeness of relations and increase the level of consistency in the work of experts and annotators. That in turn increases the value of the annotated material, and therefore its reliability.

Keywords: semantics; discourse relations; supracorpora database; connectives; corpus linguistics
DOI: 10.28995/2075-7182-2021-20-328-338

Определения дискурсивных отношений: опыт Надкорпусной базы данных коннекторов

Инькова О.Ю.
ИПИ ФИЦ ИУ РАН / Москва, Россия
Женевский университет / Женева,
Швейцария
Olga.Inkova@unige.ch

Ключевые слова: семантика; дискурсивные отношения; надкорпусная база данных; коннекторы; корпусная лингвистика

1 Вводные замечания

Настоящее исследование посвящено мало изученной проблеме определения дискурсивных отношений. В работах, им посвященным, рассматриваются, как правило, лишь вопросы их классификации, также вызывающие многочисленные дискуссии; см., например, [Roze 2013; Chiarcos 2014; Benamara, Taboada 2015; Bunt, Prasad 2016; Demberg et al. 2019, Гончаров 2021]. Хотя эти два вопроса тесно связаны, вопрос об определениях дискурсивных отношений и критериях, лежащих в их основе, заслуживает отдельного рассмотрения.

Как известно, существуют два подхода к определению отношений:

- отношение определяется на основе семантики его прототипического показателя; см., [Halliday 1985; Knott 1996; Stede, Umbach 1998; Alonso et al. 2002; Fraser 2009] и др.;
- отношение определяется на основе коммуникативного задания говорящего или семантических и прагматических характеристик контекста; см., например, [Grosz, Sidner 1986; Mann, Thompson 1988; Asher, Lascarides 2003; Prasad et al. 2017]. Тем не менее, даже в этом случае во многих корпусах предлагается список прототипических показателей, которые в процессе аннотировании помогают установить, о каком отношении идет речь; см. ANNODIS [Muller et al. 2012], [PDTB Project].

Мы начнем с краткого обзора существующих решений для определения дискурсивных отношений, а именно трех подходов, используемых для аннотирования текстов (Теория риторической структуры, Теория сегментной репрезентации дискурса, Пенсильванский корпус), а также когнитивного подхода к отношениям связности. Затем представим принципы, разработанные для определения отношений и их аннотирования в Надкорпусной базе данных коннекторов[1], и приведем несколько примеров таких определений.

2 Теория риторической структуры

В Теории риторической структуры (RST) отношение связности определяется на основе коммуникативных намерений и представлений говорящего (Г), а также представлений слушающего (С) по четырем параметрам: ограничения на сателлит (Ст); ограничения на ядро (Я); ограничения на сочетание ядра и сателлита; достигаемый эффект. Приведем в качестве примера определение отношения Evidence [2] и иллюстрирующий его пример, заимствованный с сайта RST [http://www.sfu.ca/rst/]:

Отношение: Evidence.
Ограничения на Я: С, по мнению Г, не верит в Я в достаточной степени. **Ограничения на Ст:** Г верит в Ст или считает его достаточно достоверным. **Ограничения на сочетание Я+С:** Вера С в Ст увеличивает его веру в Я. **Эффект:** Вера С в Я увеличена.

(1) 1) Darwin as a Geologist 2) he tends to be viewed now as a biologist, 3) but in his five years on the Beagle his main work was geology, 4) and he saw himself as a geologist. 5) His work contributed significantly to the field. 6) Scientific American, Sandra Herbert, May 1986, o. 116.

Согласно предлагаемому на сайте описанию, в (1) отношением Evidence связаны между собой дискурсивная единица 1), которая является заголовком статьи, и дискурсивные единицы 3)-5). Единицы 1) и 2) связаны отношением Concession ‘Уступка’. Какими отношениями связаны между собой 2) и 3), а также 3), 4) и 5) остается из анализа неясным. Неясной остается также функция дискурсивной единицы 6). Она просто оставлена на верхнем уровне схемы.

[1] Представительный фрагмент Надкорпусной базы доступен по ссылке: http://a179.frccsc.ru/PublicLingvoProjects/main.aspx.

[2] Мы предпочитаем сохранять английские названия отношений, чтобы не создавать дополнительную терминологическую путаницу, поскольку как названия отношений, так и объем даже одинаковых терминов не совпадают в разных классификациях.

Определения RST можно упрекнуть в некоторой тавтологичности. Особенно это касается многоядерных отношений. Так, отношение Disjunction определено следующим образом: а) элемент представляет собой (не обязательно исключающую) альтернативу другому; б) С распознает, что связанные элементы являются альтернативами. Отношение Sequence 'Следование': а) между ситуациями, описанными в Я, существует следование во времени; б) С распознает следование во времени между Я [http://www.sfu.ca/rst/]. Однако на каком основании – семантическом, прагматическом, ... – слушающий распознает эти отношения остается неясным. Эта тавтологичность отчасти преодолена в пособии по аннотированию RST Treebank. Из определения исчезает семантический эффект, и Sequence, например, определяется так: "A Sequence is a multinuclear list of events presented in chronological order" [Carlson, Marcu 2001: 67].

Кроме того, в RST набор отношений разделен на две группы, соответствующие grosso modo семантическим (Subject Matter Relations) и прагматическим (Presentational Relations) отношениям, внутри которых они даны списком, что не позволяет увидеть, что некоторые отношения, в том числе находящиеся в разных группах, имеют общие семантические свойства.

3 Теория сегментной репрезентации дискурса (SDRT)

В отличие от RST, SDRT ставит своей целью моделировать процесс интерпретации текста, определяя правила выведения дискурсивных отношений, которые разделены на три группы: действующие на пропозициональном уровне, или семантические (content-level relations), действующие на уровне высказывания, или прагматические (meta-talk relations), и структурные (text-structuring relations), названые так, поскольку они накладывают определенные структурные ограничения на связываемые ими дискурсивные единиц [Asher, Lascarides 2003][3]. В рамках этих групп семантическая близость входящих в них отношений не эксплицируется.

Само отношение определяется на основе семантической, прагматической и лексической информации при помощи порождающих правил. Их формализация менялась в ходе развития SDRT, но основные ее положения, сформулированные, начиная с [Asher 1996], на языке связующей логики (Glue Logic), сводятся к следующему. Отношения считаются трехместными предикатами $R(\alpha,\beta,\lambda)$, указывающими на то, что α и β (пропозиции в сочетании с характеризующими их коммуникативными намерениями) связаны отношением R в дискурсивном сегменте λ. $?(\alpha,\beta,\lambda)$ означает, что между α и β существует дискурсивное отношение, но мы еще не знаем, какое именно. Язык связующей логики включает также формулы типа [A]K, указывающие, что A является элементом пропозиционального содержания минимальной дискурсивной единицы K, а также классический оператор импликации $\rightarrow$ и условный немонотонный оператор >.

Рассмотрим, как представлено отношение Explanation, основанное, как и Evidence в RST, на причинных отношениях, на примере, приводящемся практически во всех работах, посвященных теоретическим принципам SDRT.

(2) Max fell. John pushed him. [пример из Asher 1993]

Если α и β связаны между собой и K_α описывает положение вещей, в котором некоторый субъект падает, а K_β – положение вещей, где того же субъекта толкают, мы можем предположить, на основании наших знаний о мире, что второе положение вещей может быть причиной первого, что выражается предикатом $\text{Cause}_\text{D}\ (\beta,\alpha,\lambda)$. Нижний индекс D в предикате Cause означает, что мы имеем дело с *Д*искурсивными указаниями на причинность. Формальное представление (2) дано в (3):

(3) $(?(\alpha,\beta,\lambda) \wedge [\text{Fall}(e_1,y)]K_\alpha \wedge [\text{Push}(e_2,x,y)]K_\beta) \rightarrow \text{Cause}_\text{D}\ (\beta,\alpha,\lambda)$

Затем дискурсивное отношение характеризуется его семантическим эффектом. Семантический эффект отношения Explanation – установление причинных связей между положениями вещей K_α и K_β, из чего вытекает также, что K_α непосредственно следует за K_β.

[3] На самом деле групп отношений семь, но четыре из них касаются особенностей диалогического текста.

(4) Explanation(α,β) $\Rightarrow$ (cause(e_β,e_α)
((cause(e_β,e_α) $\wedge$ event(e_α)) $\Rightarrow$ $e_\beta \prec e_\alpha$

Таким образом, определение состоит из двух частей: семантических и прагматических характеристик контекста и семантического эффекта отношения. То же можно сказать и про структурные виды отношений, к которым принадлежит Contrast. К ограничениям, которые оно накладывает на контекст, относятся изоморфизм синтаксической структуры и контрастная тема (которая может быть выражена, например, предикатами противоположной полярности или семантики; см. [Asher et Lascarides 2003]):

(5) (?(α,β,λ) $\wedge$ Structurally_similar(α,β) $\wedge$ Contrasting_themes(α,β)) > Contrast(α,β,λ)

Что касается семантического эффекта отношения Contrast, то из него следует, что α и β имеют несовместимые импликатуры, т. е. пропозициональное содержание α позволяет вывести факт, отрицание которого может быть выведено из пропозиционального содержания β [Bras 2008: 47].

Очевидно, однако, что формальный аппарат SDRT слишком сложен, чтобы использовать его при аннотировании корпусов, а набор отношений – слишком невелик: их всего 12. В работе [Reese et al. 2007: 8], которая является по сути пособием по аннотированию, к ним добавлены еще два, а формализация, особенно пропозиций, исчезает. Отношения разделяются на две группы, сочиняющие и подчиняющие, и определяются описательно. Ср.: “Explanation(α,β) holds when the main eventuality of β is understood as the cause of the eventuality in α. Explanation has temporal consequences, viz. that the eventuality described in β precedes (or overlaps) the eventuality described by α” [Reese et al. 2007: 12]. Ср. определения в аннотированном корпусе французских текстов ANNODIS [Muller et al. 2012], также использующем SDRT.

4 Пенсильванский корпус (PDTB)

PDTB дает классификацию более узкой группы дискурсивных отношений, а именно тех, которые потенциально могут быть выражены коннекторами, т.е. логико-семантических отношений (ЛСО). Они сгруппированы по степени семантической близости и разделены на четыре группы: Temporal, Contingency, Comparison, Expansion, в каждой из которых выделяются подтипы. Так, в рамках временных ЛСО выделяются две группы: синхронные и асинхронные, а в рамках последней следование и предшествование. В рамках групп Contingency и Comparison дальнейшая классификация различает ЛСО, действующие на пропозициональном уровне (Cause, Condition, Concession, Contrast) и действующие на уровне речевого акта (Pragmatic Cause, Pragmatic Condition, Pragmatic Contrast). Однако только первая группа задана на основе четкого семантического критерия, распространяющегося на всю совокупность входящих в ее состав ЛСО и позволяющего отличить эту группу ЛСО от других. Определения, данные для остальных групп, не позволяют этого сделать. Так, Contingency определяется как отношение, при котором “one argument provides the reason, explanation or justification for the situation described by other” [PDTB Research Group 2019: 19]. Под это определение явно не попадает подгруппа отношений Condition, включающая отношения, при которых “one argument presents a situation as unrealized (the antecedent), which (when realized) would lead to the situation described by the other argument (the consequent)” (с. 22).

В группу Comparison, основанную на сходстве или несходстве соединяемых ситуаций, включено, напротив, наряду с Contrast и Similarity, также Concession, в семантику которого принято включать, причинный компонент, а именно отрицание ожидаемой причинной связи между ситуациями. Разработчики PDTB (с. 24) даже подчеркивают, что именно этот компонент отличает Concession от Contrast, где этот компонент отсутствует. Место Concession – скорее в группе Contingency, где оно логично соседствовало бы с ЛСО Negative Condition, также задействующим отрицание импликативной связи между ситуациями.

Наибольшая непоследовательность наблюдается, однако, в группе Expansion, объединяющей отношения, “that expand the discourse and move its narrative or exposition forward”[4]; см. рис. 1.

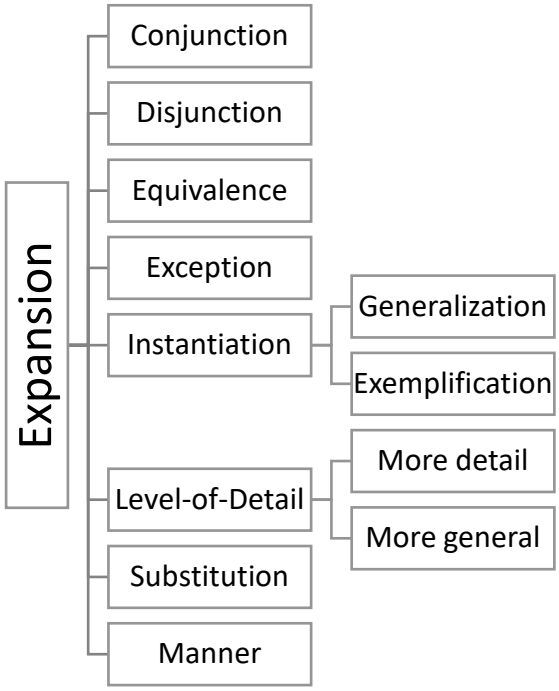


Рисунок 1. Группа ЛСО Expansion в PDTB [PDTB Research Group 2019].

Часть из отношений группы Expansion, как мы покажем в разделе 6, логично нашли бы свое место в группе Comparison: Conjunction, Disjunction, Substitution, Equivalence и подвид More general ЛСО Level-of-Detail. ЛСО Exception, подвид More detail ЛСО Level-of-Detail и обе разновидности Instantiation (Generalization, Exemplification) могут быть объединены в группу ЛСО, выражающих отношения элемента и множества или общего и частного. Кроме того, некоторые из этих ЛСО (например, Disjunction) могут быть установлены как на пропозициональном уровне, так и на уровне речевого акта, однако в классификации PDTB это никак не оговаривается, в отличие от ЛСО групп Contingency и Comparison, а Equivalence и подвид More general ЛСО Instantiation устанавливаются на метаязыковом уровне, так как касаются выбора языковых средств описания одной и той же ситуации, как и метаязыковая альтернатива и замещение по адекватности описания. Этот уровень не выделяется в PDTB. Таким образом, в группе Expansion остается только ЛСО Manner, добавленное, заметим, только в последней версии классификации (3.0).

5 Когнитивный подход к отношениям связности (CCR)

Когнитивный подход к отношениям связности, разработанный в работах Т. Сандерса (см., из последних [Sanders et al. 2018]), определяет отношения связности, которые являются логико-семантическими, через набор «когнитивно базовых» признаков:

- полярность (отрицательная vs. положительная), т. е. необходимо ли для интерпретации отношения привлекать отрицание одного из соединяемых положений вещей (к отрицательным относятся, например, уступительные отношения);
- базовая операция (аддитивная vs. причинная); причинной считается операция, основанная на импликации, соответственно, в группу причинных входят условные, причинные, целевые, следственные и уступительные отношения;

[4] Отметим в этой связи, что, например, в SDRT или в Лингвистической модели дискурса Л. Поланьи термин Narration объединяет временные отношения как раз на том основании, что они позволяют двигаться вперед сюжетной канве; ср. характеристику Narratives «giving a next episode of a story» в [Polanyi 1995: 17].

- источник связности (объективный vs. субъективный), иначе говоря, насколько говорящий «вовлечен» в устанавливаемое отношение; эта оппозиция соответствует оппозиции между семантическими (объективными) и прагматическими (субъективными) отношениями в других классификациях (см., например, выше PDTB);
- порядок сегментов (базовый vs. небазовый); этот признак применяется только к группе отношений, основанных на причинной операции, и позволяет различать следствие (с базовым порядком, где причина как хронологически, так и в линейной структуре высказывания, предшествует следствию) и причину (с небазовым порядком, где в линейной последовательности высказывания следствие предшествует причине).

Этих признаков, однако, явно недостаточно, чтобы различать отношения. Не спасает дела и введение дополнительных признаков (среди которых, например, спецификация, альтернатива, условность). Так, обе разновидности отношений Instantiation и Level-of-Detail[5] имеют одинаковый набор не только основных признаков (положительная полярность, аддитивная операция, объективный/субъективный источник связности, признак порядка не применяется), но и дополнительный (specifity ‘специфичность’) [Sanders 2018: 61].

6 Определение ЛСО в Надкорпусной базе данных коннекторов

В Надкорпусной базе данных коннекторов (НБД) используется классификация, в основу которой положены четыре базовые семантические операции, или механизма, на которые опирается то или иное ЛСО: импликация, расположение на шкале времени, сравнение, соотнесение частного и общего или элемента и множества. Классификация последовательно различает уровни, на которых может быть установлено ЛСО: пропозициональный уровень, уровень высказывания (иллокутивный), метаязыковой; подробнее см. [Инькова 2019]. Соответственно, каждое ЛСО может определяться на основе этих двух критериев, к которым добавляется еще один, характеризующий ЛСО, основанные на импликации и сравнения: полярность, т.е. устанавливается ли ЛСО непосредственно между *p* и *q* или же при его интерпретации должны быть учтены также их отрицательные корреляты ¬ *p* и ¬ *q*. Кроме того, учитываются семантические и прагматические характеристики контекста.

На основе этих параметров впоследствии были сформулированы структурированные определения ЛСО. В настоящем исследовании мы приводим несколько таких определений и примеры, их иллюстрирующие.

Соединительное ЛСО (6)

- операция сравнения, устанавливающая сходство p и q относительно некоторого «общего знаменателя»;
- пропозициональный уровень;
- *p* и *q* – положения вещей, не связанные никаким другим ЛСО[6].

(6) Но эта многопланность и полифоничность мистерии чисто формальная, *и* самое построение мистерии не позволяет содержательно развернуться множественности сознаний с их мирами. [М. М. Бахтин. Проблемы поэтики Достоевского (1963)]

[5] В версии 2.0 [PDTB Research Group 2008], которая использована в [Sanders et al. 2018], ЛСО сгруппированы несколько по-другому, чем в версии 3.0, и подвиду More detail ЛСО Level-of-Detail соответствует Specification, а подвид More general рассматривался вместе с Generalization.

[6] Т.н. «*и*-отношение» накладывает минимальные ограничения структурного и семантического характера и определяется, как правило, отрицательно. Единственное ограничение – наличие «общего знаменателя» (ср. Commun integrator, или Gemeinsame Einordungsinstanz, в [Lang 1977]).

Перифрастическое переформулирование [7] (7)	**Обобщающее переформулирование[8] (8)**
• операция сравнения, устанавливающая сходство *p* и *q*; • метаязыковой уровень; • *p* и *q* – описания одного и того же положения вещей; • *p* и *q* имеют одинаковый интенсионал.	• операция сравнения, устанавливающая сходство *p* и *q*; • метаязыковой уровень; • *p* и *q* – описания одного и того же положения вещей; • *q* имеет более бедный интенсионал (т.е. является более общим описанием).

(7) Греция неплатежеспособна, следовательно, возникает прагматический вопрос: как сократить нарастающий госдолг и повернуть вспять тенденцию сокращения объема производства? *Другими словами*: как увеличить выпуск и сократить долг? [Г. Колодко. Как избежать африканизации Греции // «Эксперт», 2015]

(8) он часто брал меня на переговоры в качестве переводчика, и тут уж приходилось соответствовать и тону беседы, и её скорости, и свободе неадекватной фразеологии (особенно со стороны отечественного партнёра) и некоторой специфики самого разговора. *Короче говоря*, приходилось выкручиваться, и делала я это весьма ловко. [Невеста // «Туризм и образование», 2000.06.15]

Пропозициональная альтернатива (9)	**Иллокутивная альтернатива[9] (10)**	**Метаязыковая альтернатива[10] (11)**
• операция сравнения, устанавливающая несходство *p* и *q*; • пропозициональный уровень; • *p* и *q* – положения вещей, имеющие статус гипотезы; • говорящий предлагает сделать выбор между *p* и *q*.	• операция сравнения, устанавливающая несходство *p* и *q*; • уровень высказывания; • *P* – речевой акт, *q* – положение вещей, имеющее статус гипотезы и ставящее под сомнение обоснованность *P*; • говорящий предлагает сделать выбор между *P* и *q*.	• операция сравнения, устанавливающая несходство *p* и *q*; • метаязыковой уровень; • *p* и *q* – возможные описания одного и того же положения вещей *r*; • говорящий предлагает сделать выбор между *p* и *q*.

(9) Квартирохозяйку тоже можно заинтересовать. *Или* запугать. [А. и Г. Вайнеры. Эра милосердия (1975)]

(10) – Моя мамаша в Петракове. Гуслинский даже сел. – А как же… а эта дама? Мы же ее называли мамашей, *или* я преспокойно сошел с ума! А? [Н. А. Тэффи. Тонкая психология (1911)]

(11) Она есть его тень *или*, по Гегелю, диалектический антитезис. [С. Булгаков. У стен Херсониса (1922)]

[7] Equivalence в PDTB.
[8] Подвид More general ЛСО Level-of-Detail в PDTB.
[9] Отсутствует в PDTB.
[10] Отсутствует в PDTB.

Пропозициональное замещение (12)	**Замещение по дескриптивной адекватности[11] (13)**
• операция сравнения, устанавливающая несходство *p* и *q*; • пропозициональный уровень; • *p* – положение вещей, осуществление которого можно было бы ожидать; • *q* – положение вещей, не соответствующее ожиданиям; • *p* отвергается, принимается *q*.	• операция сравнения, устанавливающая несходство *p* и *q*; • метаязыковой уровень; • *p* – описание положения вещей *r*, которое можно было бы ожидать; • *q* – описание того же положения вещей *r*, не соответствующее ожиданиям; • *p* отвергается, принимается *q*.

(12) Ты тоже собираешься, *вместо того чтобы* готовить обеды, сидеть за компьютером? [Елена Павлова. Вместе мы эту пропасть одолеем! // «Даша», 2004]

(13) Из голубого пластика и стекла, сверкая обтекаемыми изгибами, буфет напоминал по своим очертаниям *скорее* летательный аппарат, *чем* торговую точку. [Фазиль Искандер. Летним днем (1969)]

Исключение (14)	**Исключение из рассмотрения[12] (15)**
• операция соотнесения элемента и множества; • пропозициональный уровень; • X содержит указание на множество *P*; • Y содержит указание на элемент *q*, исключаемый из этого множества.	• операция соотнесения элемента и множества; • уровень высказывания; • X – информативный речевой акт, описывающий положение вещей *p*; • Y содержит указание на элемент *q*, который надо исключить, чтобы признать истинным *p*.

(14) В первом классе любовь к учительнице ни к чему не привела, *кроме того, что* из-за нее я очень больно ударился головой о бревно. [С. Ткачева. День влюбленных... // «100% здоровья», 2003.01.15]

(15) Едва ли кто-нибудь, *кроме* матери, заметил появление его на свет. [И. А. Гончаров. Обломов (1859)]

Экстенсиональная генерализация (16)	**Интенсиональная генерализация[13] (17)**
• операция соотнесения общего и частного; • пропозициональный уровень; • *p* – положение(я) вещей, имеющее(ие) место в некоторых обстоятельствах; • *q* – положение вещей, включающее *p* во множество $\{p_1, p_2 ... p_n\}$, а значит имеющее более широкий экстенсионал.	• операция соотнесения общего и частного; • метаязыковой уровень; • *p* – положение вещей, имеющее место в некоторых обстоятельствах; • *q* – обобщенное («без частностей») представление положение вещей, сделанное на основании свойств *p* и имеющее более широкий экстенсионал.

[11] Отсутствует в PDTB.
[12] Отсутствует в PDTB; подробнее см. [Инькова, Манзотти 2019].
[13] Отсутствует в PDTB.

(16) Никогда я не совал своего носа в литературу и в политику, не искал популярности в полемике с невеждами, не читал речей ни на обедах, ни на могилах своих товарищей... *Вообще* на моем ученом имени нет ни одного пятна. [А. П. Чехов. Скучная история (1889)]

(17) Тогда шедший впереди откровенно вынул из-под пальто черный маузер, а другой, рядом с ним, отмычки. *Вообще*, шедшие в квартиру No 50 были снаряжены как следует. [М. А. Булгаков. Мастер и Маргарита (1929-1940)]

Экстенсиональная спецификация[14] (18)	**Интенсиональная спецификация[15] (19)**
• операции соотнесения общего и частного; • пропозициональный уровень; • X содержит обобщенное понятие или положение вещей *p*; • Y содержит более частное понятие *q*, сужающее экстенсионал *p*.	• операция соотнесения общего и частного; • метаязыковой уровень; • X содержит обобщенное понятие или положение вещей *p*; • Y содержит более частное понятие *q*, сужающее интенсионал *p*.

(18) Да и зачем оно, это дикое и грандиозное? Море, *например*? [И. А. Гончаров. Обломов (1848-1859)]

(19) Подарил Марусе ценный, уникальный сувенир. *А именно* – конспиративную записку диссидента Шафаревича, написанную собственной рукой. [С. Довлатов. Иностранка (1986)]

7 Заключительные замечания

Определения, используемые в НБД, имеют, как мы видим, теоретическое значение, обладая объяснительной силой. Прежде всего, они позволяют увидеть семантическую близость некоторых ЛСО, а также тот факт, что разные ЛСО, но обладающие схожими свойствами, могут использоваться в схожих семантических структурах. Например, показатели метаязыковой альтернативы и перифрастического переформулирования могут взаимозаменяться, особенно в рамках простого предложения(20), без значительного изменения смысла, поскольку оба ЛСО установлены на метаязыковом уровне и уравнивают интенсионалы *p* и *q*. А различия в семантике этих двух ЛСО позволяют их показателям сочетаться друг с другом не создавая тавтологии(21).

(20) Можно утверждать, что сообщество есть письмо общества, *другими словами* / *или* – его «différence». [Е. В. Петровская. Безымянные сообщества (2010)]

(21) Врач общей практики, *или, другими словами*, семейный врач, – это специалист с высшим медицинским образованием, имеющий юридическое право оказывать первичную многопрофильную медико-социальную помощь независимо от возраста и пола пациентов. [В. Шпикалов. В здравоохранении эксперименты недопустимы // «Восточно-Сибирская правда» (Иркутск), 2003.06.21]

Эта семантическая близость ЛСО проявляется и при сопоставительном анализе. Переводчики выбирают иногда показатель другого, но имеющего сходные семантические параметры ЛСО, чтобы передать семантику коннектора; см. в этом отношении перевод показателя экстенсиональной генерализации *вообще* в (16) показателем ЛСО обобщающего переформулирования *bref* ‘короче говоря’ в (22):

[14] Подвид Exemplification ЛСО Instantiation в PDTB.
[15] Подвид More detail ЛСО Level-of-Detail в PDTB.

(22) Je n'ai jamais fourré le nez dans la littérature ni la politique, recherché la popularité en polémiquant contre des ignorants, prononcé de discours dans des dîners ou sur la tombe de mes confrères… *Bref*, mon nom de savant est sans tache [Tr. É. Parayre]

В определениях этих двух ЛСО есть общие элементы: оба они сигнализируют об обобщении, т.е. переходе к более широкому экстенсионалу.

В настоящее время ведется работа по интегрированию этих определений в НБД, что позволит автоматически исчислять семантическую близость ЛСО на основе наличия у них общих семантических признаков и их количества. Так, наличие общих признаков у ЛСО метаязыковой альтернативы и перифрастического переформулирования (+метаязыковой уровень, +тождество интенсионала) позволяет говорить об их большей семантической близости, чем, например, для ЛСО перифрастического переформулирования и интенсиональной спецификации, имеющих лишь один общий признак (+метаязыковой уровень).

Определения, используемые в НБД, позволяют, как нам кажется, определить на основе четких семантических критериев то или иное ЛСО, а значит преодолеть непоследовательность существующих классификаций и повысить уровень согласованности в работе экспертов и разметчиков, работающих в разных подходах и с разными языками, служа своего рода интерлингвой (о необходимости разработки такой интерлингвы см. [Sanders et al. 2018]). Это повышает и ценность аннотированного в разных подходах материала, который приобретает таким образом свойство reliability.

Литература

[1] Alonso L., Castellon I., Padro L. (2002) Lexicón computacional de marcadores del discurso. Procesamiento del Lenguage Natural. 2002. 29. P. 239–246.

[2] Asher N. (1993) Reference to Abstract Objects in Discourse. Kluwer Academic Publishers: Dordrecht.

[3] Asher N. (1996) L'interface pragmatique-sémantique et l'interprétation du discours [Pragmatic-semantic interface and discourse interpretation]. Langages. 1996. 123. P. 30–50.

[4] Asher N., Lascarides A. (2003) Logics of Conversation. Cambridge: Cambridge University Press.

[5] Benamara F., Taboada M. (2015) Mapping different rhetorical relation annotations: A proposal. Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (SEM 2015). Denver: ACL. P. 147–152.

[6] Bras M. (2008) Entre relations temporelles et relation de discours. PhD Thesis. Toulouse University; ULR: http://myriam.bras.free.fr/publis/HdR_Myriam.pdf.

[7] Bunt H., Prasad R. (2016) ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. Proceedings of the LREC 2016 Workshop "ISA-12: 12th Joint ACL – ISO Workshop on Interoperable Semantic Annotation". Bunt H. (ed.). Slovenia, Portorož, 2016. P. 45–54.

[8] Carlson L., Marcu D. (2001) Discourse Tagging Reference Manual ISI Technical Report ISI-TR-545 54, 56; URL: https://www.isi.edu/~marcu/discourse/tagging- ref- manual.pdf.

[9] Chiarcos C. (2014) Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Calzolari N., Choukri Kh., Declerck Th., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., Piperidis S. (eds.). Reykjavik: ELRA. P. 4569–4577.

[10] Demberg V., Scholman M. C. J., Asr F. T. (2019) How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. Dialogue & Discourse. 2019. 1. P. 87–135.

[11] Goncharov A. (2021) Classifications of intratextual relations: Bases and structuring principles [Klassifikatsiya vnutritekstovych otnosheniy: osnovaniya i principy strukturirovaniya]. Voprosy Jazykoznanija. 2021. 3. Pp. 97–119.

[12] Grosz B. J., Sidner C. L. (1986) Attention, intentions, and the structure of discourse. Computational Linguistics. 1986. 3. P. 175–204.

[13] Halliday M.A.K. (1985) An Introduction to Functional grammar. London: Edward Arnold.

[14] Inkova O. (2019) Logical-semantic relations: classification problems [Logiko-semanticheskie otnosheniya: problemy klassifikatsii], O. Inkova, E. Manzotti, Text coherence: mereological logical-semantic relations [Svyaznost' teksta: mereologicheskie logiko-semanticheskie otnosheniya]. Moscow: Izdatel'skii Dom YaSK. Pp. 11–98.

[15] Inkova O., Manzotti E. (2019) Text coherence: mereological logical-semantic relations [Svyaznost' teksta: mereologicheskie logiko-semanticheskie otnosheniya]. Moscow: Izdatel'skii Dom YaSK.

[16] Knott A. (1996) A Data-Driven Methodology for Motivating a Set of Coherence Relations. PhD Thesis. University of Edinburgh.

[17] Fraser B. (2009) An Account of Discourse Markers. International Review of Pragmatics. 2009. 1. P. 1–28.
[18] Lang E. (1977) Semantik der koordinativen Verknüpfung. Berlin (DDR): Akademie. (The semantics of coordination. Amsterdam: John Benjamins, 1984.)
[19] Mann W., Thompson S. (1988) Rhetorical structure theory: Towards a functional theory of text organization. Text. 1988. 8. P. 243–281.
[20] Muller P., Vergez M., Prevot L., Asher N., Benamara F., Bras M., Le Draoulec A., Vieu L. (2012) Manuel d'annotation en relations de discours du projet ANNODIS [Annotation manual in discourse relations of the ANNODIS project]. Carnets de grammaire, Rapport n°21. CLLE-ERSS. Toulouse: Université de Toulouse Jean Jaurès.
[21] Penn Discourse Treebank (PDTB) Project. University of Pennsylvania; URL: https://www.seas.upenn.edu/~pdtb/.
[22] PDTB Research Group (2008). The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania; URL: https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb- annotation- manual.pdf.
[23] PDTB Research Group (2019). The Penn Discourse Treebank 3.0 Annotation Manual, 2019; URL: https://doi.org/10.35111/qebf-gk47.
[24] Polanyi L. (1995) The Linguistic Structure of Discourse. Stanford: CSLI Publications.
[25] Prasad R., Webber B., Joshi A. (2017) The Penn Discourse Treebank: An Annotated Corpus of Discourse Relations. Handbook of Linguistic Annotation. Ide N., Pustejovsky J. (eds.). Dordrecht: Springer Science+Business Media. P. 1197–1217.
[26] Reese B., Hunter J., Asher N., Denis P., Baldridge J. (2007) Reference Manual for the Analysis and Annotation of Rhetorical Structure (version 1.0). Technical Report. Austin: University of Texas, Departments of Linguistics and Philosophy; URL: http://timeml.org/jamesp/annotation_manual.pdf.
[27] Roze C. (2013) Vers une algèbre des relations de discours [Towards a Discourse Relation Algebra], PhD Thesis, Paris Diderot – Sorbonne Paris Cité University.
[28] Sanders T., Demberg V., Hoek J., Scholman M., Asr F. T., Zufferey S., Evers-Vermeul J. (2018) Unifying Dimensions in Coherence Relations: How Various Annotation Frameworks are Related. Corpus Linguistics and Linguistic Theory; DOI: doi.org/10.1515/cllt-2016-0078.
[29] Stede M., Umbach C. (1998) DiMLex: A lexicon of discourse markers for text generation and understanding. Proceedings of the Joint 36th Meeting of the ACL and the 17th Meeting of COLING. Montreal, August 10-14. P. 1238–1242.

Divergent translation of connectives in human and machine translations

Inkova O.
Institute of Informatics Problems, FRC CSC RAS / Moscow, Russia
University of Geneva, Geneva, Switzerland
Olga.Inkova@unige.ch

Nuriev V.
Institute of Informatics Problems, FRC CSC RAS / Moscow, Russia
nurieff.v@gmail.com

Abstract

The paper is focused on divergent ways of conveying discourse relations in translation. For data collection, we used the supracorpora database of connectives storing parallel texts from the Russian-French subcorpus of the Russian National Corpus. These data show what logical-semantic relations tend to be translated using divergent ways, i.e. other than connectives (exclusion in its various gradations, propositional concomitance and substitution, the share of divergent translations ranging from 30% to 50%). Also, such data help define what causes divergent ways of translation to be used. The causes may be as follows: (a) the lack of an adequate equivalent of a given connective in the target language; (b) differences in the syntactic structure of the source and target languages; (c) usage differences; (d) contextually determined use of divergent translation. If there is a prototypical indicator of logical-semantic relations (i.e. connective) in the source text, it also occurs in translation in more than 90% of cases. The data on human translations are then compared with those on machine translations, which shows that the machine translation system also tends to keep a connective if there is one in the source text (it occurs in almost 98% of cases). However, there are cases where the machine translation system has difficulties processing a multiword connective (failing to perceive it as a whole) or a polyfunctional unit (failing to tell a connective from a non-connective) and thus uses divergent ways to translate it. Some causes of divergently translating connectives are likely to be the same for human and machine translations. These are differences in the syntactic structure of languages and usage differences. Further research of divergent means of conveying discourse relations will allow to draw a sharper border-line between explicitly expressed and implicit discourse relations. The data collected from annotated corpora (both monolingual and multilingual and parallel) will help determine what the divergent ways of expressing logical-semantic relations are and how frequently they are used. The research results can be used both in automatic text processing and automatic text generation. Also, the data on divergent translations of discourse relations can serve to improve the machine translation quality.

Keywords: semantics; discourse relations; connective; divergent translation; parallel corpora; quantitative analysis; machine translation

DOI: 10.28995/2075-7182-2021-20-339-348

Дивергентный перевод коннекторов в авторских и машинных переводах

Инькова О.Ю.
ИПИ ФИЦ ИУ РАН / Москва, Россия
Женевский университет / Женева, Швейцария
Olga.Inkova@unige.ch

Нуриев В.А.
ИПИ ФИЦ ИУ РАН / Москва, Россия
nurieff.v@gmail.com

Ключевые слова: семантика; дискурсивные отношения; коннектор; дивергентный перевод; параллельные корпуса; количественный анализ; машинный перевод

1 Введение

Проблема дивергентных средств выражения логико-семантических отношений (ЛСО) встает особенно остро при аннотировании корпусов текстов с точки зрения их связности. Как известно, дискурсивные отношения могут выражаться как коннекторами, так и другими языковыми средствами, получившими название «альтернативные лексикализации» [Prasad et al. 2010]. В дальнейшем была предложена классификация языковых средств, способных выражать дискурсивные (в данной терминологии) отношения [Das, Taboada 2013]; [Taboada, Das 2013]; [Das, Taboada 2014]. Выделяются лексические, морфологические (временные формы), семантические (синонимия, антонимия и др.), синтаксические (различные виды придаточных предложений и др.) и графические средства, которые могут сочетаться.

(1) – Слава Богу, слава Богу, – заговорила она, – теперь всё готово. *Только* немножко вытянуть ноги. [Л. Н. Толстой. Анна Каренина (1873-1877)]
– Dieu merci, me voilà prête. Il *ne me reste qu*'à étendre un peu les jambes. [Tr. H. Mongault]

В (1) *только*, выступающее в функции коннектора, устанавливает между положениями вещей «теперь почти все готово» и «(нужно) немножко вытянуть ноги» ЛСО исключения (всё готово, за исключением того, что нужно вытянуть ноги). Французский переводчик, вместо коннектора исключения, например, *sauf que*, использует два языковых средства: грамматическое (ограничительное отрицание *ne... que*, часто выступающее эквивалентом *только*) и лексическое (глагол *rester* 'оставаться', т.е. буквально 'остается вытянуть ноги').

При работе с параллельными корпусами для таких случаев используется термин «дивергентный перевод», предложенный в [Johansson 2007] и уточненный для описания средств выражения дискурсивных отношений в [Инькова 2019а]. Дивергентным считается такой перевод, когда показатель ЛСО – коннектор – переведен языковым средством, не принадлежащим к этому функциональному классу, и, наоборот, когда такое языковое средство переведено коннектором[1]. Ср. (2), где в варианте а. выбрано конгруэнтное средство перевода коннектора *а*, выражающего сопоставительные ЛСО (французский коннектор *tandis que* 'тогда как'), а в варианте б. – дивергентное: местоименный повтор подлежащего, создающий сопоставление действий двух субъектов, Обломова и Алексеева.

(2) Обломов задумался, *а* Алексеев барабанил пальцами по столу, у которого сидел.
[И. А. Гончаров. Обломов (1848-1859)]
а. Oblomov se replongea dans ses rêveries, *tandis qu*'Alexeïev tambourinait sur la table devant laquelle il était assis. [Tr. A. Adamov]
б. Oblomov devint pensif, Alexéev, *lui*, pianotait sur la table. [Tr. L. Jurgenson]

В настоящем исследовании на основе информации, полученной в Надкорпусной базе данных коннекторов (НБД)[2], мы покажем, для каких ЛСО наиболее характерны дивергентные средства перевода и проанализируем причины их использования (раздел 2). Затем мы сравним данные, полученные для переводов, выполненных профессиональными переводчиками, с данными, полученными для машинного перевода (раздел 3). Особый интерес представляет тот факт, что в обоих случаях переводился один и тот же русский контекст.

[1] Уточним, что на основании этого определения случаи, когда коннектор переведен коннектором, выражающим другое ЛСО, считаются конгруэнтными переводами.

[2] Подробнее об устройстве и возможностях использования НБД для сопоставительного количественного анализа коннекторов см. [Inkova 2021]; представительный фрагмент НБД доступен на сайте http://a179.frccsc.ru/RFH41002/main.aspx.

2 Дивергентные средства выражения логико-семантических отношений в авторских переводах

Материалом для исследования послужили 11 252 двуязычные аннотации в направлении перевода русский-французский[3]. Из них аннотаций с пометой «Дивергентное межъязыковое соответствие» (Dvrg) – 903, т. е. 8,1%. Эта цифра показательна сама по себе: она говорит о том, что при наличии показателя ЛСО переводчики стараются его переводить соответствующим показателем ЛСО. Это обстоятельство важно и при анализе машинных переводов и оценки их качества.

В НБД все виды дивергентного перевода распределены по четырем кластерам: Прочие (лексические), Прочие (грамматические), Пунктуация, Отрицание. В кластер Прочие (лексические) входят, как говорит его название, лексические средства передачи ЛСО. Например, сравнительные ЛСО часто передаются предикатами, выражающими сравнение, как в (3), где сравнительный коннектор *так*||*как* заменен в переводе на предикат *être semblable à* ‘быть похожим’.

(3) Она пела так чисто, так правильно и вместе *так… так… как* поют все девицы, когда их просят спеть в обществе; без увлечения. [И. А. Гончаров. Обломов (1848-1859)]
Son chant, si pur, si juste, était à la fois si... *semblable au* chant de toutes les autres jeunes filles quand on les prie de chanter en société: sans aucune émotion. [Tr. L. Jurgenson]

В кластер Прочие (грамматические) входят синтаксические конструкции, глагольные формы деепричастий и причастий и др. Так, в (4) временное ЛСО, выражаемое *когда*||*то*, передано деепричастием настоящего времени, также устанавливающим одновременность действий, описанных деепричастием (*en regardant bien* ‘рассмотрев’) и предикатом *s’aperçut* ‘заметили’.

(4) Но *когда* разглядели хорошенько Катерину Ивановну, *то* увидали, что она вовсе не разбилась о камень. [Ф. М. Достоевский. Преступление и наказание (1866)]
Mais *en regardant bien* Catherine Ivanovna, on s’aperçut qu’elle ne s’était nullement blessée contre une pierre. [Tr. É. Guertik]

В кластер Пунктуация входят семантически насыщенные знаки препинания: тире и двоеточие. Они совместимы с ограниченным набором ЛСО, а значит могут считаться средством их выражения. Двоеточие, например, часто передает причинные отношения, как в русском, так и во французском языках, и его переводчик использует вместо коннектора *потому что* в (5):

(5) А вы, матушка, и времени даром не теряйте, закажите ему теперь же сосновый гроб, *потому что* дубовый будет для него дорог. [Н. В. Гоголь. Шинель (1842)]
Allons, ma bonne dame, ne perdez pas votre temps inutilement; allez vite commander un cercueil de sapin*:* le chêne serait trop cher pour lui. [Tr. H. Mongault]

В кластер Отрицание изначально заносились коннекторы, которые включают в свой состав отрицание (*не то чтобы*||*а, не только*||*но и* и др.), а затем, по мере наполнения НБД и фиксации таких случаев, также отрицательные частицы (*ne*|*pas*, *ne*|*que*) или языковые единицы, в состав которых они входят (*ne pas empêcher* ‘не мешать’, *ne*|*que*_gérondif présent ‘*не*_деепричастие настоящего времени’). Дивергентными средствами перевода ЛСО, входящими в кластер Отрицание, считаются только последние. При этом кластеры Прочие (лексические), Прочие (грамматические), с одной стороны, и Отрицание, с другой, являются пересекающимися, т.е. языковые единицы, приписанные к двум кластерам Прочие, могут быть одновременно приписаны и к кластеру Отрицание. Например, *je n’irais pas jusqu’à dire* ‘букв. я не дойду до того, чтобы утверждать’, передающее ЛСО замещения, выраженное в оригинале коннектором *не то чтобы*, приписано к кластерам Прочие (лексические) и Отрицание.

(6) *Не то чтобы* меня выбросили из ресторана. Я выполз сам, окутанный драпировочной тканью. [Сергей Довлатов. Заповедник (1983)]

[3] Это направление перевода выбрано как располагающее наиболее представительным массивом аннотаций в НБД, но описанный выше подход к дивергентному переводу может быть применен и для анализа большего количества языков перевода. См. примеры такого анализа в [Кобозева, Инькова 2018] и [Инькова 2018].

Je n'irais pas jusqu'à dire qu'on m'a jeté à la porte du restaurant. J'ai rampé dehors, drapé dans le store. [Tr. Ch. Zeytounian-Beloüs]

В таблице 1 приводятся данные о распределении дивергентных средств перевода ЛСО по четырем кластерам. Они показывают, что более 55% приходится на лексические средства.

Кластер	**Всего МЭ**	**Для данного кластера**	**В %**
Прочие (грамм.)	903	219	24,2%
Прочие (лекс.)	903	520	57,6%
Пунктуация	903	28	3,1%
Отрицание	903	136	15%

Таблица 1: Распределение дивергентных переводов коннекторов по кластерам

Интересен тот факт, что некоторые дивергентные средства перевода ЛСО являются специализированными, т.е. выражают только одно ЛСО, другие, в первую очередь, грамматические, обслуживают сразу несколько ЛСО. Так, наибольшим разнообразием передаваемых ЛСО отличается определительное придаточное, которое использовано как функциональный эквивалент показателей 15 ЛСО[4]: аддитивные иллокутивные (9 аннотаций); переформулирование и спецификация (по 5 аннотаций); сопоставительные, «вопреки ожидаемому», пропозициональные причинные, сравнительные, пропозициональное сопутствование (по 3 аннотации); условные, уступительные, иллокутивное сопутствование и иллокутивные причинные (по 2 аннотации); аддитивные пропозициональные, соединительные и противопоставление (единичные употребления). Ср. (7), где оба переводчика передают ЛСО, выражаемое *потому что*, относительным придаточным, в варианте а. в сочетании с местоименным повтором.

(7) Коллежский асессор был в этом сведущ *потому, что* был посылан несколько раз на следствие еще в Кавказской области. [Н. В. Гоголь. Нос (1832-1833)]
а. L'assesseur de collège était savant en la matière, *lui qui avait été chargé d'instruire maintes affaires criminelles lors de son séjour au Caucase*. [Tr. B. De Schloezer]
б. L'assesseur, *qui avait procédé au Caucase à plus d'une enquête*, s'entendait en ces matières. [Tr. H. Mongault]

Далее по степени полисемичности следуют деепричастия настоящего времени, которые, в том числе в сочетании с местоимением *tout* 'всё', подчеркивающим одновременность действий, могут быть использованы для выражения 11 ЛСО: пропозициональное сопутствование (16 аннотаций); временные (11 аннотаций); соединительные (10 аннотаций); уступительные (8 аннотаций); «вопреки ожидаемому» (7 аннотаций); иллокутивное сопутствование (4 аннотации); условные и сопоставительные (по 2 аннотации); аддитивные пропозициональные и иллокутивные, противительно-уступительные (единичные употребления). Ср. (4) для временных отношений.

Среди лексических дивергентных средств в наибольшем спектре ЛСО задействовано наречие *même* 'даже', привносящее градационный оттенок. Оно зафиксировано для передачи 9 ЛСО: неединственности (5 аннотаций); противительно-уступительные иллокутивные, уступительные, аддитивные пропозициональные (по 2 аннотации); аддитивные иллокутивные, аналогия, коррекция, пропозициональная альтернатива, спецификация (единичные употребления). Ср. (8) для противительно-уступительных иллокутивных ЛСО в обоих вариантах перевода.

(8) Предположение, *хотя* легкое, шуточное, что она может быть несчастлива, неожиданно вызвало ее на откровенность. [И. А. Гончаров. Обломов (1848-1859)]
a. La supposition, *même* faite à la légère et en plaisantant, qu'elle pouvait être malheureuse, la poussa soudain à faire des aveux. [Trad. L. Jurgenson]

[4] В НБД и в работе используется классификация, разработанная О.Ю. Иньковой [Инькова 2019b], исходящая из базовой семантической операции, лежащей в основе того или иного ЛСО и различающая семантические уровни, на которых они могут быть установлены: пропозициональный, уровень высказывания (иллокутивный) и метаязыковой.

б. Cette hypothèse, *même* formulée à la légère, par plaisanterie, qu'elle pût être malheureuse, la poussait à avoir confiance. [Trad. A. Adamov]

В таблице 2 мы приводим распределение дивергентных переводов по ЛСО. В ней представлены только те ЛСО, для которых в НБД сформировано более 50 аннотаций[5].

ЛСО	**Всего**	**Диверг.**	**%**
исключение	277	127	45,8%
исключение из рассмотрения	69	26	37,7%
пропозициональное сопутствование	126	42	33,3%
замещение	220	73	33,2%
аналогия	102	18	17,6%
тождество	65	9	13,8%
коррекция	277	31	11,2%
спецификация	616	59	9,6%
возместительное противопоставление	131	12	9,2%
пропозициональная причина	363	33	9,1%
переформулирование	777	69	8,9%
сравнительные	307	27	8,8%
сопоставительные	373	31	8,3%
аддитивные иллокутивные	519	43	8,3%
временные	644	53	8,2%
соединительные	344	27	7,8%
иллокутивная причина	221	16	7,2%
иллокутивное сопутствование	211	15	7,1%
следствие	57	4	7,0%
условные	623	38	6,1%
уступительные	608	37	6,1%
неединственности	398	24	6,0%
противительно-уступительные	74	4	5,4%
аддитивные пропозициональные	272	12	4,4%
экстенсиональная генерализация	103	4	3,9%
несоответствие	130	5	3,8%
"вопреки ожидаемому"	1373	46	3,3%
"вопреки ожидаемому" иллокутивные	543	19	3,3%
временные метаязыковые	126	4	3,2%
отрицательная альтернатива	190	5	2,6%
пропозициональная альтернатива	577	15	2,6%
уступительные иллокутивные	76	2	2,6%
контраст	54	1	1,8%

Таблица 2: Распределение дивергентных переводов коннекторов по ЛСО

[5] По сравнению с данными из [Инькова 2019а], увеличение количества аннотаций в НБД и более детальная классификация ЛСО позволяет дать более полную панораму ЛСО и более точное представление о возможностях их перевода дивергентными средствами.

Лидируют в этом списке ЛСО исключения (почти 50% дивергентных переводов), исключения из рассмотрения, пропозициональное сопутствование и замещение, для которых доля дивергентных переводов превышает 33%. На противоположном полюсе находится ЛСО контраста с менее чем 2% дивергентных переводов. Для основной массы ЛСО доля дивергентных переводов находится в зоне 2-10%.

Поскольку, как мы видели, при наличии показателя ЛСО переводчики стараются его переводить соответствующим показателем ЛСО, то важно проанализировать причины, по которым переводчик выбирает дивергентный способ передачи ЛСО. Анализ данных позволил выявить следующие причины:

- *Отсутствие точного эквивалента коннектора в языке перевода.* Это наиболее очевидная причина выбора дивергентного перевода. Здесь, однако, следует различать два случая.
  - В обоих языках есть дивергентные средства выражения данного ЛСО, в языке перевода дивергентное средство является единственно возможными. Это случай ЛСО пропозиционального сопутствования: в обоих языках оно может быть выражено деепричастием настоящего времени, но в русском для него есть также коннектор *при этом*.
  - В языке перевода нет коннектора, являющегося точным эквивалентом, и используются близкие по семантике коннекторы или дивергентные средства; ср. в (2) варианты перевода а. и б.
- *Различия в синтаксической структуре сопоставляемых языков*. Этот случай касается, прежде всего, отношения исключения, выражаемого *кроме*. В том случае, когда множество, из которого исключается вводимый им элемент, задано отрицательным местоимением, во французском языке вместо предлога *sauf* ‘кроме’ используется семантически пустой союз *que* ‘что’.

(9) Они все против, а мне *никто*, *кроме* тебя, не нужен. [Светлана Алексиевич. Время секонд хэнд]
Ils sont contre. Mais moi, je n’ai besoin de *personne d’autre que* toi! [Tr. S. Benech]

Различиями в синтаксической структуре языков объясняется и высокая доля дивергентных средств при переводе ЛСО аналогии (17,6%), поскольку соотносительные конструкции, при помощи которых оно выражается, как и сравнительные ЛСО, менее частотны во французском языке и в целом представляют значительные трудности для перевода; подробнее см. [Inkova 2014]. Ср.(10), где при переводе сравнительных отношений на французский язык невозможно использовать коррелятивную структуру.

(10) она делает это с такой обидной снисходительностью, *так* тихо, *как* делают только с детьми или с совершенными дураками. [И. А. Гончаров. Обломов (1848-1859)]
elle le faisait avec une condescendance si blessante, avec *cette douceur qu*’on a avec les enfants ou les idiots complets. [Tr. L. Jurgenson]

- *Различия в структуре коннекторов.* Появление дивергентного средства перевода может быть связано с тем, что переведена одна из частей двухместного коннектора, как правило, первая и содержащая отрицание, или один из элементов многоэлементного коннектора, как правило, конкретизатор, в терминологии русской грамматики, т.е. семантически более насыщенный компонент. Ср. (11), где семантика двухкомпонентного коннектора коррекции *не||а* передана при помощи отрицания (при наличии во французском языке двухкомпонентного коннектора *ne|pas||mais*).

(11) он никакой *не* интурист, *а* шпион. [М. А. Булгаков. Мастер и Маргарита (1929-1940)]
Ce *n’*est *pas* du tout un touriste. C’est un espion. [Tr. Cl. Ligny]

- *Различия в узусе*. К таким случаям относятся коннекторы замещения по предпочтению, которые существуют в обоих языках, но во французском языке семантика этого ЛСО чаще

передается глаголами предпочтения (*préférer, aimer mieux* ‘предпочитать’, безличная форма *valoir mieux* ‘стоит лучше’ и др.); см. (12).

(12) *Скорее* соглашусь умереть, – сказал я в бешенстве, – *нежели* уступить ее Швабрину! [А. С. Пушкин. Капитанская дочка]
– *J’aimerais mieux* mourir, dis-je avec fureur, *que de* la céder à Chvabrine. [Tr. L. Viardot]

- *Контекстуально обусловленное использование дивергентных средств*. Использование дивергентного перевода может быть связано с особенностями конкретного контекста, где по тем или иным причинам невозможно употребление коннектора, что часто приводит к значительному изменению по сравнению с текстом оригинала.

Подчеркнем, что во всех перечисленных случаях, кроме последнего, дивергентное средство перевода является одновременно и «альтернативной лексикализацией» ЛСО в языке перевода. Поэтому изучение этих явлений представляет интерес не только для сопоставительного изучения языков, но и для изучения состава языковых средств, которые могут использоваться для выражения ЛСО.

3 Дивергентные средства выражения ЛСО в машинных переводах

Материалом для исследования дивергентных переводов, выполненных машинным переводчиком[6], послужили 7631 двуязычная аннотация в направлении перевода русский-французский. Из них с пометой «Дивергентное межъязыковое соответствие» (Dvrg) – 180, т. е. 2,4%, что указывает на корреляцию с выявленной выше тенденцией, характерной для выполненных человеком переводов: наличие показателя ЛСО в исходном тексте порождает реализацию соответствующего показателя ЛСО и в переводе.

В НБД, как и в случае «человеческого» перевода, виды дивергентного машинного перевода распределены по четырем основным кластерам: Прочие (лексические), Прочие (грамматические), Пунктуация, Отрицание. Однако наряду с этими четырьмя классами для квалификации машинного перевода также использовался кластер Прочие без спецификации (лексические/грамматические). Сюда заносились переводные варианты, реализованные в контекстах, машинный перевод которых в целом содержит многочисленные ошибки, иногда не позволяя вообще установить какие-либо параллели с исходным фрагментом текста, см. (13), где *да еще*, показатель аддитивных пропозициональных ЛСО, передается утвердительным словом *oui* (‘да’).

(13) Из всех присутствующих я узнал только музыковеда Лазарева, *да еще* фарцовщика Белугу. [Сергей Довлатов. Заповедник (1983)]
– De toutes les personnes présentes, j’ai appris que musicologue Lazarev, *oui*, фарцовщика Белугу. [Яндекс.Переводчик (14.11.2016, 16:38)]

Кластер	**Всего МЭ**	**Для данного кластера**	**В %**
Прочие (грамм.)	180	49	27,2%
Прочие (лекс.)	180	58	32,2%
Прочие	180	68	37,8%
Пунктуация	180	0	0%
Отрицание	180	5	2,8%

Таблица 3: Распределение дивергентных машинных переводов коннекторов по кластерам

Данные, приведенные в таблице 3, свидетельствуют о том, что системы машинного перевода, в отличие от человека-переводчика, вообще не используют пунктуационные возможности

[6] В НБД хранятся аннотации машинных переводов, выполненных двумя системами – «Яндекс.Переводчик» (https://translate.yandex.ru/) и «Google Переводчик» (https://translate.google.com/).

дивергентного перевода[7]. По всей видимости, это связано, в первую очередь, со структурой тренировочных материалов, задействованных для обучения систем автоматического перевода, где количество таких переводных примеров (преимущественно заимствованных из художественных переводов) оказывается нерелевантным и не принимается в расчет.

ЛСО	**Всего**	**Дивергент.**	**%**
исключение из рассмотрения	12	6	50,00 %
переформулирование	267	52	19,48%
аналогия	419	29	6,92%
аддитивные иллокутивные	246	13	5,28%
причина	323	17	5,26%
пропозициональное сопутствование	124	6	4,84%
временные	532	21	3,95%
спецификация	274	8	2,92%
иллокутивное сопутствование	233	4	1,72%
замещение	161	2	1,24%
“вопреки ожидаемому”	517	6	1,16%
аддитивные пропозициональные	495	5	1,01%
сопоставительные	102	1	0,98%
экстенсиональная генерализация	288	2	0,69%
коррекция	773	4	0,52%
уступительные	387	2	0,52%
соединительные	239	1	0,42%
неединственности	1375	1	0,07%

Таблица 4: Распределение дивергентных машинных переводов коннекторов по ЛСО

В таблице 4 приводится распределение дивергентных машинных переводов по ЛСО. Панорама ЛСО здесь отличается от той, что представлена в таблице 2. Список ЛСО значительно меньше, и лидируют в нем ЛСО исключения из рассмотрения (доля дивергентных переводов 50%) и переформулирования (доля дивергентных переводов превышает 19%). На противоположном полюсе находится ЛСО неединственности с 0,07% дивергентных переводов. Для основной массы ЛСО доля дивергентных переводов находится в зоне 0,42-7%[8].

Анализ машинных переводов позволяет выявить три группы случаев, когда реализуется дивергентный вариант.

- Система машинного перевода не распознает неоднословный коннектор как единую языковую единицу.

(14) *Как* письмо прочел, *так и* пошел... [Ф. М. Достоевский. Преступление и наказание (1866)]
En lisant la lettre, je suis allé... [Google Translate (03.09.2019, 15:06)]

В (14) при переводе контекста, содержащего показатель ЛСО аналогии – коннектор *как*||*так и*, вводится герундиальный оборот *en lisant*. Вариант, зафиксированный в машинном переводе, показывает, что, во-первых, синкретичное смысловое содержание, о котором сигнализирует коннектор (аналогия + временные ЛСО), интерпретируется в пользу временного толкования, и, во-вторых, временные отношения передаются ошибочно (предшествование заменяется одновременностью). Система машинного перевода не воспринимает двухкомпонентный коннектор как

[7] Во всяком случае, пока таких переводных вариантов зафиксировано нами не было.
[8] Очевидно, что представленная в таблице 4 панорама ЛСО будет изменяться по мере дальнейшего наполнения НБД. При актуальной структуре данных наблюдается дисбаланс в соотношении аннотированных машинных переводов для контекстов, где зафиксированы показатели тех или иных ЛСО.

единый комплекс, каждая часть которого участвует в смыслообразовании, реагируя только на один компонент. Перевод ЛСО аналогии представляет определенные трудности и для автоматического переводчика в направлении русский-французский.

- Система машинного перевода не различает разные употребления полифункциональных языковых единиц (как в (13))[9];
- Случаи, когда использование дивергентных средств может объясняться так же, как и в переводе, выполненном человеком. Оно мотивировано различиями в синтаксической структуре языков и/или узуальными предпочтениями, которые отражаются в тренировочных данных, применяемых для обучения системы машинного перевода[10].

(15) *Так как* я знал, что заботы матушки о моих занятиях ограничатся этими немногими словами, то я и не почел нужным возражать ей… [И. С. Тургенев. Первая любовь (1860)]
Sachant que les préoccupations de ma mère au sujet de mes études se limiteraient à ces quelques mots, je n'ai pas jugé nécessaire de s'y opposer… [Google Translate (03.10.2019, 21:56)]

В (15) причинный коннектор *так как* переводится причастной конструкцией *sachant que*, что является вполне естественным для французского языка. В русском языке также возможно употребление деепричастия в данном контексте (ср. *Зная, что*…), вместе с тем при редактуре русского художественного текста, например, бытует установка по возможности избегать причастных и деепричастных оборотов.

4 Заключительные замечания

Дальнейшая разработка понятий «альтернативная лексикализация» и «дивергентные средства выражения дискурсивных отношений» позволит, на наш взгляд, провести более четкую границу между эксплицитно выраженными и имплицитными дискурсивными отношениями (см., например, [Martin 1992]; [Renkema 2004]; [Taboada 2009], где к имплицитным относятся все дискурсивные отношения, не выраженные коннекторами), а также между коннекторами, прототипическими показателями дискурсивных отношений, и другими средствами их выражения (ср. в этой связи довольно широкую трактовку понятия коннектор в [Toldova et al. 2018]). Данные, полученные в аннотированных корпусах, как одноязычных, так и многоязычных и параллельных, помогут, в свою очередь, составить более четкое представление о том, какие дискурсивные отношения могут быть выражены языковыми средствами, не принадлежащими к классу коннекторов, а какие – нет, каковы причины появления дивергентных средств выражения дискурсивных отношений в тексте и какова их частотность. Полученные результаты могут быть использованы как при автоматической обработке и генерации текста, так и – в первую очередь, данные о дивергентных переводах дискурсивных отношений – для улучшения качества машинного перевода. В отношении последнего было бы интересно на примере отдельно взятой системы машинного перевода проследить во времени, как изменяется доля дивергентных переводных вариантов по мере наращивания эпох обучения, проявляется ли тенденция к уменьшению этой доли или, наоборот, к ее увеличению.

References

[1] 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.
[2] 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.
[3] Inkova O. Yu. (2018) *Voobshche* [In general], O. Inkova (ed.), Semantics of connectives: a contrastive study [Semantika konnektorov: kontrastivnoe issledovanie]. Moscow: TORUS PRESS. Pp. 80–128.

[9] После внедрения технологий нейронного машинного перевода частотность таких случаев существенно снизилась.
[10] В статье зафиксированы результаты наблюдения за использованием дивергентных средств выражения ЛСО в машинных переводах, но при этом если и называются возможные причины реализации дивергентных вариантов, то только в качестве предположения. Как правило, на реализацию такого варианта в машинном (как статистическом, так и нейронном) переводе влияет не один, а целый ряд факторов.

[4] Inkova O. Yu. (2019a) Annotirovanie parallel'nykh tekstov: ponyatie "divergentnyi perevod" [Annotation of parallel texts: the concept of divergent translation]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". Moscow, May 29–June 1, 2019. http://www.dialog-21.ru/media/4600/inkovaoyu-019.pdf.
[5] Inkova O.Yu. (2019b) Logical-semantic relations: classification problems [Logiko-semanticheskie otnosheniya: problemy klassifikatsii], O. Inkova, E. Manzotti, Text coherence: mereological logical-semantic relations [Svyaznost' teksta: mereologicheskie logiko-semanticheskie otnosheniya]. Moscow: Izdatel'skii Dom YaSK. Pp. 11–98.
[6] Inkova O. (2021) La sémantique des connecteurs: méthodes quantitatives d'analyse [Semantics of connectives: quantitative methods of analysis]. Bern / Berlin: Peter Lang. 276 p.
[7] Johansson S. (2007) Seeing through multilingual corpora: On the use of corpora in contrastive studies. Amsterdam / Philadelphia: John Benjamins. 377 p.
[8] Kobozeva I.M., Inkova O.Yu. (2018) *Kak* i ego dvukhmestnye ekvivalenty [*Kak* 'As' and its two-word equivalents], O. Inkova (ed.), Semantics of connectives: a contrastive study [Semantika konnektorov: kontrastivnoe issledovanie]. Moscow: TORUS PRESS. Pp. 168–239.
[9] Martin J. R. (1992) English Text: System and Structure. Amsterdam / Philadelphia: John Benjamins. 620 p.
[10] Prasad R., Joshi A., Webber B. (2010) Realization of Discourse Relations by Other Means: Alternative Lexicalizations. Proceedings of the 23rd International Conference on Computational Linguistics (Beijing, China – August 23–27, 2010): Posters. Pp. 1023–1031.
[11] Renkema J. (2004) Introduction to Discourse Studies. Amsterdam / Philadelphia: John Benjamins. 363 p.
[12] Taboada M. (2009) Implicit and explicit coherence relations. Discourse, of Course. An overview of research in discourse studies. J. Renkema (ed.). Amsterdam / Philadelphia: John Benjamins. Pp. 127–140.
[13] 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.
[14] Toldova S., Pisarevskaya D., Kobozeva M., Vasilyeva M. (2018) The cues for rhetorical relations in Russian: "cause–effec"" relation in Russian rhetorical structure treebank. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". Moscow, May 30–June 2, 2018. http://www.dialog-21.ru/media/4338/toldovas.pdf.

The Relation of Categories of Concreteness and Specificity: Russian Data

Ivanov V.
Innopolis University / Innopolis, Russia
v.ivanov@innopolis.ru

Solovyev V.
Kazan Federal University / Kazan, Russia
maki.solovyev@maul.ru

Abstract

The categories of concreteness and specificity are important for understanding the mechanisms of information representation and processing in human brain. These two categories are quite close, but still different. A method for quantifying the degree of correlation of these categories for the English has recently been proposed. This paper deals with a similar research of the Russian. Ratings from the Concreteness/Abstractness Dictionary (RDCA) are taken as a measure of the words' concreteness. The degree of a word specificity is estimated by its location in the RuThes thesaurus. The paper represents the comparison with the English data and shows the similarity of the results for Russian and English.

Keywords: concreteness, specificity, thesaurus, RuThes, WordNet
DOI: 10.28995/2075-7182-2021-20-349-357

Соотношение категорий конкретности и специфичности – данные русского языка

Иванов В. В.
Университет Иннополис / Иннополис, Россия
v.ivanov@innopolis.ru

Соловьев В. Д.
Казанский федеральный университет / Казань, Россия
maki.solovyev@maul.ru

Аннотация

Категории конкретности и специфичности имеют важное значение для понимания механизмов представления и обработки информации в мозге человека. Эти две категории достаточно близки, но все же различаются. Недавно был предложен метод количественной оценки степени корреляции этих категорий для английского языка. В настоящей работе мы проводим аналогичное исследование для русского языка. В качестве меры конкретности слов берутся рейтинги из словаря конкретности/абстрактности (RDCA). Степень специфичности слова оценивается по его расположению в тезаурусе RuThes. Приведено сопоставление с данными для английского языка, показано, что результаты для русского языка схожи с результатами для английского.

Ключевые слова: конкретность, специфичность, тезаурусы, RuThes, WordNet

1 Introduction

The categories of abstractness/concreteness and specificity/genericity are the focus of cognitive research on the organization of information in human brain. Modern approaches to the study of concreteness/abstractness originate from the fundamental papers [21, 18].

There seems to be a correlation between these categories, and they are not always distinguished. Let us say that the concept "furniture" is more generic than the concept "sofa", and "furniture" is simultaneously more abstract than "sofa". The main goal of this paper is to find out to what extent these two

categories are correlated. The first study of such kind based on empirical material for the English was conducted in [4]. This paper shows that there is a correlation, but it is moderate − 0.361, according to Spearman. We set the same goals as those that were in the abovementioned work, but the research is done for the Russian, and, of course, the external linguistic resources, which have been used, are changed. We strive to reproduce the methodology of the study of the paper [4] as accurately as possible to ensure comparability of the results. In particular, we analyze only nouns and ignore word combinations. This limitation is also due to the fact that it is for nouns that the hierarchical relationships are described in the most detail.

It should be mentioned that the words “category” and “concept” will be used as synonymous, although there can be a difference between them as in Barsalou’s work [1]. Concreteness is usually defined in published papers as the ability to perceive members of this category through the senses [5]. A detailed discussion of abstractness is given in Barsalou’s work [2], while reviews of abstractness/concreteness studies can be found in the papers [7, 23].

It is generally assumed that abstractness/concreteness is not a binary but a continuous category, and the degree of concreteness of a concept is estimated by a number in a certain interval.

Surveys of respondents are conducted to assess the degree of abstractness/concreteness, resulting in a dictionary with ratings of abstractness/concreteness of words. The first four-thousand-word dictionary for the English was described in the article [8] and is available at https://websites.psychology.uwa.edu.au/school/mrcdatabase/uwa_mrc.htm. A dictionary of 40,000 words was created later on [6]. The RDCA (Russian Dictionary of Concreteness/Abstractness) dictionary for the Russian language for one thousand words was created at the Kazan Federal University [26].

The category of specificity/genericity reflects paradigmatical semantic relations between hyponym (as a subtype) and hyperonym (as a supertype); such relations underlying thesaurus lexicon classifications. The category of specificity/genericity is not binary. In the following sequence of concepts: "a doberman" - "a dog" - "a predator" - "a mammal" each next one is more general than the previous one. Our useage of terms is consistent with work [4], which explains the difference between Specificity и Concreteness: “… Specificity (which operationalizes the process of categorical abstraction) and Concreteness (which operationalizes the perceptibility of a referent associated with a concept)” (p. 368). An important contribution to its study was made by the classical works of Rosch [21]. After the creation of the WordNet thesaurus [11], the degree of specificity/genericity is usually evaluated by the place of the concept in the WordNet thesaurus [9]. The structure of WordNet and its relevance to linguistic facts is presented in [14]. The closer the concept represented by the synset (synonymous sets) of WordNet is to the upper levels of the thesaurus, the more generic it is. It can be automatically quantified. Three formulas for calculating the specificity/genericity measure are proposed and compared in [4]. The authors concluded that the most successful formula is: $(1 + d) / D$, where d is the total amount of hypernyms (direct and indirect) of a target word and D is the maximum distance from synset leaves to the top node. For WordNet, this value is 20. In this study, we use this approach with the replacement of WordNet with the freely available Russian thesaurus RuThes [16].

Generally speaking, the perception of such concepts as “concrete/abstract” and “specific/generic” should not depend on a language at least in close cultures of the modern globalized world. In our paper [25], we compared the concreteness ratings for Russian words and their equivalents in English and showed that they are mostly close, although there are significant differences, primarily related to the polysemy of words. Thus, the expected results are similar to the results for the English. However, in addition to differences in languages, the difference in the structure of the RuThes and WordNet thesauri can also influence the results.

Research objectives.

Q1. Do measures of concreteness and specificity correlate? In particular, will generic concepts be more abstract than specific ones?

Q2. Does the division of concepts into abstract entities and physical entities, presented in RuThes, correspond to the concreteness indices?

Q3. Which concepts have extreme values for the combination of concreteness and specificity parameters? To what extent are the WordNet and RuThes structures conformed in this aspect? If the results for English and Russian will differ considerably, what features of the structure of WordNet and RuThes can be responsible for this?

2 Data and methods

All dictionaries with human concreteness/abstractness ratings were created as follows. The respondents were asked to rate the degree of concreteness/abstractness of a word on a 5 or 7-point scale. In this work, we use the results of surveys on a 5-point scale. Recently, surveys have been conducted using crowdsourced platforms. For each word, at least 30 scores are obtained, which were averaged. At the same time, special measures are taken to screen out the ratings of unscrupulous respondents. The most frequent and/or well-known words are selected for rating. For more information on the methodology for creating dictionaries with human ratings, see [6]. Dictionaries for other properties of concepts have also been created, for example, Imagability [8].

The dictionary of the Russian language with the ratings of abstractness/concreteness, mentioned in the introduction, is insufficient for many studies. Therefore, the program was developed that extrapolates people's assessments of those words, which do not have ratings [24]. A machine dictionary containing 22,000 words of the Russian language is available at https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html. Its quality was evaluated, and a high level of correlation between machine and human ratings was shown (about 0.8 according to Spearman) [24]. This dictionary was used in the research.

RuThes Thesaurus (http://www.labinform.ru/pub/ruthes/index.htm) contains more than 31.5 thousand concepts, 111.5 thousands of different text inputs (words and expressions of the Russian language), more than 130,000 polysemantic words, taking into account their meanings. RuThes was created on the basis of a large body of news texts on socio-political issues as a resource for automatic text processing. The general structure of RuThes corresponds to the structure of WordNet – a set of concepts represented by synsets and connected by semantic relations. In this study, only one type of relationship between synsets is needed – hyponymy/hyperonymy.

The intersection of the set of nouns from RuThes and our dictionary with concreteness indexes contains 14,294 words. The study described in [4] covered 13,518 words of English.

To estimate the degree of synset specificity, we use the formula given in the introduction: $(1 + d)/D$, where d is the total amount of hypernyms (direct and indirect) of a target word and D is the maximum distance from synset leaves to the top node. For RuThes the value of D is equal to 13. The problem in calculating the degree of synset specificity is as follows. A synset can have several hyperonyms and, accordingly, several paths, possibly of different lengths, to the top of the hierarchy. Moreover, a word can have several meanings represented by different synsets. In [4], the first one in the WordNet list is selected from several options, which is usually the most frequent. Another possibility – averaging over all path lengths – is used in [13]. An overview of the various distance measures in WordNet is provided in [9]. We chose the second method for two reasons. First of all, the meanings of words in RuThes are arranged in random order, not ordered by frequency of use. The second argument is of a more theoretical nature. When determining the concreteness ratings, a word, but not particular meanings of the word, are represented to respondents. It is probable that the word has different meanings, some of them are concrete and others are abstract. This possibility and its relation to the metaphor are discussed in [20]. The respondents' responses, which may reflect different meanings of the word, are averaged when calculating the concreteness rating. Due to the fact that we still do not have ratings of the concreteness of particular meanings of words, we found it quite possible to calculate also the average ratings of specificity.

Both ratings – concreteness and specificity – are standardized and reduced to a 5-point scale, where 5 is the highest level of concreteness and specificity.

3 Results

3.1 Correlation coefficient

Determining the degree of correlation between specificity and specificity for the Russian and comparing it with the English is our first result. First of all, let us compare the distribution of specificity values in Russian and English.

Judging by the specificity histogram in Figure 2 from the article [4], WordNet shows a noticeable bias towards genericity of concepts and lower specificity. In particular, the median is M = 2.192, SD = 0.378. For RuThes, the corresponding values are M = 2.840, SD = 0.569. This difference can be explained as

follows: there are very long chains from the leaves to the top node – 20 nodes – in WordNet. However, not all chains, of course, are so long. So, for example, for a *vintage* leaf with a very specific value, the distance to the node 'Entity' is 6. When calculating the specificity formula, we get a fairly small number 7/20, which is more typical for abstract concepts; i.e., some specific concepts, according to this formula, are pulled up to the top of the hierarchy, shifting towards genericity.

This effect is reduced for RuThes by the fact that the maximum branch length is the denominator in the specificity formula = 13, not 20. As a result, the ratio of general and specific concepts in RuThes is more balanced than in WordNet, as can be seen in Figure 1 in comparison with Figure 2 from [4].

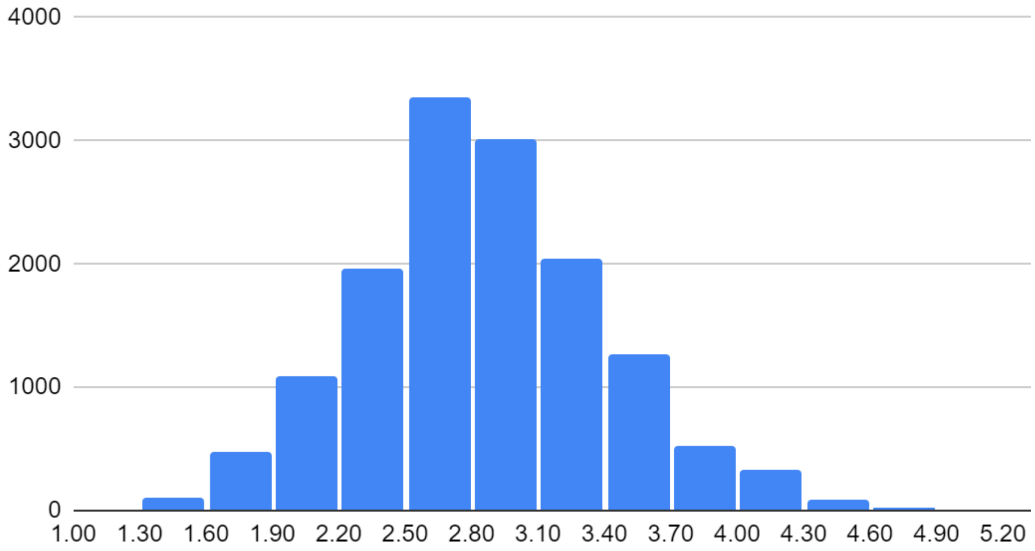


Fig. 1. Histogram of specificity ratings

The values of concreteness and specificity for all of the 14,294 words of the Russian language considered by us are given in the 'Concreteness Ratings in RuThes' file on the project website (https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html). Spearman's correlation coefficient = 0.264, Pearson's = 0.256 ($p<0.001$). For English, the coefficients are 0.361 and 0.354, respectively [4].

3.2 Abstract entities vs Physical entities

There are two high-level nodes (or concepts) - the PHYSICAL ENTITY and the ABSTRACT ENTITY in RuThes. Let us analyze the nodes of the thesaurus that are located below them. In the case of polysemy, when a word in one of the meanings is an abstract entity and a physical one in the other meaning, it is excluded from consideration. The average values and standard deviations are given in Table 1 and the histograms are shown in figures 2-5. For the t-test, the difference in the average values of both ratings for the ABSTRACT ENTITY and PHYSICAL ENTITY groups is statistically significant ($p<0.0001$).

	Concreteness	**Specificity**
All words under the labels ABSTRACT ENTITY or PHYSICAL ENTITY (n=9377)	M=3.570 SD=0.990	M=2.625 SD=0.653
Words under the label ABSTRACT ENTITY (n=2952)	M=2.553 SD=0.912	M=2.471 SD=0.626
Words under the label PHYSICAL ENTITY (n=6425)	M=4.037 SD= 0.595	M=2.734 SD=0.663
p-value of 1-tailed t-test	<0.0001	<0.0001
Cohen's d (effect size)	1.9267	0.3527

Table 1. Average values of the concreteness and specificity indices of the words under the label ABSTRACT ENTITY and PHYSICAL ENTITY

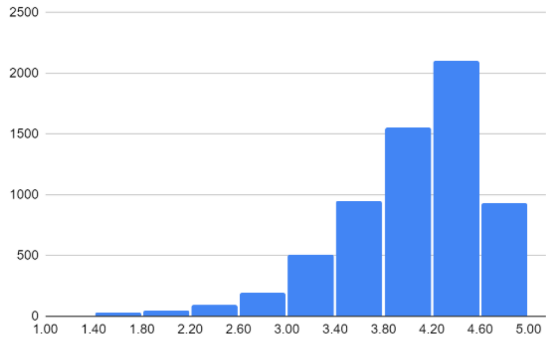


Figure 2. Histogram of distribution of words under the label PHYSICAL ENTITY by the concreteness index

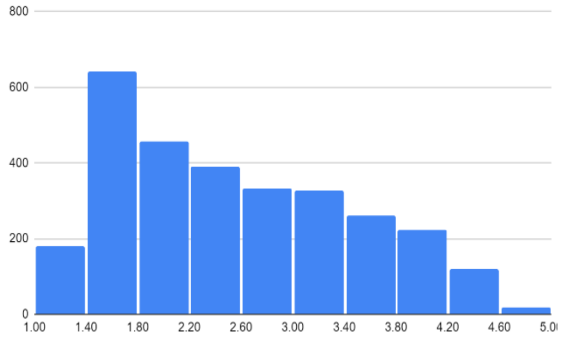


Figure 3. Histogram of distribution of words under the label ABSTRACT ENTITY by concreteness index

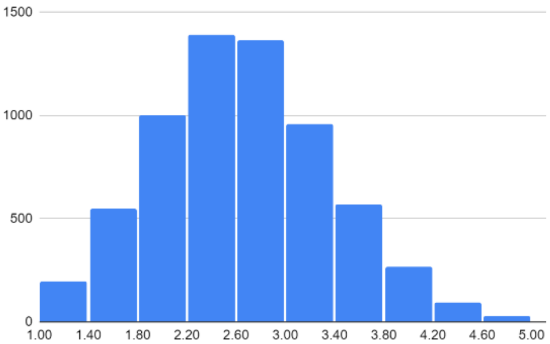


Figure 4. Histogram of distribution of words under the label PHYSICAL ENTITY by specificity index

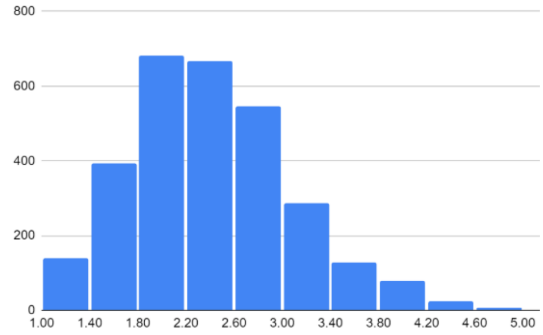


Figure 5. Histogram of distribution of words under the label ABSTRACT ENTITY by specificity index

The following patterns can be noted:

1. The average values of the concreteness and specificity indices are higher for words under the PHYSICAL ENTITY node than under the ABSTRACT ENTITY node (fig. 2 and 4 vs. fig. 3 and 5).
2. The average values of the concreteness and specificity indices for words under the ABSTRACT ENTITY node are close (fig. 3 and 5), although the average values of these indices differ significantly for words under the PHYSICAL ENTITY node (fig. 2 and 4).
3. Average values of indices of specificity of words under the nodes of the PHYSICAL ENTITY and of the ABSTRACT ENTITY of the Russian language is close to the same for English, respectively 4.037 vs 4.311 and 2.553 vs 2.754 [4].
4. Average values of indices of the specificity of the words under the nodes of the PHYSICAL ENTITY and of the ABSTRACT ENTITY for the Russian language is substantially higher than those for English, respectively 2.734 vs. 2.192 and 2.471 vs. 1.944 [4].

3.3 The distribution across the 4 quadrants

Let us consider how the distribution of concepts across the quadrants, representing combinations of the analyzed parameters, look like: highly specific and highly concrete, highly specific and highly abstract, highly generic and highly concrete, and highly generic and highly abstract, and compare it with distribution in the English.

The distribution of English words across the four quadrants, obtained by crossing the variables Specificity and Concreteness is shown in Figure 6.

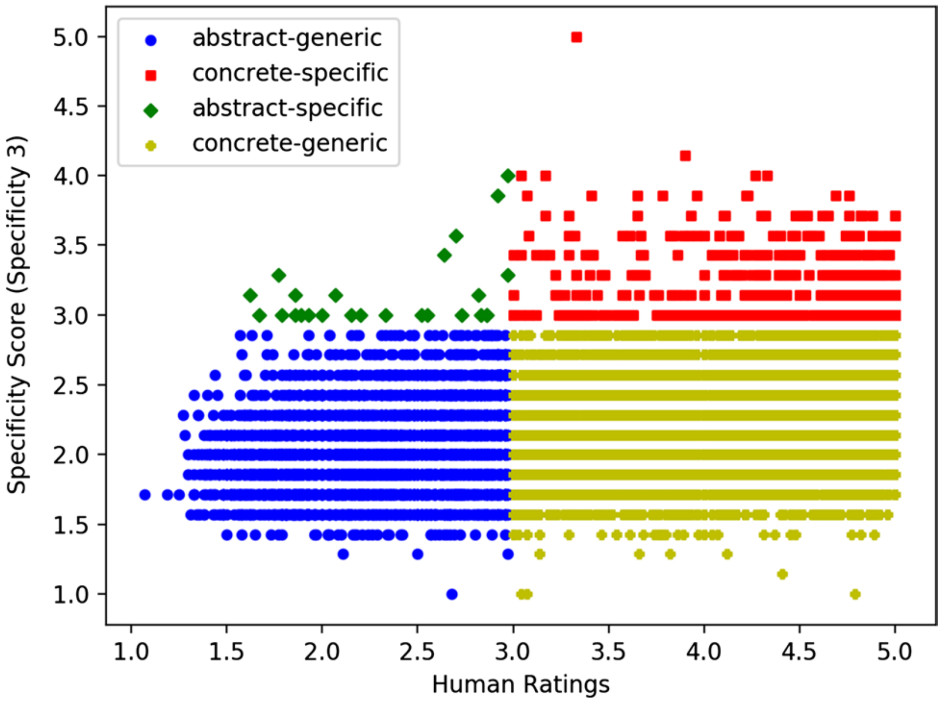


Fig. 6. The distribution of the English nouns across the four quadrants, obtained by crossing the variables Specificity and Concreteness [4], reproduced with permission of the authors.

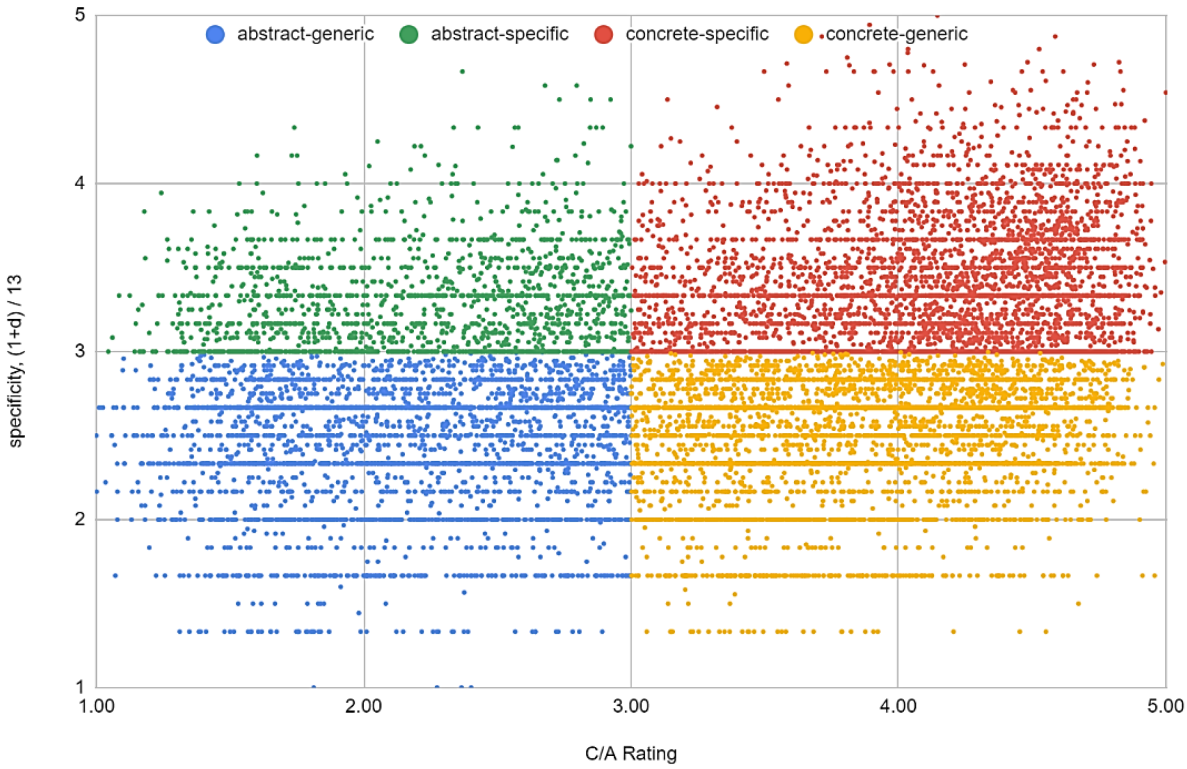


Fig. 7. The distribution of the Russian nouns across the four quadrants, obtained by crossing the variables Specificity and Concreteness

The distribution of Russian words across the four quadrants is shown in Figure 7. As we can see, there is a serious discrepancy in the configuration of the points on the rafts. For English, there is a large bias towards low specificity, i.e. high genericity. The probable reason for this was described in Section 3.1.

Let us see which words have the extreme values of the concreteness-specificity parameters.

Upper right quadrant: highly specific and highly concrete. Words in this sector were expected to denote typical concrete objects of the physical world that can be seen or touched. In English, the words *karaoke, epinephrine, aspirin, heifer, triglyceride, glucose, chloroform, fructose,* and *petroleum* have extreme meanings; in Russian – *травмпункт (emergency room), мегафон (megaphone), бомбардировщик (bomber), психбольница (mental hospital), радиотелефон (radiotelephone), мобильник (mobile phone), домофон (intercom), монитор (monitor), госпиталь (hospital, usually for military men), ноутбук (laptop), больница (hospital), горбольница (city hospital), медпункт (medical center), эвакуатор (tow truck), холера (cholera)*. As you can see, not all of these words refer to concrete material objects – *karaoke and холера (cholera)* do not apply to them. It seems interesting that in both cases, a significant proportion of the words refer to diseases/medicines/medical institutions. Apparently, in the minds of people, these are very concrete entities. The semantic field of communication is also distinguished in this quadrant in Russian.

Lower left quadrant: highly generic and highly abstract. Typical abstract entities, which are not perceived by the senses, were expected in this sector. The expectations were confirmed. In English they are: *absurdity, adaptability, ambiance, ambivalence, amorality, applicability, aptitude, authenticity, belief, circumstances, commitment, contradiction, desire, destiny,* and *idea*. In the Russian – *воздействие (effect, impact), непохожесть (otherness, dissimilarity), бремя (burden), пребывание (staying), претворение (implementation), различие (difference)*. The article [12] highlights the following 9 domains of abstract concepts: cognition, action, shapes, communication, relations, states, events, time, and motives. Most of the above words belong to domains relations or states. The domain cognition is also provided for English, but not for Russian.

Upper left quadrant: highly specific and highly abstract. In English this sector includes: *cakewalk, fundamentalism, and vintage, bootleg, finisher, general, mankind, monotheism, polytheism,* and *summons*. In [4], these words are characterized as referring to social reality. The following set of words is obtained for the Russian language: *идолопоклонство (idolatry), кощунство (blasphemy), поругание (desecration), святотатство (sacrilege), плодородие (fertility), сретение (Candlemas), роскошество (addiction to luxury or expensive venture), помрачение (obscuration, confusion) , заикание (stuttering, impediment in one's speech), роскошь (luxury), царствование (reign, kingship)*; most of them also relate to social reality. Moreover, a significant part of the words in both lists is related to issues of religion and faith.

Lower right quadrant: highly generic and highly concrete. For English, these words are: *ground, people, ribbon, seafood, ashes, breath, cloth, college, daytime, fabric, forest*. For Russian: *могильщик (gravedigger), зад (butt), бедро (hip), спина (back of the human body), снежинка (snowflake), подоконник (window sill), затылок (back of the head), бычок (young bull), задница (backside or ass), ягодица (buttock), подбородок (chin), фоторобот (identikit)*. In this sector, no connection between English and Russian words is found. As for the Russian, it is easy to see that most of the words refer to body parts.

4 Discussion and conclusion

Let us formulate what answers we can give to the questions announced at the beginning in accordance with the results of the study.

Q1. Do measures of concreteness and specificity correlate? In particular, will generic concepts be more abstract than specific ones?

The correlation coefficient established by us, although positive and statistically significant, is classified as weak [10]. Thus, the results of the work [4] are confirmed on the material of the Russian, which indicates the independence of the parameters of concreteness and specificity and the need to study them independently. Both generic and specific concepts can be abstract and concrete as well.

The difference between the categories of abstractness and specificity raises important questions in the field of cognitive science. A large number of cognitive and neurophysiological studies are devoted to the representation and processing of concrete/abstract concepts in human brain. There is the so-called

"concreteness effect", demonstrating greater ease of processing concrete words in the human mind [15]. Several theories have been proposed to explain the concreteness effect. The most developed and frequently cited are the following two theories: the dual-coding theory (DCT) [19] and the context-availability theory (CAT) [22]. A number of studies have found specific brain structures responsible for the representation of specific words [17]. It would be interesting to investigate whether these results are relevant to specificity.

Q2. Does the division of concepts into abstract entities and physical entities, presented in RuThes, correspond to the concreteness indices?

We have shown that the concreteness indices for words under the PHYSICAL ENTITY node are statistically much higher than for words under the ABSTRACT ENTITY node. Thus, the structure of RuThes corresponds well enough to native speakers' intuitive ideas of the degree of concreteness of words, expressed in their concreteness ratings.

Compared to the data for the English language, the average concreteness value has little differences, which indicates a good consistency of the concreteness ratings in these two languages. At the same time, the average specificity value is different. For the Russian language, it is larger, which confirms the above-mentioned difference in the structures of the WordNet and RuThes thesauri (also discussed below).

Q3. Which concepts have extreme values for the combination of concreteness and specificity parameters? To what extent are the WordNet and RuThes structures conformed in this aspect? If the results for English and Russian will differ considerably, what features of the structure of WordNet and RuThes can be responsible for this?

For the Russian, the distribution of words in the four octants (highly specific and highly concrete, highly specific and highly abstract, highly generic and highly concrete, and highly generic and highly abstract) is almost even, which also confirms the independence of these two parameters. Here we see significant differences between Russian and English, with a predominance of generic concepts in the latter. This is probably due to some differences in the structure of WordNet and RuThes – 1.5 times longer chains of hypo-hyperonymic relations in WordNet. We are going to conduct additional research by modifying the formula for calculating specificity so that we can eliminate this difference in the structure of thesauri. The difference in the algorithms for calculating the specificity index can also impact the results. In our algorithm, we took the average length of paths from the synset to the top concept and in [4] it was the length of the path with the most probable (frequency) values. It is possible that the more generic values are the most frequent in the case of polysemy (which is quite natural). It could also explain the shift in ratings towards genericity.

WordNet and RuThes are well aligned in another respect. Namely, the classes of words that simultaneously have extreme values of the parameters under consideration are largely similar. Prototypical abstract concepts were also among the most common in terms of their position in the hierarchies of both thesauri. For RuThes, this is, for example, *воздействие (effect, impact), непохожесть (otherness, dissimilarity) различие (difference)*. Prototypical concrete entities that can be seen and touched were also among the most specific. For example in Russian they are: *мобильник (mobile phone), ноутбук (laptop)*. In another quadrant – highly specific and highly abstract – words in both languages refer to social reality. And some of the words are related to questions of religion and faith. Taking into account the independent creation of WordNet and RuThes resources, as well as dictionaries with concreteness ratings, we can consider the degree of their consistency to be very high. There is no connection between the words of these thesauruses only in one quadrant.

The comparison of the categories of specificity and concreteness has important implications for cognitive science. In [4], it is suggested that specificity reflects the nature of the structuring of the World by language, while concreteness reflects the structuring of the World by consciousness for the construction of mental representations. The difference between these two categories (a low degree of correlation) is an argument against the strong version of the Sepir-Whorf hypothesis of linguistic relativity, which assumes that language determines thinking.

Acknowledgements

This research was financed by Russian Foundation for Basic Research, grant 19-07-00807.

References

[1] Barsalou L.W. (1983), Ad hoc categories. Mem Cognit 11(3):211–227.
[2] Barsalou L.W. (2003), Abstraction in perceptual symbol systems. Philos Trans R Soc Lond B Biol Sci 358(1435):1177–1187.
[3] Bolognesi M., Steen G. (eds.) (2019), Perspectives on abstract concepts: from cognitive processing to semantic representation. Benjamins Publishing Company, Amsterdam.
[4] Bolognesi M., Burgers Ch., Caselli T. (2020), On abstraction: decoupling conceptual concreteness and categorical specificity. Cognitive Processing. 21:365–381. https://doi.org/10.1007/s10339-020-00965-9.
[5] Borghi A.M., Binkofski F. (2014), Words as social tools: an embodied view on abstract concepts. Springer, New York.
[6] Brysbaert M., Warriner A.B., Kuperman V. (2014), Concreteness ratings for 40 thousand generally known English word lemmas. Behav Res Methods 46:904–911.
[7] Burgoon E., Henderson M., Markman A. (2013), There are many ways to see the forest for the trees: a tour guide for abstraction. Perspect Psychol Sci 8:501–520. https ://doi.org/10.1177/17456 9161349796 4.
[8] Coltheart M. (1981), The MRC Psycholinguistic Database, Quarterly Journal of Experimental Psychology, 33A, 497 – 505.
[9] Devitt A. and Vogel C. (2004), The Topology of WordNet: Some Metrics. GWC 2004, Proceedings, pp. 106-111.
[10] Evans J.D. (1996), Straightforward statistics for the behavioral sciences. Brooks/Cole Publishing, Pacific Grove.
[11] Fellbaum C. (ed.) (1998) WN: an electronic lexical database. MIT Press, Cambridge.
[12] Feng S., Cai Z., Crossley S.A., McNamara D.S. (2011), Simulating Human Ratings on Word Concreteness. In: FLAIRS Conference.
[13] Iliev R., Axelrod R. (2017), The paradox of abstraction: precision versus concreteness. J Psycholinguist Res 46(3):715–729.
[14] Miller G.A. (1998), Nouns in WordNet. In: Fellbaum C (ed.) Word-Net—an electronic lexical database. The MIT Press, Cambridge.
[15] Montefinese M. (2019), Semantic representation of abstract and concrete words: a minireview of neural evidence. J. Neurophysiol. 121, 1585–1587. doi: 10.1152/jn.00065.2019
[16] Loukachevitch N.V. (2011), Thesauri in information retrieval problems. M.: Publishing house of Moscow University.
[17] Orena E. F., Caldiroli D., Acerbi F., Barazzetta I., Papagno C. (2018), Investigating the functional neuroanatomy of concrete and abstract word processing through direct electric stimulation (DES) during awake surgery. Cognitive Neuropsychology. 36:3-4, 167–177. doi: 10.1080/02643294.2018.1477748.
[18] Paivio A. (1965), Abstractness, imagery, and meaningfulness in paired-associate learning. J. Ver-bal Learn. Verbal Behav. 4, 32–38. doi: 10.1016/s0022-5371(65)80064-0.
[19] Paivio A. (1990), Dual coding theory. In: Mental Representations: A Dual Coding Approach. Broadbent, D. E., McGaugh, J. L., Mackintosh, N. J., Posner, M. I., Tulving, E., Weiskrantz L. (eds). Oxford University Press, Oxford. 53–83. doi:10.1093/ac-prof:oso/ 9780195066661.003.0004.
[20] Reijnierse W.G., Burgers C.F., Bolognesi M., Krennmayr T. (2019), How polysemy affects concreteness ratings: the case of metaphor. Cogn Sci 31(8):e12779. https://doi.org/10.1111/cogs.12779.
[21] Rosch E. (1975), Cognitive representations of semantic categories. J Exp Psychol Gen 104(3):192–233.
[22] Schwanenflugel P. J., Shoben E. J. (1983), Differential context effects in the comprehension of abstract and concrete verbal materials. J. Exp. Psychol. Learn. Mem. Cogn. 9, 82–102. doi: 10.1037/0278-7393.9.1.82
[23] Solovyev V. (2021), Concreteness/Abstractness Concept: State of the Art. In: Velichkovsky B.M., Balaban P.M., Ushakov V.L. (eds) Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics. Intercognsci 2020. Advances in Intelligent Systems and Computing, vol 1358. Springer, Cham. pp. 275-283. https://doi.org/10.1007/978-3-030-71637-0_33.
[24] Solovyev V., Ivanov V. (2020), Automated Compilation of a Corpus-Based Dictionary and Computing Concreteness Ratings of Russian. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). - Vol. 12335 LNAI, - P.554-561.
[25] Solovyev V., Andreeva M., Solnyshkina M., Zamaletdinov R., Danilov A. and Gaynutdinova D. (2019), Computing Concreteness Ratings of Russian and English Most Frequent Words: Contrastive Approach. 12th International Conference on Developments in eSystems Engineering (DeSE), Kazan, Russia, pp. 403-408, doi: 10.1109/DeSE.2019.00081.
[26] Solovyev V. D., Ivanov V. V., Akhtiamov R. B. (2019), Dictionary of Abstract and Concrete Words of the Russian Language: A Methodology for Creation and Application. Journal of Research in Applied Linguistics. vol. 10, 215 -227.

Data pseudo-labeling while adapting BERT for multitask approaches

Karpov Dmitry
Moscow Institute of Physics
and Technology
Dolgoprudny, Russia
dimakarp19962008@yandex.ru

Burtsev Mikhail
Moscow Institute of Physics
and Technology
Dolgoprudny, Russia
burtcev.ms@mipt.ru

Abstract

Nowadays, BERT models have found wide use in the NLP field. However, standard BERT architecture training can be stifled by the lack of labels for different tasks while treating multitask settings as a one-task multilabel setting. For every example, we have labels from this example's source task but not from other tasks. This article addressed this issue, exploring eight different data pseudo-labeling approaches in the GLUE 4-task setting. These approaches do not require changes in samples or model architecture. One of the presented techniques excels results on RTE from the original article, by 6.2 %, and falls behind the original article on QQP, MNLI, and SST only by 0.5-1.2 %. This way also excels other pseudo-labeling approaches explored in the article by 0.5-2% on average if we consider similar tasks. However, for tasks that are dissimilar to each other, different proposed approach yields the best results.

Keywords: BERT, multitask, data augmentation, pseudo-labeling
DOI: 10.28995/2075-7182-2021-20-358-366

Псевдоразметка данных при адаптации архитектуры BERT для многозадачных подходов

Карпов Дмитрий
dimakarp19962008@yandex.ru

Бурцев Михаил
burtcev.ms@mipt.ru

Московский физико-технический институт
Долгопрудный, Россия

Аннотация

В настоящее время в обработке естественного языка широко используются модели типа BERT. Однако обучение стандартной архитектуры BERT при многозадачном подходе бывает затруднено недостатком меток для разных задач. В статье исследуются восемь различных способов псевдоразметки данных при обучении на нескольких задачах типа GLUE, не требующих изменений ни в наборе примеров, ни в архитектуре. В частности, представлен такой способ псевдоразметки данных для обучения оригинальной модели для решения четырех задач типа GLUE, который превосходит результаты из оригинальной статьи на датасете RTE на 6.2 % и отстает от нее на QQP, MNLI и SST только на 0.5-1.2 %. Способ превосходит другие методы псевдоразметки, рассмотренные в статье, в среднем на 0.5-2% на похожих задачах, но на разнородных задачах лучше работает другой из предложенных способов.

Ключевые слова: BERT, многозадачная модель, дополнение данных, псевдоразметка

1 Introduction

Transformer-based models, such as BERT, have found their wide use in the task of text classification. Conditions of learning such models are described in the original article [3]. These conditions suppose fitting every model for its task. In such a way, if we need to solve several classification tasks in parallel, we should keep several models for prediction, which increases the demand for computation power. This problem leads us to the idea of training one model that can yield the result for several tasks simultaneously. We explore the ways of training such a model without architecture changes.

2 Literature review

Article [8] shows the approach to auto-selecting tasks while training linear models. However, the authors evaluated the way proposed in the article only for the custom binary classification dataset. This way also cannot be directly compared to the more novel results due to the evaluated model's simplicity. Articles [9], [6] and [10] show the ways of fitting BERT on several tasks at once. However, the ways described there still require utilizing more complex architectures compared to single-task learning. The most basic way of fitting a single-task BERT, described in this article, is the following. We tune the multitask BERT without changing the base model architecture and data the model used for fine-tuning. We only change the available labels and possibly freeze some weights. In [1] while processing images, authors used predictions of models fit on 1 percent of training data for pseudo-labeling (assigning labels for unlabeled samples). In [2] authors used such a pseudo-labeling method (while having models fit for different languages) as data translation from one language to another and backward, and in [7] authors explored the impact of pseudo-labeling approaches for the computer vision tasks as well as for the machine translation. However, we did not find research comparing different pseudo-labeling ways for fitting BERT on the GLUE benchmark[5], so this article fills in this gap. The simplest way of fitting multitask BERT, described in this article, is fitting the multitask model without changing its architecture and data given as an input, but only with editing an array of labels for the multiclass task.

3 Experiments setting

In this article, we have researched different methods of training the BERT model for solving various classification tasks simultaneously. The unique feature of every considered approach is that, unlike multitask learning methods such as [10] and [9], we do not change model architecture. But the only thing we change is the array of labels sent as input. In every approach, we trained the model that solves the multilabel classification problem. Specifically, this model predicts probability from 0 to 1 for every class. In this work, we research the quality of pseudo-labeling for such tasks. We have evaluated the model on the following classification tasks: MNLI, Quora Question Pairs (further - QQP), SST-2, and RTE for the GLUE set of tasks[4]. We have chosen MNLI, QQP, and SST-2 as their datasets were large enough (>= 50000 samples for every task). We also have chosen RTE as we need to get entailment in the same way as in MNLI. We have reproduced original article results for each of these tasks. Note that the BERT model in the original article was not multilabel. The examples from all tasks were shuffled and sampled randomly. We should note that we used the BERT-Base model as a benchmark due to the computational restrictions.

4 Notations

We use the following labels in the formulas described in the article:

- $+, -$: labels *positive* and *negative* for SST dataset;
- $d, !d$: labels *duplicate* and *not duplicate* from Quora question pairs dataset;
- e, c, n: labels *entailment*, *contradiction* and *neutral* from MNLI dataset;
- $\varepsilon, !\varepsilon$: labels *entailment* and *not entailment* from RTE dataset;
- $MNLIpred, RTEpred, QQPpred, SSTpred$ - predictions of the model trained on the following task (MNLI, RTE, QQP, SST) for the label from the lower formula index;
- I denotes rounding of probability vector predicted by the original model: we round the largest element of the probability vector to 1 and all other elements to 0;
- $MNLIpred^{!e}$ is the prediction of the plain MNLI model with entailment set to zero. So, we set the predicted probability of $entailment$ to zero and then treat the 3-class classification as 2-class while predicting the plain single-label MNLI model (for classes entailment, contradiction, and neutral);
- $prob_{task}^{label}$ is the vector with probabilities from 0 to 1 that we need to assign to the example from $task$, which was labeled as $label$;
- P_{label} is the probability P of label $label$, where P is from 0 to 1.

It means that, for example:

- $MNLIpred_e$ - is the probability of entailment label, predicted by the model trained on MNLI;
- $I(MNLIpred)_e$ means that it is 1 if entailment is most likely predicted class in the MNLI task, and 0 otherwise.

$$MNLIpred_n^{!e} = MNLIpred^n/(MNLIpred^n + MNLIpred^c) \tag{1}$$

$$MNLIpred_c^{!e} = MNLIpred^c/(MNLIpred^n + MNLIpred^c) \tag{2}$$

We denote the components of probability vectors using brackets.

5 Multitask approaches explored

We considered different approaches for fitting multitask models. We present these methods below.

5.1 Independent labels

In this approach, we fit the model on the united array of RTE, MNLI, QQP, and SST-2. For every example, we consider label arrays for each of the tasks to be independent. In other words, we set for every sample the probability of absolutely all classes, except for already known, as 0. We also set the likelihood of an already known class as 1 (or 100 percent) for every sample. There are nine classes: 3 classes for the MNLI task and two classes for each other tasks. It means that default probability vector for this setting is:

$$prob_{default} = [0_\varepsilon, 0_{!\varepsilon}, 0_e, 0_c, 0_n, 0_d, 0_{!d}, 0_+, 0_-] \tag{3}$$

So for every label, we change in this default vector only the probability of the "correct" label to 100 percent, for example:

$$prob_{RTE}^{\varepsilon} = [1_\varepsilon, 0_{!\varepsilon}, 0_e, 0_c, 0_n, 0_d, 0_{!d}, 0_+, 0_-] \tag{4}$$

Other equations can be written in an analogous way.

5.2 Soft independent labels

This approach is analogous to the **Independent labels**. However, it has the following difference: we do not take down to zero probabilities of absolutely all classes we do not know for every sample. Instead, we take down to zero only the probability of all classes except for the known label for the "own" task. The probabilities of all other classes are labeled to be the same. To label them, we used the following rule: the sum of probabilities of all other classes must be equal to 1, and the probabilities of all other classes (for each task) must be equal to each other. We can quickly obtain probability coefficients for every class from the "other" task if we know these conditions. It means that default probability vector for this setting is:

$$prob_{default} = [1/2_\varepsilon, 1/2_{!\varepsilon}, 1/3_e, 1/3_c, 1/3_n, 1/2_d, 1/2_{!d}, 1/2_+, 1/2_-] \tag{5}$$

So for every label, we change in this default vector only the probability of the "correct" label to 100 percent and the probability of the "incorrect" labels from this task to 0 percent, for example:

$$prob_{MNLI}^{e} = [1/2_\varepsilon, 1/2_{!\varepsilon}, 1_e, 0_c, 0_n, 1/2_d, 1/2_{!d}, 1/2_+, 1/2_-] \tag{6}$$

We can write other equations analogously.

5.3 Augmented independent labels

This approach is similar to the **Independent labels** and **Soft independent labels**. However, it has the following difference. For every sample, we do not consider the probability of every class from a "different" task to be the same, but instead, we define it by the prediction of the base model. The base model was trained preliminarily on this "different" task to reproduce the original article results.

It means that default probability vector for this setting is:

$$prob_{default} = [RTEpred_{\varepsilon}, RTEpred_{!\varepsilon}, MNLIpred_{e}, MNLIpred_{c}, MNLIpred_{n}, QQPpred_{d}, QQPpred_{!d}, SSTpred_{+}, SSTpred\_] \quad (7)$$

So for every label, we change in this default vector only the probability of the "correct" label to 100 percent and the probability of the "incorrect" labels from this task to 0 percent, for example:

$$prob^{d}_{QQP} = [RTEpred_{\varepsilon}, RTEpred_{!\varepsilon}, MNLIpred_{e}, MNLIpred_{c}, MNLIpred_{n}, 1_{d}, 0_{!d}, SSTpred_{+}, SSTpred\_] \quad (8)$$

We can write other equations analogously.

5.4 Soft probability assumption

In this setting, as well in the settings **Independent labels**, **Soft independent labels** and **Augmented independent labels**, we trained the model on the united array of RTE, MNLI, QQP, and SST-2. However, we considered labels for these tasks to be dependent. Specifically, we downsized the number of classes for the model to 5: *positive*, *negative*, *entailment*, *contradiction*, *neutral*. We have transferred classes of that datasets to the probabilities of these five classes by the following rules. We consider labels from RTE, MNLI, and QQP to be 50 percent positive and 50 percent negative, and in that time:

- We use MNLI labels "as it is" for classes *entailment*, *contradiction* and *neutral*: one of the classes *entailment/neutral/contradiction* has probability 100 percent and other two have probability 0 percent.
- We consider QQP label *duplicate* to be *entailment* with probability 100 percent and *neutral/contradiction* with 0 percent probability. The label *not duplicate* is considered to be *neutral/contradiction* with probabilities 0.5 and 0.5 (as they need to be equal to each other and their sum must be equal to 1), and its probability to be *entailment* is set to 0.
- We consider RTE label *entailment* to be entailment with probability 100 percent and neutral/contradiction with zero probability. The label *not entailment* is considered to be *neutral/contradiction* with probabilities 0.5 and 0.5(as they need to be equal to each other and their sum must be equal to 1), and its probability to be *entailment* is set to 0.

We consider all labels on SST-2 as belonging to classes *entailment/neutral/contradiction* with the same probability equal to 1/3. At the same time, we assign labels positive/negative according to the initial SST-2 dataset.

It means that default probability vector for this setting is:

$$prob_{default} = [1/3_{e}, 1/3_{c}, 1/3_{n}, 1/2_{+}, 1/2\_] \quad (9)$$

And for examples from SST task and MNLI task, we just set in the default vector the probabilities of "correct" labels to 1 and of "incorrect" labels to 0, respectively, for example:

$$prob^{SST}_{+} = [1/3_{e}, 1/3_{c}, 1/3_{n}, 1_{+}, 0\_] \quad (10)$$

$$prob^{c}_{MNLI} = [0_{e}, 1_{c}, 0_{n}, 1/2_{+}, 1/2\_] \quad (11)$$

For RTE task and QQP task, we handle entailment in an analogous way to the:

$$prob^{\varepsilon}_{RTE} = prob^{d}_{QQP} = prob^{e}_{MNLI} = [1_{e}, 0_{c}, 0_{n}, 1/2_{+}, 1/2\_] \quad (12)$$

However, we handle "not entailment" from RTE and "not duplicate" from QQP differently:

$$prob^{!\varepsilon}_{RTE} = prob^{!d}_{QQP} = [0_e, 1/2_c, 1/2_n, 1/2_+, 1/2_-] \quad (13)$$

5.5 Soft predicted labels

This approach is analogous to the **Soft probability assumption** with the following difference. We obtain the missing labels (*contradiction/neutral* and positive/negative on tasks RTE and QQP, positive/negative on the MNLI task, *entailment/contradiction/neutral* on the SST-2 task) by the additional labeling made by the model for each task, specifically:

- If an example is not from the SST-2, we get positive/negative labels from the SST-2 model. Otherwise, we get them from the original dataset;
- If an example is from MNLI, we get labels *entailment*, *contradiction*, or *neutral* from the original dataset;
- If an example is from RTE with the label *entailment* or from QQP with the label *duplicate*, we assign the label *entailment* with probability 1 in an analogous way to the previous point;
- If an example is from RTE with the label *not entailment* or from QQP with the label *not duplicate*, we assign the probability of label *entailment* as 0. In that way, we also take probabilities of label *contradiction* or *neutral* from predictions of the model trained on MNLI, and we normalize that probabilities for the sum of probability of *contradiction* and the probability of *neutral*.

It means that default probability vector for this setting is:

$$prob_{default} = [MNLIpred_e, MNLIpred_c, MNLIpred_n, SSTpred_+, SSTpred_-] \quad (14)$$

For the SST task and MNLI task, we just set in this vector the probabilities of "correct" labels to 1 and of "incorrect" labels to 0, respectively, in the same way as in previous approach, for example:

$$prob^n_{MNLI} = [0_e, 0_c, 1_n, SSTpred_+, SSTpred_-] \quad (15)$$

And the formulas for probabilities of RTE and QQP tasks look as the following:

$$prob^{\varepsilon}_{RTE} = prob^{e}_{MNLI} = prob^{d}_{QQP} = [1_e, 0_c, 0_n, SSTpred_+, SSTpred_-] \quad (16)$$

$$prob^{!\varepsilon}_{RTE} = prob^{!d}_{QQP} = [0_e, MNLIpred^{!e}_c, MNLIpred^{!e}_n, SSTpred_+, SSTpred_-] \quad (17)$$

5.6 Hard predicted labels

This approach is analogous to the **Soft predicted labels** approach with the following changes. From labels received from the original model prediction ($MNLIpred$, $SSTpred$, $QQPpred$, $RTEpred$), the maximal probability for each task is rounded to 1, all other probabilities are rounded to 0.

It means that default probability vector for this setting is:

$$prob_{default} = [I(MNLIpred_e), IMNLIpred_c), I(MNLIpred_n), I(SSTpred_+), I(SSTpred_-)] \quad (18)$$

And for the SST task and MNLI task, we make in this vector the same changes as in previous approaches. Changes for QQP task and RTE task look in the following way:

$$prob^{\varepsilon}_{RTE} = prob^{d}_{QQP} = [1_e, 0_c, 0_n, I(SSTpred_+), I(SSTpred_-)] \quad (19)$$

$$prob^{!\varepsilon}_{RTE} = prob^{!d}_{QQP} = [0_e, MNLIpred^{!e}_c, MNLIpred^{!e}_n, I(SSTpred_+), I(SSTpred_-)] \quad (20)$$

Setting name	Average by 4 tasks	RTE	QQP	MNLI	SST
Plain(reproduced)	81.3	64.6	90.8	77.3	92.7
Independent labels	82.8	**78.3**	90.6	75.8	92.0
Soft independent labels	82.2	69.7	89.5	75.9	**92.6**
Augmented independent labels	81.4	68.1	90.5	75.6	92.4
Soft probability assumption	**84.2**	78.2	**90.7**	76.2	91.9
Soft predicted labels	83.2	76.3	90.5	76.0	92.2
Hard predicted labels	82.9	77.4	90.6	75.3	90.7
Independent labels frozen head	82.5	76.1	90.5	75.7	91.4
Soft independent labels frozen head	82.6	74.4	90.4	**76.7**	91.2

Table 1: Best accuracy on validation data (best learning rate for every setting, average by 3 runs)

5.7 Independent labels frozen head

This approach is the same as **Independent labels**, with the following exception: the head of the model (linear layer for classification) does not learn; only the body does. Formulas for this approach are the same as for **Independent labels**.

5.8 Soft independent labels frozen head

This approach is the same as **Soft independent labels**, with the following exception: the head of the model (linear layer for classification) does not learn; only the body does.

Formulas for this approach are the same as for **Soft Independent labels**.

6 Results

We have made four reproduction attempts for every approach described above, including reproducing the original article results. In a similar way to the original article, these attempts had learning rates 2e-5, 3e-5, 4e-5, and 5e-5 accordingly. As the final learning rate, we chose the learning rate for which we had the maximal accuracy on the validation set. We have defined accuracy on the validation set as the average accuracy for all four tasks. We restricted the learning to 3 epochs for all tasks. Complete obtained data or the validation set are attached below in the Appendix A. Results on the validation set and the test set are described below in Table 1 and Table 2. We should note that we achieved all these results with 10-13 % fewer parameters and without any changes to basic architecture despite the yielding lower results than [10].

7 Discussion

As we can see, considered methods yield results similar to the original BERT model results, or even better if we descry the RTE task. The reason for this exceeding for the RTE task is its similarity with other tasks from the GLUE benchmark, for which we have much more data than for RTE.

From all considered methods, **Soft probability assumption** method yields the best results on the most similar tasks: RTE and MNLI. This result shows that uniting labels while solving similar tasks is justified.

However, on different tasks, such as SST and QQP, **Augmented independent labels** method yield the best result, which is explained by the effect of knowledge transfer while solving different tasks. Nonetheless, this effect was weakly expressed or even absent while we united labels, and its reason remains unclear. Also, the absence of the accuracy growth on QQP while uniting labels can tell that unification in this task was too rough. It shows the constraints for the proposed method with label unification.

Setting name	Average by 4 tasks	RTE	QQP	MNLI(m/mm)	SST
Plain(from original article)	78.8	66.4	71.2	84.6/83.4	93.5
Plain(reproduced)	77.6	62.7	71.0	83.1/ 82.7	93.5
Independent labels	79.0	71.5	70.9	82.7/81.7	91.3
Soft independent labels	78.9	69.3	71.3	82.8/ 82.1	92.6
Augmented independent labels	77.6	64.2	**71.8**	81.2/ 80.7	**93.2**
Soft probability assumption	**79.7**	**72.7**	70.7	**83.4/82.3**	92.5
Soft predicted labels	78.8	70.3	70.7	81.7/ 81.7	92.5
Hard predicted labels	79.1	71.3	71.1	81.7/ 81.4	92.6
Independent labels frozen head	78.2	66.9	**71.8**	82.6/81.8	91.9
Soft independent labels frozen head	79.1	70.0	71.5	83.0/ **82.3**	92.4

Table 2: Best accuracy on test data (best F1 on the QQP task)

Notably, the similarity of tasks poses some constraints on applying the **Soft probability assumption**. If they were entirely dissimilar, we could not unite the labels in this task. Therefore, in that case, **Augmented independent labels** would have been the best choice, as it can be expanded to a great variety of tasks. Exploring the broader range of architectures for which this conclusion remains valid will be the subject of future research. We also leave unexplored the impact of different sampling ways on the process of learning the model. Looking at how the result varies when we try the same sampling ways as in [4] is also a subject of future research.

8 Conclusion

After considering eight different data pseudo-labeling approaches in the GLUE 4-task setting, we can single out the method **Soft probability assumption** as the best for similar tasks such as RTE and MNLI. This method excels results on RTE from the original article by 6.2 % and falls behind the original article on QQP, MNLI, and SST only by 0.5-1.2 %. However, method **Augmented independent labels** works as the best method for solving different tasks such as SST and QQP.

Acknowledgements

This work was supported by National Technology Initiative and PAO Sberbank project ID 0000000007417F63000.

References

[1] Arachie Chidubem, Huang Bert. Constrained Labeling for Weakly Supervised Learning // arXiv preprint arXiv:2009.07360. — 2021.

[2] Aroyehun Segun Taofeek, Gelbukh Alexander. Aggression Detection in Social Media: Using Deep Neural Networks,Data Augmentation, and Pseudo Labeling.Proceedings of the First Workshop on Trolling // Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying. — 2018. — P. 90:97. — Access mode: https://www.aclweb.org/anthology/W18-4411.pdf.

[3] BERT: Pre-training of deep bidirectional transformers for language understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // 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). — 2019. — P. 4171:4186. — Access mode: https://arxiv.org/abs/1810.04805.

[4] Dynamic Sampling Strategies for Multi-Task Reading Comprehension / Ananth Gottumukkala, Dheeru Dua, Sameer Singh, Matt Gardner // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — P. 920:924. — Access mode: https://www.aclweb.org/anthology/2020.acl-main.86/.

[5] GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding / Alex Wang, Amanpreet Singh, Julian Michael et al. // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. — 2018. — P. 353:355. — Access mode: https://arxiv.org/abs/1804.07461.

[6] Multi-Task Deep Neural Networks for Natural Language Understanding / Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — P. 4487:4496. — Access mode: https://www.aclweb.org/anthology/P19-1441/.

[7] Müller Rafael, Kornblith Simon, Hinton Geoffrey. When Does Label Smoothing Help? // Proceedings of the 33th NeurIPS. — 2020. — Access mode: https://proceedings.neurips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf.

[8] Pentina Anastasia, Lampert Christoph H. Multi-Task Learning with Labeled and Unlabeled Tasks // Proceedings of the 34th International Conference on Machine Learning. — Vol. 70. — 2017. — P. 2807:2816. — Access mode: http://proceedings.mlr.press/v70/pentina17a.html.

[9] Pilault Jonathan, Elhattami Amine, Pal Christopher. Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters Less Data. — 2020. — Access mode: 2009.09139.

[10] Stickland Asa Cooper, Murray Iain. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning // Proceedings of the 36th International Conference on Machine Learning. — Vol. 97. — 2019. — P. 5986:5995. — Access mode: https://arxiv.org/abs/1902.02671.

9 Appendix A. Validation set accuracies for different attempts

RTE valid accuracies

Setting name	Attempt 1	Attempt 2	Attempt 3	Attempt 4
Plain(reproduced)	64.6	57.8	63.2	62.7
Independent labels	66.8	68.2	66.4	74.0
Soft independent labels	71.1	69.3	65.0	66.8
Augmented independent labels	67.5	67.1	64.6	65.3
Soft probability assumption	76.9	76.9	71.8	71.1
Soft predicted labels	72.6	74.4	70.8	73.3
Hard predicted labels	73.3	71.8	72.9	72.5
Independent labels frozen head	70.0	72.9	70.8	69.3
Soft independent labels frozen head	71.8	66.4	65.0	67.9

SST valid accuracies

Setting name	Attempt 1	Attempt 2	Attempt 3	Attempt 4
Plain(reproduced)	92.7	92.1	91.7	89.3
Independent labels	91.6	91.1	90.7	90.4
Soft independent labels	90.8	91.4	91.5	89.5
Augmented independent labels	92.3	91.5	91.5	91.7
Soft probability assumption	92.3	91.6	90.5	90.4
Soft predicted labels	92.1	91.9	91.2	90.5
Hard predicted labels	92.4	91.4	90.5	90.8
Independent labels frozen head	89.9	90.6	90.8	91.9
Soft independent labels frozen head	91.6	90.9	89.1	90.5

QQP valid accuracies

Setting name	Attempt 1	Attempt 2	Attempt 3	Attempt 4
Plain(reproduced)	90.8	90.5	87.1	89.8
Independent labels	87.9	89.1	87.5	90.4
Soft independent labels	90.4	89.8	86.0	89.1
Augmented independent labels	90.3	89.6	90.4	90.3
Soft probability assumption	90.5	90.3	90.0	89.6
Soft predicted labels	90.0	90.4	90.1	90.0
Hard predicted labels	90.1	89.9	89.3	89.9
Independent labels frozen head	90.1	90.5	90.5	89.6
Soft independent labels frozen head	90.2	90.5	89.9	89.1

MNLI valid accuracies

Setting name	Attempt 1	Attempt 2	Attempt 3	Attempt 4
Plain(reproduced)	76.7	77.3	76.4	72.7
Independent labels	73.8	75.0	72.6	76.3
Soft independent labels	76.5	75.8	72.62	74.5
Augmented independent labels	75.5	75.1	75.3	75.0
Soft probability assumption	77.0	76.4	75.4	75.3
Soft predicted labels	75.9	76.0	76.0	75.5
Hard predicted labels	75.9	75.4	75.0	74.5
Independent labels frozen head	76.1	76.0	76.3	73.6
Soft independent labels frozen head	76.8	76.2	75.0	73.5

Adjunct role labeling for Russian

Roman Kazakov
National Research University
Higher School of Economics
Moscow, Russia
roman-k2000@mail.ru

Olga Lyashevskaya
National Research University
Higher School of Economics;
V. V. Vinogradov Russian Language
Institute of RAS
Moscow, Russia
olesar@yandex.ru

Abstract

The task of the semantic role labeling usually focuses on identifying and classifying the core, obligatory arguments of the predicate. The adjuncts of Time, Location, etc. (non-core, modifier arguments) are considered on the periphery of the task [30] and even doing the easy part of it [44], despite the fact that they are highly integrated into the clause structure and may non-trivially interact with the meaning of the verb [4, 32]. In this paper, we present experiments on labeling the adjunct roles of LOCATION, TIME, MANNER, DEGREE, REASON, and PURPOSE, based on the manually annotated Adjuncts-FrameBank data set. The results show an average F1-score of 0.94 on the gold adjunct phrase annotations using the word2vec representations of adjuncts, word2vec representations of predicates, and the moprhosyntactic marking of adjuncts. Our findings generally corroborate the theoretical hypothesis on the structural and semantic autonomy and lexico-morphosyntactic specialization of adjuncts. Yet, more complicated organization of their network is revealed, pointing to the diversity of adjuncts in terms of their distribution and behavior.

Keywords: semantic role labeling, adjunct role labeling, adjunct and predicate embeddings, Russian FrameBank, Russian language

DOI: 10.28995/2075-7182-2021-20-367-377

Определение семантических ролей сирконстантов для русского языка

Роман Казаков
НИУ Высшая Школа
Экономики
Москва, Россия
roman-k2000@mail.ru

Ольга Ляшевская
НИУ Высшая Школа
Экономики;
Институт русского языка
им. В. В. Виноградова РАН
Москва, Россия
olesar@yandex.ru

Аннотация

Задача разметки семантических ролей (semantic role labeling, SRL), как правило, строится вокруг идентификации и классификации ядерных, обязательных аргументов предиката. Сирконстанты времени, места и т. п. (неядерные, модифицирующие аргументы предиката) вытеснены на периферию [30] и даже признаются составляющими самую простую часть задачи [44]. Вместе с тем, они глубоко интегрированы в структуру клаузы и могут нетривиально взаимодействовать со значением глагола [4, 32]. В этой статье мы представляем эксперименты по определению роли сирконстантов МЕСТА, ВРЕМЕНИ, ОБРАЗА ДЕЙСТВИЯ, СТЕПЕНИ, ПРИЧИНЫ и ЦЕЛИ, на основе аннотированного вручную набора данных Adjuncts-FrameBank. Модель на основе признаков word2vec репрезентаций сирконстантов и предикатов и морфосинтаксического оформления сирконстантов показывает среднюю F1-меру 0,94 на данных, в которых вручную размечены границы предикатов, актантов и сирконстантов. Наши результаты в целом подтверждают теоретические предположения о структурной и семантической автономии и лексико-морфосинтаксической специализации адъюнктов. Тем не менее, обнаруживается более сложная организация их структуры, что указывает на разнообразие адъюнктов с точки зрения их распределения и поведения.

Ключевые слова: классификация семантических ролей, определение семантических ролей сирконстантов, векторные представления сирконстантов и предикатов, ФреймБанк, русский язык

1 Introduction

Adjunct role labeling is a sub-task of **semantic role labeling** (SRL) that addresses the identification and classification of the non-core arguments of a predicate. Among semantic roles that communicate "Who does What to Whom and Why and How and When and Where?", Who and What are the core arguments that are defined and constrained by the semantics and governing properties of the particular predicate, while other elements (adjuncts) are less strongly associated, functionally and formally, with it. One can assume that any event takes place in a certain setting, namely, Location and Time, may be explained by a certain Purpose, Reason, or Condition, and may be characterised by aspects such as Frequency, Manner, Measure, Evaluation, or Modality. As the non-core arguments are not necessary to complete the meaning of the predicate, they tend to occur only sporadically in the phrase and thus are considered semantically and syntactically non-obligatory (circumstances in Tesnière's terms [47]).

At the same time, both formal and functional schools claim that adverbs and other non-core arguments have their own selectional and structural preferences, are highly integrated into the structure of the verb phrase and may non-trivially interact with the predicate semantics and grammar, as well as with the verb phrase structure in general and other adjuncts [14, 9, 5, 27, 4, 10, 16, 17].

Over the past twenty years the task of SRL was mainly focused on detecting and labeling the obligatory arguments, overshadowing adjuncts, their essence and types. In most analyses, any information about adjuncts is omitted or briefly mentioned. Our study, by contrast, narrowly addresses the distributional properties of adjuncts as a guide to build feature-based labelers for their semantic roles in the verb phrase.

Somewhat simplifying, the task of the adjunct role labeling in a sentence can be subdivided into three subtasks:

1. whether or not the element A and the predicate P are related;
2. whether or not A is an adjunct of P;
3. of which particular type this adjunct relation is.

This paper concerns the third subtask, assuming that (1) and (2) are identified correctly. We frame adjunct classification as a supervised one-of-N classification problem. Specifically, we investigate what kind of linguistic information about the form and meaning of adjuncts and verbs is relevant to the identification of the adjunct roles.

2 Related works

SRL as a computational linguistic task has become widespread since the active development of machine learning. The pioneering work of Gildea and Jurafsky [12] used supervised machine learning to predict semantic roles in English FrameNet [3], with syntactic features having the most discriminative power. Since then, a variety of methods has been applied to SRL and SRI (semantic role induction): global optimisation [11], semi-supervised learning [11], and unsupervised learning and graph similarity [20]. More recently, various neural architectures have been found effective for the task [22, 24, 7, 46, 28].

As regards the SRL for Russian, Dialing [43], ETAP3 [13] and Compreno [45] should be mentioned among the early applications. For example, Dialing was based on the method of full variants and rules. [41] combined the dictionary-based approach with a data-driven transition-based model trained on the automatically enriched SynTagRus treebank.

After the SRL-labeled resource Russian FrameBank was published [15, 26], a number of supervised methods were evaluated on it. [18, 19] suggested an SVM-based labeling model that used hand-crafted features extracted from corpus including syntactic features and clustered lexicon. [42] combined information available in annotations (morphosyntactic features of arguments, lemmas of predicates, syntactic labels of arguments, relative positions of arguments) with the word2vec embeddings of arguments and predicates. [36] used Bi-GRU and attention to extract the potential features of arguments and then voting ensemble over three models that took both extracted and basic features. In order to overcome biases and scarcity in available annotated data, [37] suggested using the pretrained contextual embeddings and introduced two models to process the argument structures of known and unknown predicates. [1] extended [37] approach with cross-lingual transfer learning and showed that pretraining on English FrameNet slightly improves the results.

One more SRL resource for Russian was presented recently — PropBank [33], but it is not in use for now, because of its small size.

Nevertheless, the large amount of work on SRL doesn't change the fact that adjuncts are nowhere in sight for researchers. Unfortunately, they are rarely mentioned or even classified. The article of [30] can serve as a weak counterexample: adjuncts were classified into preposition phrases and adverbs. Perhaps, the reason for such a strange bypassing of adjuncts is somewhere in their nature. More detailed and deep work on adjuncts was performed in the article of [44], based on Chinese PropBank data [29]. Along with core arguments authors consider different non-core ones (arguments-modifiers): location, temporal, condition, frequency etc. Their classification architecture was based on the SVM method. One more system that works with adjuncts is Compreno model [6], but it does not distinguish core arguments and adjuncts because of another development purposes: it tries to translate a phrase in a natural language to the universal language, so the type of the valency is not important in this scope. This model is applied mainly in automatic translation. However, this model involves the structure of slots' levels where circumstantial valencies are at the higher levels and actant valencies are at the lower ones [31].

Some works are focused on automatic systems for the identifying of specifically temporal [38, 25] or locative [39] relationships.

3 Theoretical background

The main distinction between arguments and adjuncts lies in their relations with predicates that are called **valencies** [47]. Valencies are divided into two types: semantic and syntactic. Semantic valencies are those valencies of the word which attach syntactically dependent words to it, and each of them corresponds to the variable in the interpretation of the word's meaning [2]. Syntactic valencies are capacities to enter in syntactic connections with other elements [23]. Y. Testelets proposes a simple table to understand the difference between arguments and adjuncts (Table 1) [48].

Table 1. The correlation of valency types and core arguments / adjuncts [48]

	Semantic valency	Syntactic valency
core arguments	+	+
adjuncts	-	+

Consequently, core arguments fill in the semantic valency of a predicate while adjuncts do not. In other terms, (core) arguments are called binding valencies of the predicate, while adjuncts — non-binding ones. There are no ideal criteria differentiating between arguments and adjuncts distinction and there is a vast theoretical literature on borderline cases between non-obligatory core valency roles and circumstances [32, 21]. Plungian and Rakhilina suggest the criterion of the compatibility control: 1) a binding valency manifests itself in the relevant compatibility of the predicate; if it does not, there are two outcomes: there is not such a variable in the semantic representation of the predicate or, in rare cases, there are some rules that prohibit its usage; 2) the compatibility of the relevant binding valency is «non-trivial». It means that adjuncts may be used with «any» predicate, but, apparently, this does not hold empirically.

Note that compatibility frequently depends on the semantic types of adjuncts, which are not a heterogeneous class. There are different classifications of adjuncts [9]. Classes in them are more or less stable (for example, temporal adjunct can be found in all of them). For our pilot study, we used data in which predicates, core arguments and adjuncts are manually labelled. It will be described in the next section.

4 Data set

Russian FrameBank includes examples from the Russian National Corpus in which the verb predicates and their core arguments map to the dictionary of the verb constructions. Non-core elements which semantically relate to the verb but do not correspond to the argument slots in the dictionary are labelled as adjuncts, matrix predicates, or modal elements. Adjuncts include adverbs and particles, prepositional phrases, case phrases, subordinate clauses, infinitive and gerundive verb phrases for the most part, which can be either syntactically dependent on or independent of the predicate.

Table 2. Some data set strings

Phrase	Form	Role	KeyLexeme (predicate)
сильно (‘hard, really’)	ADV	DEGREE	*беспокоить* (‘to bother’)
в редакции (‘in the editorial office’)	*в* (‘in’) + S.LOC	PLACE	*брать* (‘to take’)
на ходу (‘on the move’)	*на* (‘on’) + S.LOC	MANNER	*менять* (‘to change’)
...	...	...	...

Only the pairs of adjuncts and the corresponding verb predicates were taken to compile the Adjuncts-Framebank data set. Each line (e. g. Table 2) represents an adjunct, its form, its type, and a predicate. A form can be a part of speech (e.g. ADV — an adverb), a part of speech and morphosyntactic tag (e.g. V.INF — a verb in infinitive), or a more complex structure (*за* + S.ACC *до* + S.GEN — a prepositional phrase (PP) in the accusative case with the preposition *за* (‘over’) that governs another PP with the preposition *до* (‘before’)). Adjuncts can be sentential and of any length (17% groups has length greater than 3), there are not any restrictions on their form.

Some types (explications) assigned to adjuncts in FrameBank were combined to further simplify classification (Table 3). It was decided to divide adjuncts into six groups: PLACE, TIME, MANNER, DEGREE, PURPOSE, REASON. After aggregating and cleaning, the data set includes 7860 adjunct-verb entries (976 unique verbs, 2819 unique adjuncts).

Table 3. Comparison of the FrameBank classes of adjuncts and generalized classes used in the model

Generalized class	Frequency, %	FrameBank class	Frequency, %	Example
PLACE	12.06	PLACE	10.13	*в огороде* (‘in the garden’)
		FINAL POINT	1.27	*в Госдуму* (‘to the state Duma’)
		INITIAL POINT	0.33	*отсюда* (‘from here’)
		DISTANCE	0.33	*издали* (‘from afar’)
TIME	32.02	DURATION	12.25	*долго* (‘for a long time’)
		TIME	10.99	*вчера* (‘yesterday’)
		FREQUENCY	8.62	*порой* (‘sometimes’)
		MOMENT OF TIME	0.1	*в этот час* (‘at this hour’)
		TIME – LIMIT	0.07	*по сей день* (‘to this day’)
		DURATION – LIMIT	0.05	*до рассвета* (‘until dawn’)
MANNER	7.5	MANNER	4.75	*порывисто* (‘gusty’)
		MEANS	2.15	*на глаз* (‘to eye’)
		SOUND	0.67	*громко* (‘loudly’)
DEGREE	29.44	DEGREE	29.44	*безмерно* (‘immensely’)
REASON	16.58	REASON	16.58	*от пота* (‘with sweat’)
PURPOSE	2.33	PURPOSE	2.33	*для чая* (‘for tea’)

It must be noted that there is one more classification of adjuncts. They can be divided into modifiers of sentences and modifiers of predicates [34]. They are not distinguished in this work. Hence, there are adjuncts of both types in the data set.

Figure 1 represents the correspondence analysis (CA) plot for the class of adjuncts and the parts of speech of the content-word head of the adjunct phrase. The first two dimensions of the CA plot explain ca. 84% variance. We see that PART (intensifying particles) are strongly associated with DEGREE, whereas S (prepositional phrases), V (gerundive and finite clauses), and ADV (adverbials) are rather neutral. SPRO (mostly personal and demonstrative pronominals) are associated with PURPOSE and PLACE. Other parts of speech are less frequent.

Another CA plot shows most frequent head words of adjuncts associated with the class of adjuncts (Figure 2). TIME is associated with adverbs such as *теперь* (‘now’), *потом* (‘later’), *тогда* (‘then’), *сейчас* (‘now’) and nouns such as *время* (‘time’), *год* (‘year’), *день* (‘day’), *раз* (‘time’). PLACE is associated with the adverbs *там* (‘there’), *здесь* (‘here’), *где* (‘where’), *высоко* (‘high’). DEGREE is associated with the particles *даже* (‘even’) и *и* (‘and’), adverbs *сильно* (‘strongly’), *очень* (‘very’),

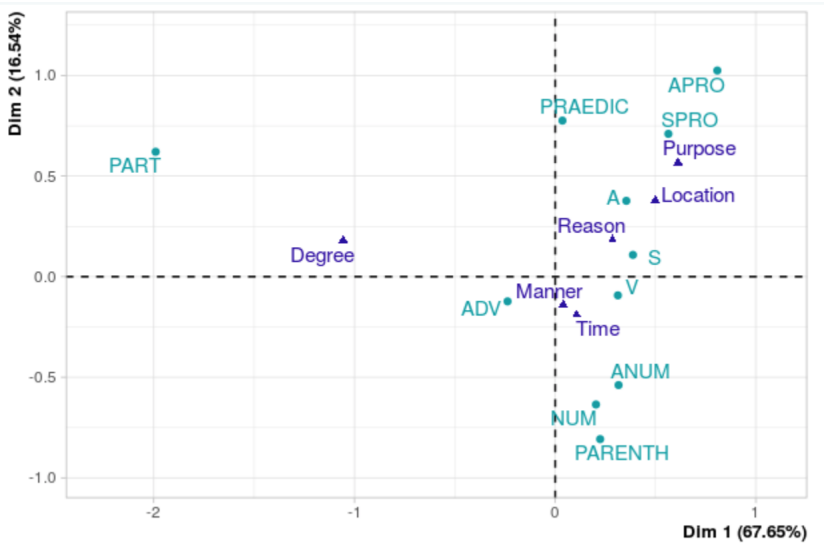


Figure 1. CA map: adjunct class vs. head POS

больше (‘more’). MANNER — with the adverbs *так* (‘this way’), *как* (‘how’), *просто* (‘just’). Interestingly, PURPOSE and REASON do not have frequent head words associated with them. REASON is also more neutral, which can indicate that it is associated with words of different classes.

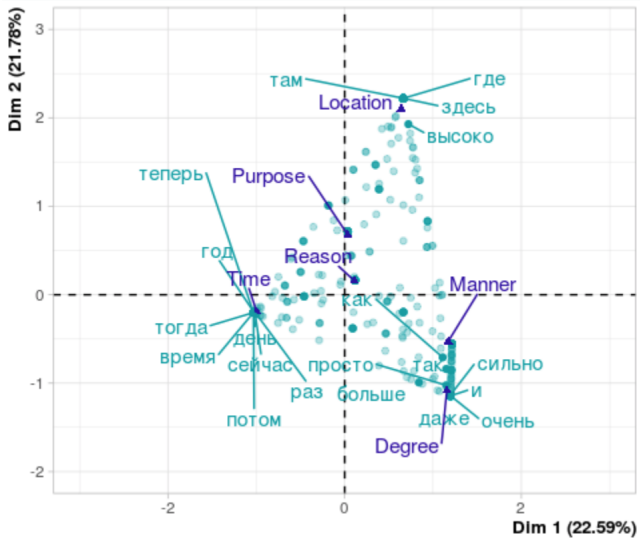


Figure 2. CA map: adjunct class vs. head word

Finally, CA plot on Figure 3 illustrates associations between the most frequent verbs and the class of adjuncts. There are three clusters:

- centered around DEGREE;
- centered around MANNER and PLACE;
- centered around TIME, REASON, and PURPOSE.

We can see some fascinating regularities, but the classifying experiment probably will make them clearer.

5 Method

The classifier predicts the role of an adjunct in the input sentence, given three features:

1. a word2vec representation [49] of an (lemmatized) adjunct; if an adjunct consists of more than one word, the program calculates the mean of all meaningful constituents’ scores;

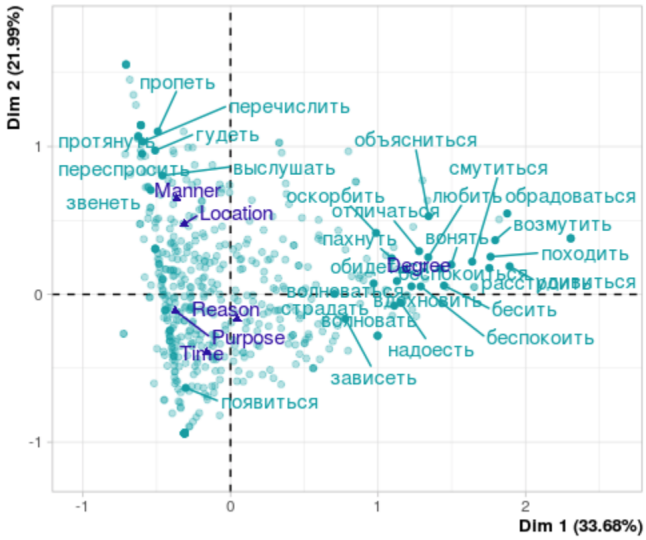


Figure 3. CA map: adjunct class vs. verb

2. a word2vec vector score of a corresponding (lemmatized) predicate;
3. a form (see Section 3); it is a categorical feature, so all forms' types (27) were encoded to lists of 0 and 1 by one-hot encoding.

In this work, a word2vec model[1] was used, pre-trained on Taiga Corpus [40] (about 5 GW) and supplied with Universal POS Tags. The model architecture was continuous skip-gram, with the vector size of 300 and the context window size of 2.

Then, after preprocessing, the data were divided into two samples: a training set (80%) and a test one (20%).

We tested four classification models on our data: Decision Tree, Random Forest and two Gradient Boosting models: basic one from Scikit-learn [35] Python library and advanced one from CatBoost [8] Python library. As a baseline, two models were used that choose the most frequent tag for the adjunct (character string) and the adjunct form, respectively. Table 4 represents the quality metrics of these models and the baseline models.

Table 4. Model performance

Model	Precision	Recall	Macro F1-score
Baseline	–	–	0.09
Decision Tree	0.88	0.87	0.87
Random Forest	**0.95**	0.89	0.91
Gradient Boosting	0.92	0.91	0.91
CatBoost	**0.95**	**0.94**	**0.94**

The CatBoost model performed best (F1-score is 0.94).

6 Results

A summary of the CatBoost model performance on the test set is presented below (Table 5). The macro-F1 is 0.94 and micro- and weighted F1 is 0.95, which indicates the high quality of the model. Whereas the F1-score for the roles of TIME and PLACE is 0.97 and higher, the model performs poorly on the minority class, REASON (F1 = 0.9). It has rather low recall (0.86).

Figure 4 reports on the feature importance metric calculated on the CatBoost and Random Forest models. Surprisingly, the adjunct word2vec feature emerges as more important than the two others used in

[1] The model was taken from RusVectōrēs (URL: `https://rusvectores.org/ru/models/`).

Table 5. Detailed metrics of the CatBoost model

	Precision	Recall	Macro F1-score
PLACE	0.94	0.97	0.96
TIME	0.98	0.98	0.98
MANNER	0.92	0.95	0.93
DEGREE	0.97	0.93	0.95
PURPOSE	0.95	0.95	0.95
REASON	0.94	0.86	0.90
Macro average	0.95	0.94	0.94
Weighted average	0.95	0.95	0.95

both model. However, importance of the form feature is different (0.098 for CatBoost and 0.022 for Random Forest). Probably, it is because the CatBoost algorithm handles categorical data better than other algorithms with one-hot encoding. Perhaps it is the main reason why the CatBoost model performed better than others.

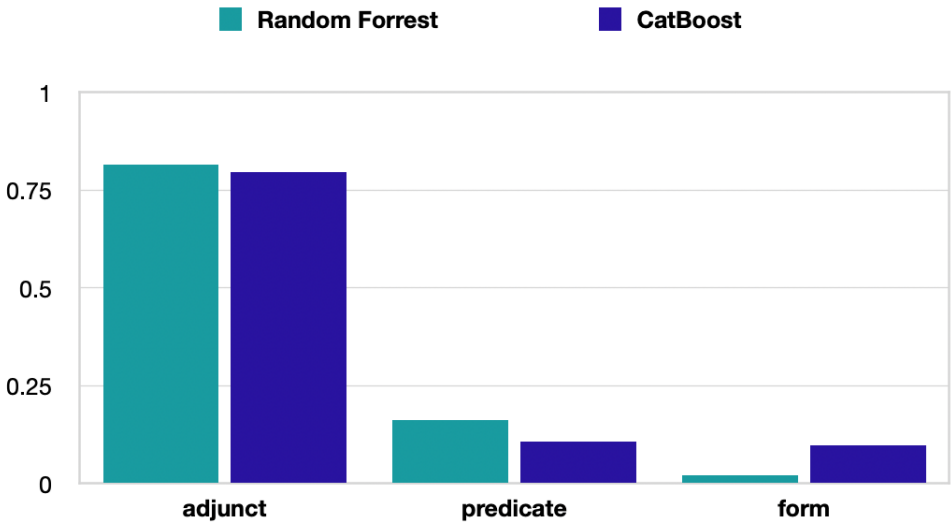


Figure 4. Feature importance of the CatBoost and Random Forest models

7 Analysis

It turned out that the adjunct word2vec feature is the most important feature. This indicates that adjuncts by themselves are strongly associated with certain semantic roles and, to some extent, semantically independent of the predicates. Furthermore, the word2vec vector score of a predicate feature is relatively low (hereinafter the CatBoost model: 0.106), but not around zero, and this confirms that adjuncts do not fill in semantic valencies of predicates, though at the same time adjuncts may not be used with any predicate. Perhaps this depends on the type of adjunct.

The lowest F1-score was shown by the group of REASON adjuncts (F1=0.9). Interestingly, there is a wide gap between precision and recall. This means that the classifier leaves a lot of true negatives and some types resemble them. This is most likely due to the heterogeneity of REASON adjuncts and the small amount of them in our data sample (only 132 elements in the test set). In addition, adjuncts of REASON and PURPOSE are often longer than others (they are expressed sometimes by subordinate clauses). By contrast, temporal adjuncts performed almost perfectly (the F1-score is 0.98) due to their semantic similarity. Thus, more homogeneous types demonstrate better results. This situation had been described in detail via a correspondence analysis: it presented three distinct clusters including temporal,

locative, manner and degree adjuncts (Figure 2). It is noteworthy that higher results were expected from locative adjuncts (the F1-score is 0.96), because this type seemed to be extremely detached.

PLACE, MANNER, REASON are classes that have the lowest precision: it seems that the word2vec embeddings models are not confident enough to distinguish metaphorical shifts, e.g. assigning PLACE to adverbs used in temporal meaning (cf. *здесь* 'here, in this moment') and MANNER to adverbs used as intensifiers (cf. *нежно* 'gently, slightly').

Besides that, the lemmatized word2vec models are content words-biased and fail to identify cases in which the semantics of the preposition and other closed-class words shifts the interpretation of the whole phrase (e.g. *минуту* 'for a minute', TIME - *ради такой минуты* 'for the sake of such a moment', PURPOSE).

8 Discussion

Considering the feature importance metrics, it may seem like the representation of an adjunct is the only significant feature, but it is not. There are some examples (*под бомбами* 'under the bombs' — PLACE) where meaningful words do not refer to the type of the adjunct. Such cases provide the idea that form and predicate also affect the final prediction of the classifier.

9 Conclusion

In this work, we performed the labeling experiments for Russian adjuncts with rather high efficiency (average F1-score is 0.94). The proposed method can further be implemented in the SRL systems to classify elements that are not core arguments of the predicate. Owing to its interpretability, it promises to be helpful, in different areas, for example, information retrieval, information extraction, QA-systems, and automatic translation.

The word2vec representations of adjuncts turned out to be a powerful effect in the task. Contextual embeddings and/or syntactic features are obvious candidates to be added to the model in order to improve its quality and ensure more accurate labeling of long adjuncts.

The theoretical assumptions that adjuncts do not fill in semantic slots of predicates based on the distinction of arguments and adjuncts are confirmed for the most part. Adjuncts demonstrate high independence from predicates, but apparently it is not an absolute truth.

References

[1] Alimova I., Tutubalina E., Kirillovich A. Cross-lingual Transfer Learning for Semantic Role Labeling in Russian // Fourth International Conference Computational Linguistics in Bulgaria. — 2019. — P. 72.

[2] Apresyan Yu. D. Lexical semantics [Leksicheskaya semantika]. — 2 edition. — Moscow : Yazyki russkoj kul'tury, 1995.

[3] Baker C. F., Fillmore C. J., Lowe J. B. The Berkeley FrameNet Project // 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. — Montreal, Quebec, Canada : Association for Computational Linguistics, 1998. — P. 86–90. — Access mode: https://www.aclweb.org/anthology/P98-1013.

[4] Boguslavsky I. M. Scope of lexical units [Sfera dejstviya leksicheskih edinic]. — Moscow : Shkola "Yazyki russkoj kul'tury", 1996.

[5] Cinque G. Adverbs and Functional Heads: A Cross-Linguistic Perspective. — Oxford : Oxford University Press, 1999.

[6] The Compreno Semantic Model as Integral Framework for Multilingual Lexical Database / E. Manicheva, M. Petrova, E. Kozlova, T. Popova // Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon. — Mumbai, India : The COLING 2012 Organizing Committee, 2012. — Dec. — P. 215–230. — Access mode: https://www.aclweb.org/anthology/W12-5117.

[7] Deep Semantic Role Labeling with Self-Attention / Zh Tan, M. Wang, J. Xie et al. // CoRR. — 2017. — Vol. abs/1712.01586. — 1712.01586.

[8] Dorogush A. V., Ershov V., Gulin A. CatBoost: gradient boosting with categorical features support. — 2018. — 1810.11363.

[9] Ernst T. The Syntax of Adjuncts. — Cambridge University Press, 2001. — Vol. 96 of Cambridge Studies in Linguistics.

[10] Filipenko M. V. Adverbs in "Lexicograph" system // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2004" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2004"]. — Moscow : Nauka, 2004. — P. 650–655.

[11] Generalized Inference with Multiple Semantic Role Labeling Systems / P. Koomen, V. Punyakanok, D. Roth, W. Yih // Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). — Ann Arbor, Michigan : Association for Computational Linguistics, 2005. — P. 181–184. — Access mode: https://www.aclweb.org/anthology/W05-0625.

[12] Gildea D., Jurafsky D. Automatic Labeling of Semantic Roles // Computational Linguistics. — 2002. — Vol. 28, no. 3. — P. 245–288. — Access mode: https://www.aclweb.org/anthology/J02-3001.

[13] Interfacing the Lexicon and the Ontology in a Semantic Analyzer / I. Boguslavsky, L. Iomdin, V. Sizov, S. Timoshenko // Proceedings of the 6th Workshop on Ontologies and Lexical Resources. — Beijing, China : Coling 2010 Organizing Committee, 2010. — P. 67–76. — Access mode: https://www.aclweb.org/anthology/W10-3308.

[14] Jackendoff R. Semantic interpretation in generative grammar Cambridge // MA: MIT. — 1972.

[15] Kashkin E. V., Lyashevskaya O. N. Semantic roles and construction net in Russian FrameBank [Semanticheskie roli i set'konstrukcij v sisteme FrameBank] // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog"]. — Vol. 12. — 2013. — P. 1.

[16] Khrakovsky V. S. Consept of adjunct and its status [Ponjatie sirkonstanta i ego status] // Semiotika i informatika. — 1998. — Vol. 36. — P. 141–153.

[17] Khrakovsky V. S. Verbocentric approach to constructions and/or Construction Grammar [Verbotcentricheskij podkhod k konstruktsijam i/ili grammatika konstrukcij] // Smysly, teksty i drugie zakhvatyvajuschie sjuzhety. Sbornik statej v chest' 80-letija I. A. Mel'čuka. — 2012. — P. 288–300.

[18] Kuznetsov I. Semantic Role Labeling for Russian Language Based on Russian FrameBank // Analysis of Images, Social Networks and Texts / Ed. by M. Yu. Khachay, N. Konstantinova, A. Panchenko et al. — Cham : Springer International Publishing, 2015. — P. 333–338.

[19] Kuznetsov I. Automatic semantic role labelling in Russian language, PhD thesis : Ph.D. thesis / I. Kuznetsov ; Higher School of Economics. — Moscow, 2016.

[20] Lang J., Lapata M. Unsupervised Semantic Role Induction with Graph Partitioning // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. — Edinburgh, Scotland, UK. : Association for Computational Linguistics, 2011. — P. 1320–1331. — Access mode: https://www.aclweb.org/anthology/D11-1122.

[21] Lazard G. L'Actance. — Paris : Presses Universitaires de France, 1994.

[22] Learning Structured Natural Language Representations for Semantic Parsing / J. Cheng, S. Reddy, V. Saraswat, M. Lapata // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Vancouver, Canada : Association for Computational Linguistics, 2017. — P. 44–55. — Access mode: https://www.aclweb.org/anthology/P17-1005.

[23] Linguistic Encyclopedic Dictionary [Lingvisticheskij enciklopedicheskij slovar'] / Ed. by V. N. Yartseva. — Moscow : Sovetskaya enciklopediya, 1990.

[24] Linguistically-Informed Self-Attention for Semantic Role Labeling / E. Strubell, P. Verga, D. Andor et al. // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. —

Brussels, Belgium : Association for Computational Linguistics, 2018. — P. 5027–5038. — Access mode: https://www.aclweb.org/anthology/D18-1548.

[25] Llorens H., Saquete E., Navarro-Colorado B. Automatic system for identifying and categorizing temporal relations in natural language // International Journal of Intelligent Systems. — 2012. — Vol. 27, no. 7. — P. 680–703. — https://onlinelibrary.wiley.com/doi/pdf/10.1002/int.21542.

[26] Lyashevskaya O. N., Kashkin E. V. FrameBank: A Database of Russian Lexical Constructions // Analysis of Images, Social Networks and Texts / Ed. by M. Yu. Khachay, N. Konstantinova, A. Panchenko et al. — Cham : Springer International Publishing, 2015. — P. 350–360.

[27] Maienborn C., Schäfer M. Adverbs and adverbials // Semantics: Lexical Structures and Adjectives / Mounton de Gruyter. — 2019. — P. 477–514.

[28] Munir K., Zhao H., Li Z. Adaptive Convolution for Semantic Role Labeling // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2021. — Vol. 29. — P. 782–791.

[29] Palmer M., Gildea D., Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles // Comput. Linguist. — 2005. — Mar. — Vol. 31, no. 1. — P. 71–106. — Access mode: https://doi.org/10.1162/0891201053630264.

[30] Palmer M., Gildea D., Xue N. Semantic Role Labeling // Synthesis Lectures on Human Language Technologies. — 2010. — Vol. 3, no. 1. — P. 1–103. — https://doi.org/10.2200/S00239ED1V01Y200912HLT006.

[31] Petrova M. The Compreno Semantic Model: The Universality Problem // International Journal of Lexicography. — 2013. — 05. — Vol. 27. — P. 105–129.

[32] Plungian V. A., Rakhilina E. V. The valence paradoxes // Semiotika i informatika. — 1998. — Vol. 36.

[33] The Russian PropBank / S. Moeller, I. Wagner, M. Palmer et al. // Proceedings of the 12th Language Resources and Evaluation Conference. — Marseille, France : European Language Resources Association, 2020. — P. 5995–6002. — online; accessed: https://www.aclweb.org/anthology/2020.lrec-1.734.

[34] Russian grammar [Russkaya grammatika] / Ed. by N. Yu. Shvedova. — Moscow : Nauka, 1980. — Vol. 2.

[35] Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort et al. // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.

[36] Semantic Role Labeling For Russian Language Based on Ensemble Model / X. Zheng, B. Zhou, J. Huang et al. // 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). — 2019. — P. 1263–1268.

[37] Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates / D. Larionov, A. Shelmanov, E. Chistova, I. Smirnov // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — Varna, Bulgaria : INCOMA Ltd., 2019. — Sep. — P. 619–628. — Access mode: https://www.aclweb.org/anthology/R19-1073.

[38] Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations / N. UzZaman, H. Llorens, L. Derczynski et al. // Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). — 2013. — P. 1–9.

[39] Semeval-2015 task 8: Spaceeval / J. Pustejovsky, P. Kordjamshidi, M.-F. Moens et al. // Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015) / ACL. — 2015. — P. 884–894.

[40] Shavrina T., Shapovalova O. To the methodology of corpus construction for machine learning:“Taiga” syntax tree corpus and parser // Proceedings of “CORPORA-2017” International Conference. — 2017. — P. 78–84.

[41] Shelmanov A., Smirnov I. Methods for semantic role labeling of Russian texts // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2014”

[Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2014"]. — 2014. — P. 607–619.

[42] Shelmanov A. O., Devyatkin D. A. Semantic role labeling with neural networks for texts in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog"]. — 2017. — P. 245–256.

[43] Sokirko A. V. Semantic Dictionaries in the Natural Language Processing: Based on the DIALING system [Semanticheskie slovari v avtomaticheskoj obrabotke teksta: Po materialam sistemy DIALING] // Cand. Tech. Sc. Dissertation. — Moscow, 2001.

[44] Sun H., Jurafsky D. Shallow Semantic Parsing of Chinese // Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. — Boston, Massachusetts, USA : Association for Computational Linguistics, 2004. — P. 249–256. — Access mode: https://www.aclweb.org/anthology/N04-1032.

[45] Syntactic And Semantic Parser Based On Abbyy Compreno Linguistic Technologies / V. P. Selegej, K. V. Anisimovich, F. R. Minlos et al. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2012"]. — Vol. 2. — 2012. — P. 91–103.

[46] Syntax for Semantic Role Labeling, To Be, Or Not To Be / Sh. He, Z. Li, H. Zhao, H. Bai // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 2018. — P. 2061–2071. — Access mode: https://www.aclweb.org/anthology/P18-1192.

[47] Tesnière L. Elements of Structural Syntax [Eléments de syntaxe structurale]. — Paris : Klincksieck, 1959.

[48] Testelets Ya. G. Introduction to general syntax [Vvedenie v obshchij sintaksis]. — Moscow : Rossijskij gosudarstvennyj gumanitarnyj universitet, 2001. — P. 156–228.

[49] word2vec / T. Mikolov, K. Chen, G. Corrado et al. // URL https://code.google.com/p/word2vec. — 2013. — Vol. 22.

A New Electronic System for Comparative Analysis of Verse and Prose

Evgeny Kazartsev
National Research University Higher School of Economics (HSE University), Faculty of Humanities / Moscow, 105066, Staraya Basmannaya St, 21/4
kazar@list.ru

Tatiana Zemskova
National Research University Higher School of Economics (HSE University), Faculty of Humanities / Moscow, 105066, Staraya Basmannaya St, 21/4
tatzem98@gmail.com

Abstract

This paper will focus on the development of a new computational system, *Prosimetron*, which enables comparative statistical studies of the rhythm of verse and prose in different languages (currently 10 languages are operative, with the possibility of adding more). The results of the analysis can be used not only for studying the processes for the genesis, expansion, and modification of various versification systems, but also for commenting on and interpreting the verse rhythm in different national poetic traditions in comparison with their foreign sources and language prosody. In addition, the possibility to model various processes of poetic speech generation and to analyze rhythmic vocabularies of prose allows hypotheses about the cognitive mechanisms of verse generation. This system operates in a semi-automatic mode and, by minimizing errors and enabling the processing of large amounts of data, provides a unique tool for computer research on the rhythm of different modes of speech.

Keywords: rhythm; prose and poetry; comparative statistical analysis; versification models
DOI: 10.28995/2075-7182-2021-20-378-384

Новая компьютерная система сравнительного анализа стиха и прозы

Евгений Казарцев
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), факультет гуманитарных наук / г. Москва, 105066, Старая Басманная, 21/4
kazar@list.ru

Татьяна Земскова
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), факультет гуманитарных наук / г. Москва, 105066, Старая Басманная, 21/4
tatzem98@gmail.com

Аннотация

В работе описываются элементы новой компьютерной системы – Прозиметрон, которая позволяет осуществлять сравнительно-статистический анализ ритма стиха и прозы на разных языках (на настоящий момент доступны 10 языков, планируется расширение этого списка в дальнейшем). Результаты исследования могут быть использованы не только для изучения процессов становления, распространения и эволюции систем стихосложения, но и для изучения ритма стиха в разнообразных поэтических традициях в сравнении с их иностранными источниками и языковыми характеристиками. Кроме того, возможность моделирования процессов порождения стихотворной речи позволяет выдвигать гипотезы о когнитивных процессах, связанных с генерацией стиха. Предложенная система функционирует в полуавтоматическом режиме и, позволяя обрабатывать большое количество данных, представляет собой уникальный инструмент для компьютерного анализа ритма стиха и прозы.

Ключевые слова: речевой ритм; проза и стих; сравнительный статистический анализ; модели стихосложения

1 Introduction

This research is devoted to the study of poetic rhythm based on a new computational system for the analysis of prosodic structures in different languages. The system works on the so-called "Russian" or "linguo-statistical method"[1], and is called *Prosimetron*,[2] because it enables the analysis of rhythm both in metrically organized texts, and in texts that are free from metrical organization — thus both in verse and in prose. Currently, this system is still being developed; however, there are already some results from this project that correspond with the results and calculations in previous studies that were carried out manually or with the use of basic computational tools. This allows us to argue that the new system works well and provides reliable information. Some of its interesting findings are presented in this paper.

The development of this new computational system for verse analysis is an important step towards studying peculiarities in the organization of poetic speech against the background of prose rhythm in various languages [2: 155]. This system is comprehensive: it can store and process various texts in relation to their rhythm.

All texts entered into the system are pre-attributed. For each text, a separate data cell is created, which indicates such parameters as the author, year of creation, title, language of the work, type (prose or poetry), bibliographic description of the source, name of the person who puts the text into the system, verse parameters for poetic texts, and some others. The system is already a unique repository of both fairly common texts and rare ones, including several editions of the same text, which can be used to study changes in the rhythm of an idiolect or the development of rhythmic inertia in verse. To enable further rhythmic analysis special rhythmic markings are added to all texts: the border of rhythmic words and the position of the stress are indicated. For metrical texts, special phenomena such as caesura are also noted.

Currently there are two modes for marking texts: automatic and manual. Automatic indication of tresses and boundaries of rhythmic words has so far been implemented only for Russian, but in the future it will be expanded to all working languages of the system. The accuracy of this marking is now about 90%.[3] Whichever way the text has been marked, it then goes through a semi-automatic three-stage process of checking and correcting the markup. The first stage involves a search for technical errors — when the stress falls on consonants, spaces and other signs — and is carried out automatically. At the second stage all verse lines are checked for compliance with the meter of the entire poem (the meter is determined automatically). If a line does not match the meter of the entire text, either the text is designated as polymetric, or the line is marked as invalid and needs to be checked by a user.

The third and the most important stage in validating the markup involves the possibility of a manual search for all forms of a particular word in a rhythmic context and manually correcting the position of stress or other attribution, followed by the automatic recalculation of all related parameters. While a part of the validation phase, this toolkit also provides an important means for research. Seeing at what position in the poetic line a particular word or its form is most often found, we can draw conclusions regarding its stress status both within a given text and in general (for example, the controversial status of some rhythmic clitics can be revised). Automatic recalculation of all results allows users to quickly change the entire markup in case it has been decided to change the stress status of a particular word.

In the future, it is planned to create an algorithm that will be able to automatically search for texts on the network, mark them up and then enter them into the corpus with minimal human participation. This will expand the applicability of the techniques developed using the current material to an unlimited number of texts, enabling the Prosimetron system to carry out big data analysis.

As already mentioned, Prosimetron currently works with two categories of texts: prose and poetry. In total, 10 working languages have now been implemented: Ancient Greek, English, German, Dutch, French, Swedish, Russian, Belarusian, Ukrainian, and Polish. All of these are represented by samples of both types of texts. The total volume of the poetic corpus is about 250,000 lines, while the prose corpus

[1] That terms are introduced into the international scientific use by James Bailey [1].

[2] The authors of this article, together with Boris Maslov and Viktor Vashchenkov, as well as a group of HSE-University students led by Evgeny Kazartsev and Tatyana Zemskova, are actively participating in the development of that system, especially in the framework of a project supported by Russian Science Foundation in the period 2016–2020. The name *Prosimetron* was suggested in 2020 by Boris Maslov.

[3] Two algorithms are used to prosodic marking of verse, a preliminary one based on the Zaliznyak dictionary and a final one based on a specially trained recurrent neural network trained on a large corpus of metrical texts.

contains about 390,000 rhythmic (phonetic) words.[4] At the moment the poetic corpus consists primarily of clearly metrical poems, written in iambic, trochaic, dactylic, etc., but there are also several syllabic and polymetric poems. In principle, there are no restrictions on the type of versification.

2 Results

Let us show some analytic possibilities that the Prosimetron can provide. One of the results from processing a prose text is the compilation of rhythmical vocabularies. In other words, it is possible to find out which rhythmic words (taking into account their length and stress placement) are most often found in a particular text and, more broadly, in a particular language. It is important to note that in this context we are dealing specifically with rhythmic words: thus, when rhythmic clitics (prepositions, conjunctions, articles, particles etc.) combine with rhythmically independent words they form groups of syllables, united by a single stress vertex. For example, *the táble*, *in a tíme*, *dróp it*, or Russian combinations *like moi dóm*, *na dne′*, *pered léśom* etc all comprise single rhythmic words.[5] With the help of this toolkit, it becomes clear that the non-fixed Russian stress generally tends to be in the middle of a word rather than closer to the ends.[6] This result — previously observed in some texts of Russian prose — is now confirmed and strengthened thanks to the computational analysis of big data.

Russian words look longer than English or German, and some scholars believe that the relatively low number of realized stresses in Russian metrical verse as compared with English or German is predetermined by the word length. Russian words seem to be longer, and therefore stress cannot be realized at as many strong positions of verse. Figure 1, which shows the average word length (in syllables) found in fiction samples from different languages, indicates that this is not the case. Granted, the average number of syllables for rhythmic words in the Slavic languages is slightly larger, but this difference is not so great that by itself it could cause the number of realized stresses in the Slavic poetic tradition to differ so sharply from, for example, the German: in German verse, about 75% of iambic tetrameter lines are fully-stressed, in Russian a little less than 30% are, and in Ukrainian still fewer. Most likely, the key factors are the author's intention and internal laws of the poetic tradition, while the language itself does not necessarily predetermine the verse rhythm. For example, in English iambic tetrameter verse, omissions of metrical stresses occur almost as often as in the Russian or Ukrainian iamb, although English words are shorter than Russian or Ukrainian.[7]

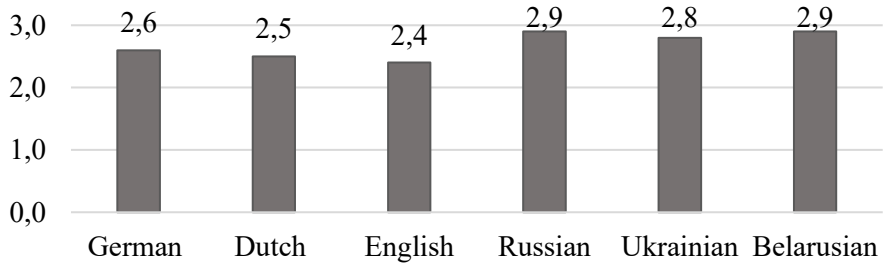


Figure 1: Average Length of Rhythmic Words[8]

[4] The corpus and the Prosimetron-system are in a state of formation, now the largest part of corpus, about 66%, is represented by Russian texts. An increase in the share of other languages is underway.

[5] Thus, a rhythmic word is identical to the concept of a phonetic word, for Russian and other East Slavic languages it is, as a rule, a complex of syllables united by one stress, for German, Dutch or English it is the same complex, but united by one main stress (Hauptton), for example *zur Straśsenbahnlinie*. In this paper, speaking about the word, we will mean a rhythmic/phonetic word.

[6] For example, in Pushkins "The Captain's Daughter" among the three-syllable words, words with an accent on the second syllable prevail, their 16% of all words in the text, three-syllable words with an accent on the first syllable are only 6%, and three-syllable words with an accent on the third 9%. In Pasternak's "Doctor Zhivago" are practically the same figures: 15%, 7% and 8%. A similar picture is observed in longer words, with 4, 5 and 6 syllables, words with a non-extreme position of stress also prevail in them. And that is in all Russian texts.

[7] For more information, see Kazartsev's paper [3].

[8] The average length of rhythmic words in every language was calculated on the corresponding prose corpus in Prosimetron-system in the following way: the total number of syllables in prose corpus (of selected language) is divided by the number of rhythmic words in this corpus.

We have already said that the system allows one to consider individual words in all possible metrical and rhythmical contexts. For example, we can find all iambic tetrameter verse lines in which irregular, non-metrical stress on the odd syllables — first, third, and fifth — is present. It is known that, in principle, the continental model of metrical versification, unlike the insular English tradition, largely prohibits non-metrical stressing — that is, the replacement of iambic line fragments with, for example, trochaic ones. The Prosimetron allows us to see all cases when stress at the odd positions is present in a verse line of, for example, Russian and Belarusian iambic verse. Thus, if we look for all such cases in the Russian iambic tetrameter, we will see that, in general, examples of non-metrical stress are extremely rare around all strong positions, and only the first strong position is a place where a trochee may occasionally replace an iamb. In general, this is not surprising, because Russian versification, of course, inherits the continental model.

	Syllables in Iambic Tetrameter			
	1	**2**	**3**	**4**
Russian	4,7%	0,8%	0,7%	0,7%
Belorussian	4,9%	4,2%	3,8%	5,0%

Table 1: The Number of Violations on Weak Positions of Iambic Tetrameter

So, in the Russian iambic tetrameter of the 19th and 20th centuries, only about 5% of the lines have a stress on the first syllable, the rest - about 1% or even less. But further from the center of Europe through Russia to the south, this rule gradually loses its strength. If the same analysis is carried out for the Belarusian iambic tetrameter of the corresponding period, we will see that violations occur with approximately the same frequency — 4-5% — on all odd syllables. Weak positions in the iambic line here receive more stresses. That is, the purity in the realization of the continental model of iambic verse gradually dissipates.

Prosimetron allows us to identify, in the context of intercultural influences, rhythmic features inherent to a tradition, as well as stages of internal development. Thus, for the Russian iambic tetrameter, which was discussed above, one can show the evolution in how the number of stresses on strong positions changes; in other words, the evolution of the so-called verse stress profile (figure 2).

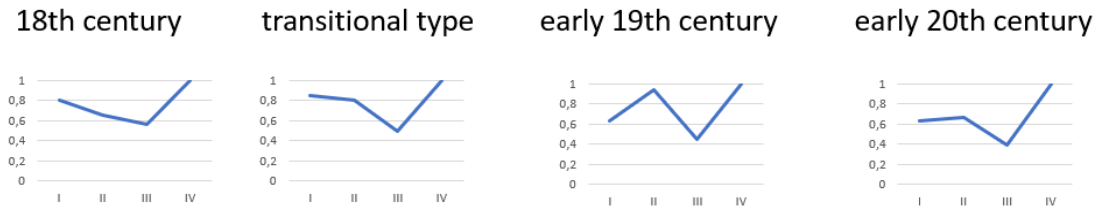


Figure 2: The Evolution of Metrical Realization in Russian Iambic Tetrameter

These data, obtained automatically by Prosimetron, correspond to the data collected manually by Taranovskii in the 1930s and 1940s, and describe the most important stages in the evolution of the Russian iambic tetrameter (first published in 1953) [6]. They show the extent to which the rhythmical profile has varied. In some periods, it looks more like a frame. In some, it resembles a jagged streak of lightning: for instance, an alternating rhythm was typical for Alexander Pushkin's poems in the Russian tetrameter after 1825. Derived by analyzing many lines of verse, these statistics allow us to speak not only about the author's rhythmic originality, but also about some of the internal laws of the iambic tetrameter's development, including the tendency towards the distribution of strong ictuses (those heavily stressed). In addition, this technique can be used for the attribution of texts, since the rhythm of certain authors can be quite characteristic.

In the context of studying national literatures, Prosimetron makes it possible to obtain qualitatively new data. The more texts used in the analysis, the more unexpected results that can be obtained. Let's cite two specific examples. The complete rhythmic similarity of Lev Loseff's poem "Iosif Brodsky or

an ode to 1957" and Alexander Pushkin's "Eugene Onegin" (1826) was completely unexpected. It would seem that if Loseff's text has a subtext, then it is necessary to look for it in Brodsky's work, but the rhythmic structure (extremely uncharacteristic for Loseff on average) coincides with the structure of Eugene Onegin and thus highlights a second important layer of sources for this text (see figure 3).

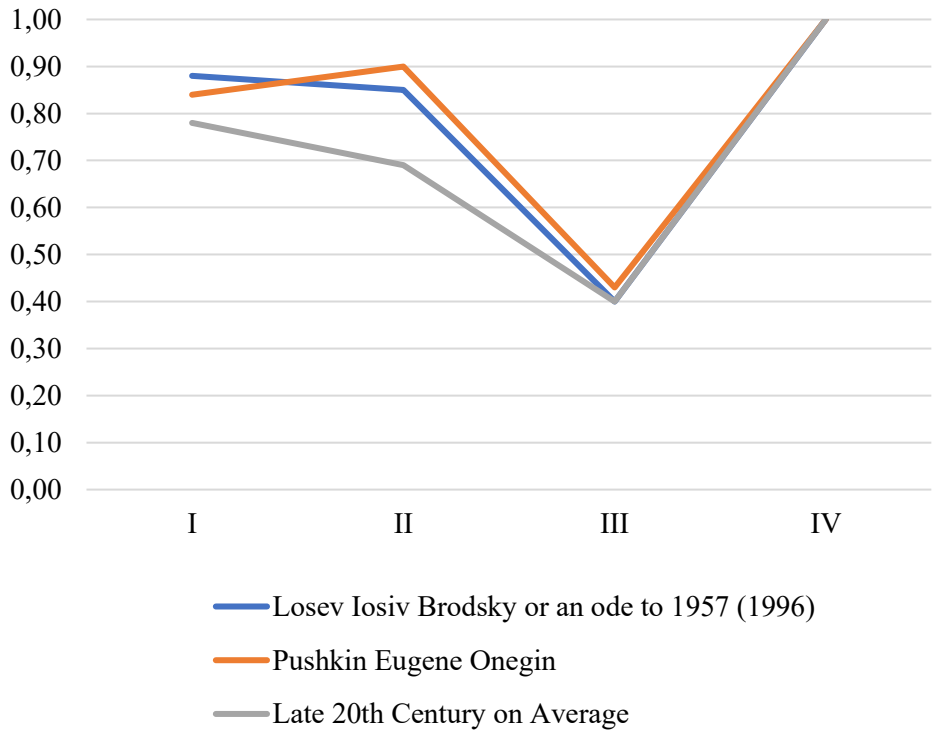


Figure 3: Three Stress Profiles of Russian Verse

These three graphs show that Lev Loseff's verse from the second half of the 20th century significantly diverges from the corresponding tradition of his era and turns out to be surprisingly close to Pushkin's verse from the mid-20s of the 19th century.[9] Of course, such parallels can be important both for philological interpretive work and for comparing the influence of different languages on each other, as well as for comparing different modes and styles of speech.

The next important example of a Prosimetron finding involves a comparative analysis of Vasilii Zhukovsky's iambic pentameter in translations from Johann Peter Hebel. His translations of this particular Alemannic author are known to contain, for example, the first instance of Russian iambic pentameter without caesura, namely, the poem "Vergänglichkeit". It is this meter that will later be used in "Boris Godunov" and many other significant texts. But the rhythmic influence of Hebel on Zhukovsky actually runs much deeper. For example, in the text "Morning Star", Zhukovsky imitates the rhythm of Hebel directly, but functionally. What I mean is that where Hebel uses the same iambic rhythmic form for several lines, Zhukovsky also does not change the rhythmic pattern; where Hebel's form changes, Zhukovsky makes a change as well. This virtuoso game with rhythm is most likely not perceived by readers, but there is reason to believe that it was created by Zhukovsky, and it certainly influenced the developing Russian iambic verse tradition (a similar technique will appear later in the works of Pushkin and other authors).

[9] Despite the fact that the Loseff's poem is much shorter than Pushkin's verse, obviously, such a tendency could not have developed by chance, the centuries-old experience of Russian studies of metrics shows that, as a rule, the tendencies inherent in large poems are preserved in small texts. Lev Loseff copies Pushkin's manner not only in rhythm, but in the style of the text.

An Example of a Similar Rhythmic Structure of Hebel's Text and Zhukovsky's Translation:[10]

[78] :	Form 1 (-\-\-\-\ ):	Was wa<ndlet| dö<rt| im Mo<rge-| Stra<hl|	[78] :	Form 4 (-\-\---\ ):	Но кто< там| в у<тренних| луча<х|
[79] :	Form 1 (-\-\-\-\ ):	mit T[ue]<ch| und Cho<rb| dur's Ma<tte-| Tha<l?|	[79] :	Form 4 (-\-\---\ ):	Мелькну<л| и спря<тался| в куста<х?|
[80] :	Form 1 (-\-\-\-\ ):	's sind d'M[ei]<dli| [ju]<ng,| und fli<nk| und fro<h,|	[80] :	Form 4 (-\-\---\ ):	С ветве<й| посы<палась| роса<.|
[81] :	Form 1 (-\-\-\-\ ):	s[ie] bri<nge| we<ger| d'Su<ppe| scho<,|	[81] :	Form 4 (-\-\---\ ):	Не ты< ли,| де<вица-| краса<,|
[82] :	Form 1 (-\-\-\-\ ):	und 's A<nne| M[ei]<li| vo<rnen| a<,|	[82] :	Form 4 (-\-\---\ ):	Душе<| сказа<лася| мое<й|
[83] :	Form 1 (-\-\-\-\ ):	es la<cht| mi scho<| vo wi<tem| a<.|	[83] :	Form 4 (-\-\---\ ):	Весе<лой| пре<лестью| свое<й?|

At the moment, the Prosimetron system provides the ability to download all statistical data for a given subsample of texts sorted by date of creation, author, length, language or other parameter. If for Taranovsky it took half of his life to create such reference materials, this computer system makes it possible to process a larger amount of data over a predictable period of time. The possibility for constant rechecking, clarification, and expansion of data at any stage of the analysis minimizes the possibility for errors and their influence on the result.

This system also makes it possible to compare the rhythm of verse and prose. In this regard, the "rhythm of prose" means modeling the potentially possible rhythm of verse on the basis of the rhythmic variability of prose — that is, on the basis of a rhythmic dictionary, created through the use of the Kolmogorov model, also called the language model. Modeling, on the one hand, allows one to obtain comparable rhythmic data for two different modes of speech, that is, for the poetic and the prosaic, and on the other hand, it allows one to assess the differences that can rhythmically distinguish a poem from prose. One of the results of this work can be described as the confirmation of a previously stated hypothesis — the "prosaicization" of the Russian iambic tetrameter at the beginning of the 20th century — through analyzing a large volume of data from poems of the period. [6: 367–97]. The prosaicization of poetic rhythm in this context is defined as its gradual approaching the percentages in the language model; that is, the growing tendency to write rhythmically "non-poetic" poems.

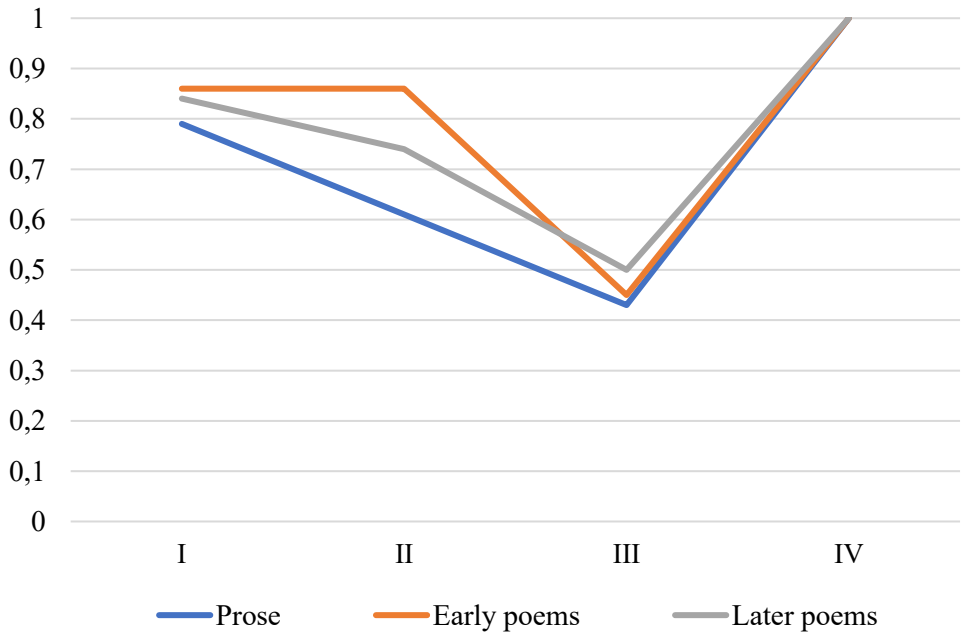


Figure 4: Rhythm of the Early and Late Poems of Boris Pasternak in Comparison with the Prose Prosody (Language Model)

Let us consider this phenomenon using as an example the rhythm of the poems of Boris Pasternak. Note the difference between the red and gray lines, which represent the rhythm of the early and late

[10] The sign "<" indicates in the Prosimetron the position of an accent, the sign "|" indicates a border between rhythmic words. Square brackets indicate that several graphemes denoting vowel sounds convey one complex phoneme, the case of diphthong, they should be calculated as a center of one syllable.

verses, respectively (see figure 4). While the rhythm of his early poetry is itself far from the typical verse of Pushkin's time and is an example of a transitional type, the rhythm of his later poetry is even farther from "poetic" rhythm and much closer to that of prose (this model is based on Pasternak's own prose). Interestingly, the model of speech based on the rhythm of occasional iambs in Pasternak's prose sharply moves away from his probability language model, and approaches the values found in his early verse and Pushkin's early poems [4: 64].

3 Conclusion

The Prosimetron-system, working with different languages, can test hypotheses about the proximity of verse rhythm to the linguistic preconditions, about the influence of one poetic tradition on another, and about the evolution of poetic forms within one of the traditions. The results presented here make it possible to see the utility of this new system for the comparative analysis of verse prosody. The system enables us to obtain data regarding the distribution of the rhythmical elements in different texts — in particular, the frequency of stresses on the metrically strong positions — as well as to simulate the rhythm of the verse according to the probability parameters for the distribution of rhythmical words. The analytical tools within this system allow for procedures of varying complexity — from the creation of dictionaries and stress profiles, to modeling and cross-language comparative analysis of rhythmic structures — and thus make it a truly comprehensive program that can be useful for both poetry theorists and linguists.

Acknowledgements

This publication was prepared as a result of research on the project no 20-04-023 supported by Research Foundation of HSE University in 2021, as a part of "5-100 Program" for leading universities of Russia.

References

[1] Bailey James. Toward a Statistical Analysis of English verse. — Lisse: Peter de Ridder Press, 1975.
[2] Kazartsev Evgeny. Computer Models of Verse Prosody. // CEUR Workshop Proceedings. — 2020. — P. 155–165.
[3] Kazartsev Evgeny. Language and Meter in the Early English, Dutch, German and Russian Iambic Verse. — Comparative Literature Studies, 2015. — Vol. 52(4).
[4] Kazartsev Evgeny. The Rhythmic Structure of «Tales of Belkin» and the Peculiarities of a Poet’s Prose. — “A Convenient Territory”: Russian Literature at the Edge of Modernity: Essays in Honor of B. P. Scherr. — Slavica, 2015.
[5] Kazartsev Evgeny. Comparative Study of Verse: Language Probability Models. — Style, 2014. — Vol. 48(2).
[6] Taranovskii K. F. (2010), Russian Iambic and Trochaic Verses. Articles about Verse [Russkie dvuslozhnye razmery. Stat’i o stikhe], Slavic Culture Languages [Iazyki slavianskoi kul’tury].

BERT for Russian news clustering

Khaustov S. V.
MTS AI
Moscow, Russia
haustovsv@gmail.com

Gorlova N. E.
MTS AI
Moscow, Russia
n.gorlova@yandex.ru

Kalmykov A. V.
MTS AI
Moscow, Russia
takiholadi@gmail.com

Kabaev A. S.
MTS AI
Moscow, Russia
askabay3@mts.ru

Abstract

This paper provides results of participation in the Russian News Clustering task within Dialogue Evaluation 2021. News clustering is a common task in the industry, and its purpose is to group news by events. We propose two methods based on BERT for news clustering, one of them shows competitive results in Dialogue 2021 evaluation. The first method uses supervised representation learning. The second one reduces the problem to binary classification.

Keywords: BERT, text clustering, news clustering, text classification, Russian text clustering, RuBERT
DOI: 10.28995/2075-7182-2021-20-385-390

BERT для кластеризации русскоязычных новостей

Хаустов С. В.
МТС ИИ
Москва, Россия
haustovsv@gmail.com

Кабаев А. С.
МТС ИИ
Москва, Россия
askabay3@mts.ru

Горлова Н. Е.
МТС ИИ
Москва, Россия
n.gorlova@yandex.ru

Калмыков А. В.
МТС ИИ
Москва, Россия
takiholadi@gmail.com

Аннотация

Статья описывает результаты участия в соревновании по кластеризации русскоязычных новостей Dialogue Evaluation 2021. Кластеризация новостей часто встречается в индустрии и основной целью является группировка новостей по событиям. Мы предложили два метода основанных на модели BERT, один из них показал конкурентный результат в соревновании. Первый метод использует обучение с учителем для получения оптимальных векторных представлений для кластеризации. Второй метод сводит задачу к бинарной классификации.

Ключевые слова: BERT, кластеризация текстов, кластеризация новостей, классификация текстов, кластеризация русских текстов, RuBERT

1 Introduction

This paper describes models used in Dialogue Evaluation 2021 competition in Russian news clustering [7]. Our team was called naergvae and took second place among thirteen participants. The competition aims to collect and compare approaches in news clustering task and the task of selecting the best headline for the resulting clusters. This paper focuses only on the first part of the competition - clustering. News clustering often occurs as a practical problem in news aggregators. The current problem aims to group news by one event into one cluster. These groups can be used for event importance estimation, news picture of the day visualization, and other tasks. The other important task is to filter similar content. In MTS AI we use this approach when developing a news service which is then used by a smart assistant. The smart assistant can read news from several sources, and it is better if the assistant does not deliver similar consecutive news about one event.

2 Related work

A natural approach to the competition task leads to review related work from the two following perspectives: choosing good news document representation(text embeddings) and choosing a good clusterization/classification scheme.

Some of the algorithms were based on unsupervised learning algorithms to cluster news articles, followed by supervised learning algorithms to classify recent articles, such as the K-Means Clustering algorithm. In [3] it is described several algorithms for news clustering, such as similarity measures. It was also found that k-means with knowledge from WordNet give better aggregate results when it comes to efficiency and the WordNet-enabled W-k means clustering algorithm significantly improves standard k-means generating. In [11] a novel clustering method was announced for an incoming stream of multilingual documents into monolingual and cross-lingual story clusters which consider a small and known number of labels.

Recent EMNLP work [14] evaluates the quality of news document embeddings and reports BERT outperforms Word2vec, GloVe, fastText, ELMo on Reuters and 20 Newsgroups datasets.

Using monolingual Russian pre-trained model RuBERT [9] is better than multilingual BERT for Russian language tasks. BERT [2] has set a new state-of-the-art performance on sentence-pair regression tasks like semantic textual similarity, which is very similar to the clustering task. However, this is computationally inefficient, and this problem is solved in Sentence-BERT [13] by deriving semantically meaningful sentence embeddings that can be compared using cosine-similarity.

Other papers using Sentence-BERT on news clustering: [6], [10]. Recently in the text clustering problem [15], Supporting Clustering with Contrastive Learning (SCCL) was proposed - a novel framework to leverage contrastive learning. Considering the landscape of the embedding clustering methods, there are no recent game-changers to our knowledge. Algorithms from a decade ago [1], [5] are still at the forefront of many near state-of-the-art results [12].

3 Method description

As for embeddings, we focus on the power of pre-trained Transformers encoders, specifically BERT, since it has been dominated on a wide range of NLP tasks.

In recent years, models based on transformers have become very popular in different NLP tasks. Our approach follows this trend. We introduce two models based on Bidirectional Encoder Representations from Transformer BERT. The first model is trying to learn good news text representation embeddings for subsequent clustering. The second one uses binary classification to classify if two news texts are from the same group or not. For both of these models, input is tokenized concatenation of headline and news text.

3.1 Embeddings using BERT triplet networks

The first approach was inspired by [13]. The idea is to train a model which will produce a fixed-size embedding representation vector for news text. These vector representations are then used for unsupervised clustering with cosine distance metric. The scheme of model inference is shown in Fig.1.

First, tokenized news text goes to the pre-trained BERT layers. Then, BERT output embeddings for each token are averaged by the pooling layer. These averaged vectors then proceed to the fully connected layer size of 768, followed by L2 normalization layer. For model training, hard triplet loss is used [8]. BERT weights do not freeze while training and initialized from pre-trained model RuBERT [9]. We trained the model on GPU Tesla V100 for ten epochs with a learning rate of 1e-6 and a batch size of 16. For the final result, the model is used to make representation embeddings for each news text, which is then used for agglomerative clustering with average linkage and cosine distance. Our model was compared on the public leaderboard with two models without fine-tuning and our model was better. To obtain vector representations of texts, the Universal Sentence Encoder (USE [4]) model and SBERT distilbert-multilingual-nli-stsb-quora-ranking model were applied, the comparison is shown in Table 1.

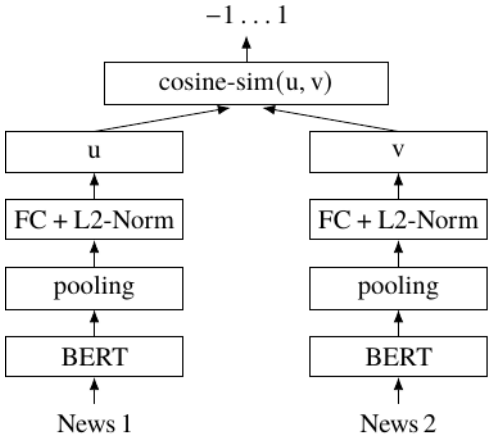


Figure 1: Model inference

Model	F1-score public LB
USE	89.4%
SBERT	88.1%
Our	91.7%

Table 1: Embeddings comparison on public leaderboard.

3.2 BERT classifier

The second approach is based on the transfer from clustering problem to news pair binary classification problem. We train a model which can classify whether the news texts pair is from the same cluster or not. To solve this problem, we used an approach similar to BERT Next Sentence Prediction problem in [2]. The general scheme of the approach is shown in Fig.2. The BERT inputs are a tokenized sequence of both news texts separated by a special token [SEP] preceded by a special token [CLS]. A token segments vector is needed for the model to understand which token belongs to which news text. The input passes through the BERT layers, which make the vector representations of each token. We use BERT pooled output which corresponds to [CLS] special token and follows this output with a softmax classification layer. The model is trained with cross-entropy loss. BERT weights do not freeze while training and are initialized from the pre-trained model RuBERT. Google bert-base-multilingual model and Sberbank AI sbert-large-nlu-ru model were also tried for weights initialization, but RuBERT performed better on the public leaderboard, the comparison is shown in Table 2. We trained a model on GPU Tesla V100 with the batch size of 8 for 6 epochs with a learning rate of 1e-5.

Model	F1-score public LB
RuBERT	96.7%
bert-base-multilingual	96.1%
sbert-large-nlu-ru	96.2%

Table 2: Initial weights comparison on the public leaderboard.

4 Results

We use the dataset with 15K training news document pairs provided by the competition organizers. This dataset is collected from the same data sources as Telegram Data Clustering Contest (2021) but specified with additional human annotations crowdsourced via Yandex.Toloka. Annotators were asked to conduct

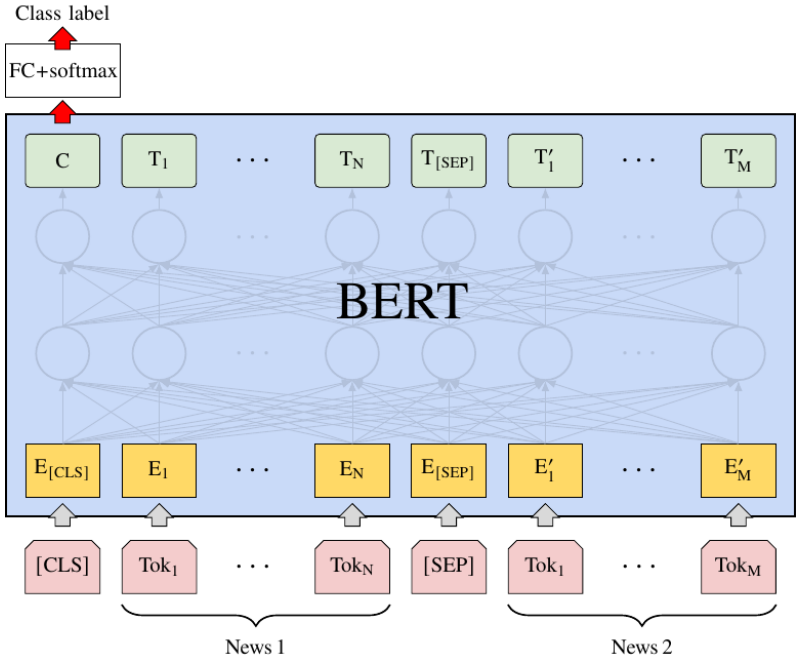


Figure 2: BERT NSP task architecture, depicted from the original paper.

a pairing task, judging pairwise news clustering relatedness. Competition leaderboard (8.5K public, 8.5K private) evaluates in terms of F1-score for positive examples (calculated only on truly positive sentence pairs, i.e., representing the same news documents).

We considered two models: BERT embeddings with agglomerative clustering and BERT classifier in Next Sentence Prediction setting. The results are reported in Table 3. As the table shows, our second approach clearly outperforms the first one. At the end of the competition, this approach took the second place. However, the advantage of our first approach is comparative computational efficiency. It represents a better practical suite for the production inference. While NSP-like Bert classifier requires at inference pairwise news documents comparison (quadratic from the expectedly huge number of news documents), BERT embeddings with clustering require only a single inference and a single neighbor search (allows efficient approximate nearest neighbor search with such tools as Facebook's Faiss, Spotify's Annoy).

We conclude that using a common trick from recommender systems practice, such as generating candidates shortlist with a less complicated model and later reranking with a more advanced one, may yield the solution that will perform well enough in general.

Model	F1-score public LB	F1-score private LB
BERT Embeddings + clusterization	91.7%	91.27%
BERT Classifier	96.7%	95.98%
Competition winner	96.9%	96.04%

Table 3: F1 score on public/private leaderboards and comparison with the competition leader.

5 Error analysis

We analyzed the results of our best-performing model (BERT classifier) on the public test data of 8.5K samples to identify the potential common causes of errors. The confusion matrix shows there were 263 misclassifications: 148 false positive and 115 false negative. Detailed investigation of a random subsample of the errors can be summarized as follows. Table 4 provides some examples.

error type	first news fragment	second news fragment
false positive, doubtful labels	Столичные отделения загс зарегистрировали за апрель около 4 тыс. браков, что на 24 % меньше, чем за аналогичный период прошлого года. Количество разводов в апреле этого года сократилось на 65 % по сравнению с апрелем 2019 года - говорится в сообщении.	В апреле нынешнего года количество разводов в российской столице уменьшилось на 65 % по сравнению с апрелем прошлого года ... В апреле в отделах загс и дворцах бракосочетания москвы поженились около четырех тысяч пар, на 24 % меньше, чем в апреле прошлого года - сообщили агентству в пресс-службе.
false positive, addition of continuation	Житель вологды дмитрий губин подал в суд на кадырова из-за комендантского часа в чечне.	Верховный суд чечни не стал рассматривать иск вологжанина дмитрия губина, в котором он пытался оспорить спецмеры из-за пандемии коронавируса, введенные рамзаном кадыровым.
false positive, same topic but different place/time	В туле покупатели устроили давку в очереди за дешевыми кастрюлями	В башкирии устроили давку из-за кастрюль за 99 рублей
false negative, usage of hypernym/hyponym relations	Существует несколько способов борьбы с сонливостью , но самые частые методы взбодриться — это кофе и физическая нагрузка . команда ученых из канады решила проверить , какой из способов самый действенный.	Ученые из лаборатории университета западного онтарио изучают, как физические упражнения могут улучшить различные показатели здоровья, один из которых — когнитивные способности.

Таблица 4: Error analysis.

We identify 45% of false positive samples to have questionable labels: even though they are labeled as a different news stories, they are strongly the same in our view. We observe another 15% of false positive also relates to the same news main event, but with the addition of continuation or person's comments; such cases are indeed errors due to the competition terms. The rest 40% of false positive samples definitely belong to different news stories but shares similar topics or context, often with different locations and dates: weather forecasts, news about financial exchange rates, announcements of football matches or their results, etc.

Using the previous errors classification terms, we find that false negative samples follow the next distribution: 33% have questionable labels, 48% are supported with additional recap/continuation/person's comment, and 19% are surely the model's errors. Interestingly we noticed several cases of the same news that classifier confused due to the usage of different hypernym/hyponym relations, specifically geographical toponyms. Another pattern that leads to errors is probably the abundance of quoted speech in the texts.

6 Conclusion

In this paper, we presented and compared two approaches for news clustering based on BERT, one of them showing competitive results in Dialogue 2021 evaluation. The first approach is supervised representation learning followed by clustering. This approach is computationally efficient and can be easily applied in real life to a large set of documents. We showed that representation learning was able to outperform unsupervised approaches from baselines. The second method with a binary classifier shows the superiority of supervised learning over unsupervised methods. This

method has shown promising results, but it can hardly be applied without modifications in real life due to performance.

References

[1] Recent Developments in Document Clustering : Rep. : TR-07-35 / Computer Science, Virginia Tech ; Executor: Nicholas O. Andrews, Edward A. Fox : 2007. — 11.

[2] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. — 2018. — 1810.04805.

[3] Bouras Christos, Tsogkas Vassilis. A clustering technique for news articles using WordNet // Knowledge-Based Systems. — 2012. — Vol. 36. — P. 115–128. — URL: https://www.sciencedirect.com/science/article/pii/S0950705112001864.

[4] Daniel Cer Yinfei Yang Sheng-yi Kong Nan Hua Nicole Limtiaco Rhomni St John NoahConstant Mario Guajardo-C espedes Steve Yuan Chris Tar et al. Universal Sentence Encoder. — 2018. — 1803.11175.

[5] Minaee Shervin, Kalchbrenner Nal, Cambria Erik et al. Deep Learning Based Text Classification: A Comprehensive Review. — 2021. — 2004.03705.

[6] Saravanakumar Kailash Karthik, Ballesteros Miguel, Chandrasekaran Muthu Kumar, McKeown Kathleen. Event-Driven News Stream Clustering using Entity-Aware Contextual Embeddings. — 2021. — 2101.11059.

[7] Gusev Ilya; Smurov Ivan. Russian News Clustering and Headline Selection Shared Task // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”. — 2021.

[8] Hermans Alexander, Beyer Lucas, Leibe Bastian. In Defense of the Triplet Loss for Person Re-Identification // CoRR. — 2017. — Vol. abs/1703.07737. — 1703.07737.

[9] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. — 2019. — 1905.07213.

[10] Linger Mathis, Hajaiej Mhamed. Batch Clustering for Multilingual News Streaming. — 2020. — 2004.08123.

[11] Miranda Sebastiao, Znotins Arturs, Cohen Shay B., Barzdins Guntis. Multilingual Clustering of Streaming News. — 2018. — 1809.00540.

[12] Pugachev Leonid, Burtsev Mikhail. Short Text Clustering with Transformers. — 2021. — 2102.00541.

[13] Reimers Nils, Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. — 2019. — 1908.10084.

[14] Sia Suzanna, Dalmia Ayush, Mielke Sabrina J. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! — 2020. — 2004.14914.

[15] Zhang Dejiao, Nan Feng, Wei Xiaokai et al. Supporting Clustering with Contrastive Learning. — 2021. — 2103.12953.

LowResourceEval-2021: a shared task on speech processing for low-resource languages

Elena Klyachko
HSE University
RAS Iling

Daniil Grebenkin
NSU
NSU SDAML

Daria Nosenko
NSU
NSU SDAML

Oleg Serikov
HSE University
DeepPavlov, MIPT

Abstract

This paper describes the results of the first shared task on speech processing for low-resource languages of Russia. Speech processing tasks are notoriously data-consuming. The aim of the shared task was to evaluate the performance of state-of-the-art models on low-resource language data as well as draw the attention of experts to field linguistics data (using Lingovodoc project data). The tasks included language identification and IPA transcription, with three teams participating in them. The paper also provides a description for the datasets as well as an analysis of the participants' solutions. The datasets created as a result of the shared task can be used in other tasks to enhance speech processing and help develop modern NLP tools for both speech communities and field linguists. **Keywords:** automatic speech recognition, language identification, minority languages, low-resource languages

DOI: 10.28995/2075-7182-2021-20-391-402

LowResourceEval-2021: дорожка по обработке речи для малоресурсных языков

Е. Клячко НИУ ВШЭ ИЯз РАН

Д. Гребенкин НГУ ЛАПДиМО

Д. Носенко НГУ ЛАПДиМО

О. Сериков НИУ ВШЭ DeepPavlov МФТИ

Аннотация

В статье описываются результаты первого соревнования по обработке речи для малоресурсных языков России. Задания по обработке речи, как правило, требуют больших объемов данных. Задачей соревнования было оценить качество работы современных моделей на данных малоресурсных языков, а также привлечь внимание экспертов к полевым данным (на примере данных проекта Lingvodoc). Задачи соревнования включали идентификацию языка и транскрипцию в МФА. В соревновании участвовали три команды. В статье описываются наборы данных, подготовленные в рамках соревнования, а также анализируются решения участников. Наборы данных могут переиспользоваться для улучшения обработки речи и развития инструментов NLP для языковых сообществ и лингвистов. **Ключевые слова:** автоматическая обработка речи, идентификация языка, малые языки, малоресурсные языки

1 Introduction

The paper describes the results of the first shared task on speech processing for low-resources languages of Russia.

Speech processing tasks are notoriously data-consuming. However, for most of the world's languages little spoken data such as news collections or audiobooks is available, not to speak of manually curated collections. Nevertheless, there are so-called field linguistics datasets, e. g. Paradisec[1], DOBES[2], ELAR[3], Lingvodoc[4]. These datasets have been created primarily for the purpose of language documentation and are often used in linguistic typology or for dialectological and historical linguistics studies. A

[1] https://www.paradisec.org.au
[2] https://dobes.mpi.nl
[3] https://elar.soas.ac.uk
[4] http://lingvodoc.ispras.ru

Shared task	Language	Training set (hours)	Test set (hours)	Number of teams
Interspeech-2018	Tamil	45	4.2	14
Interspeech-2018	Telugu	45	4.2	18
Interspeech-2018	Gujarati	45	5	18
GermEval-2020	Swiss German	70	4	3
Interspeech-2020	non-native English	≈51	≈2.5	7—9 (different tracks)
Sigtyp-2021	16 languages	≈5.5 per language	≈0.7 per language	3

Table 1: Other shared tasks

number of issues, namely, unstable recording quality (with background noise such as dogs barking or cars passing by) as well as vague licensing conditions and non-standard annotation, make field recordings largely unknown and unpopular within the speech processing community. However, these resources are often rich in dialect, age and gender variation and thoroughly annotated by language experts. We can therefore hope that they can be used by the NLP community, too. It is also worth noting that language documentation tasks, being crucial for both theorizing about languages and language revitalization, involve a lot of tedious annotation effort, making it even more important to develop automatic annotation tools. We organized a shared task on low resource speech processing[5], which was active from January to March 2021 and had the following goals:

1. evaluate the quality of modern speech processing methods on field data collections;
2. create a field recording dataset for ASR;
3. promote field data collections and low-resource language data among speech processing experts.

2 Related work

2.1 Other shared tasks on low-resource speech processing

Several competitions in various speech processing tasks have been organized recently in low-resource settings. However, the definition of what "low-resource" means varies from task to task. In 2018, Microsoft organized a shared task in low-resource automatic speech recognition ([6]), releasing data for Telugu, Tamil, and Gujarati, which was provided by Speechocean and Microsoft itself. In 2020, the participants of the GermEval-2020 shared task([13]) had to build a speech-to-text model for Swiss German, using recordings made in the parliament of Bern. Another low-resource speech processing task of 2020 was a challenging automatic speech recognition task for non-native children's speech ([15]), where records of Italian students speaking English were used.

In 2021, SIGTYP has organized a shared task on predicting language IDs (name, genus, and family) from speech[6], which is described in detail in [18]. In contrast with our task, SIGTYP-2021 involved a greater number of typologically diverse languages coming from a greater number of families, though limiting the participants to the language identification task only. They mostly used CMU Wilderness data, which is based on the sounding bible collection ([4]), for the training data as well as Common Voice[7] and OpenSLR[8] for the validation and test data. Moreover, some field data from the Paradisec is also used. The diversity of the data sources is meant to check the robustness of the participants' models. Three teams took part in SIGTYP-2021, with two of them performing better than the baseline. The winning team (Lipsia, [5]) transformed the MFCCs distributed by the organizers into spectrograms and then applied a ResNet-50 CNN based model to them. Another team was NTR ([2]), which used a solution similar to the one submitted to our shared task (see the description for the NTR system below). Finally, the Anlirika ([1]) system combines convolutional and LSTM layers in their approach

The details for the above-mentioned low-resource tasks are summarized in the table below (table 1).

[5] https://lowresource-lang-eval.github.io/content/shared_tasks/asr2021_en.html
[6] https://sigtyp.github.io/st2021.html
[7] https://commonvoice.mozilla.org
[8] https://openslr.org

A low-resource end-to-end speech translation task has been organized in 2021, focusing on two Swahili varieties and French and English[9]. However, its results will only be available later this year.

2.2 Using field linguistic data in speech processing tasks

Most papers on speech recognition for field linguistics datasets are aimed at facilitating language documentation itself, e. g. [12] (for Japhug) or [16](for Samoyedic languages). Moreover, tools for training speech recognition on field datasets have been developed, for instance Persephone ([21]) followed by Elpis ([19]). In [10], Gina-Anne Levow endorses using endangered language data in shared tasks. In [11], the authors show an exemplar case of using field data from the ELAR archive to create datasets for the speaker diarization and identification tasks. The paper also deals with the dual use of linguistic data and its potential consequences.

3 Shared task description

We offered three tasks: number of speakers detection, language identification, and automatic transcription. However, only two latter tasks were actually completed by three teams. The small number of teams is actually comparable to the other low-resource ASR shared tasks, which is perhaps due to the task difficulty and absence of public datasets. The shared task was hosted at the CodaLab automatic scoring platform[10]. The evaluation script is available online[11]

3.1 Number of speakers detection

The track was not completed due to the lack of participants. The aim of the track was to identify the number of speakers in a short recording. The track is crucial for field data processing as recordings often contain dialogues between a linguist and a language consultant. The dataset recordings were thus annotated with a corresponding number of speakers. The participants were to predict the number.

3.2 Language identification task

The shared task participants had to identify the language spoken, its genus, and the language family. The test dataset included surprise languages belonging to the same genera as the previously seen languages. The participants had to classify them as “unknown” (X). We scored the accuracy of classification across all the fields.

3.3 Automatic IPA transcription

The participants were to automatically transcribe speech using IPA. As in the language identification task, the test set included both previously seen and previously unseen data. We scored the length-normalized CER of the transcriptions.

4 Evaluation datasets

The datasets were based on the Lingvodoc platform ([7]), developed at the Institute of System Programming, RAS. The voiced data was compiled by linguists from Russian scientific institutions and processed in a unified way. The project focuses on collecting wordlists and corpora in various dialects of (predominantly) Uralic and Altaic languages, which are usually used for dialectological and historical linguistics studies.

4.1 Dataset preparation

Dataset preparation involved both scraping Lingvodoc and additionally annotating it. Lingvodoc is a joint effort of multiple teams so it is not surprising that the data can suffer from variation in annotation approaches, which has to deal with how wordlists are collected in language documentation projects. It is often a case that a linguist first pronounces the stimulus (a word or a phrase) in an auxiliary language

[9] https://iwslt.org/2021/low-resource
[10] https://competitions.codalab.org/competitions/30008
[11] https://github.com/lowresource-lang-eval/asr_evaluation_scripts

(Russian in case of Lingvodoc). The language consultant then pronounces the translation of the stimulus in their native tongue, sometimes repeating it. These repetitions are sometimes accompanied by a linguist asking the native speaker to pronounce the word again. When uploaded to Lingvodoc, these repetitions are not always split. The stimuli pronounced by the linguist are also not cut off in some cases. Both decisions (whether to split the recording into several repetitions and whether to include the stimuli or not) are usually made by the particular team uploading data to Lingvodoc. This variation in approaches is not crucial for manual processing but can hamper automatic processing. We therefore decided to specify whether there can potentially be repetitions or Russian stimuli in the data. The annotation is also available in the test dataset as two additional columns. We also checked if the transcriptions were IPA-valid using `ipapy`[12], excluding non-IPA transcriptions and normalizing some standard ways of transcribing which are not genuinely IPA, e.g. ɛ́:ʏ-ak became ɛ́ːɣak

The datasets in the converted format can be found online[13].

4.2 Dataset statistics

Dataset was split into two subsamples, *Train* and *Test* respectively. The tasks were evaluated on the *Test* subsample, which contained both previously seen and surprise languages. The surpise languages were chosen on the following grounds: all the families and genera from the datasets had to be represented.

While there is some difference in recordings lengths across the language groups 1, no specific handling has been applied regarding these distributions.

5 Participants and results

Three teams took part in the shared task, choosing either the classification track (team NTR/TSU)(see [3]in this volume) or the transcription track (teams DG and DN).

5.1 System description

5.1.1 NTR/TSU

NTR/TSU uses a convolutional neural network with a self-attentive pooling layer for the classification task. The input for the network are mel-frequency spectral coefficients calculated from the original audio files. The architecture of the network is QuartzNet ASR. They also used several augmentation techniques, namely, shifting samples in range (-5ms; +5ms), SpecAugment, and adding background noise to the audio files.

5.1.2 Team DG

Daniil Grebenkin, one of the authors of this paper, contributed to this model. The experiment included the following stages:

1. Data preprocessing. The audio files had different sample rates and numbers of channels. Therefore, DG converted the files into mono and set them to the same sample rate (16000 kHz) using sox[14].
2. Creating a multilingual acoustic model for getting the transcriptions from lattices' files. DG used Kaldi ASR ([9]), which is a modular system allowing to add new data to either the language model or the acoustic model at a time without changing the other model. They applied a multilingual model trained on VoxForge[15] corpus to make transcriptions of the train dataset. Then, they trained a language module with these new words and new transcriptions. Finally, they created a new version of a multilingual model with an upgraded language module.
3. Getting the phoneme sequences of the competition train set utterances with the multilingual model. At this stage, they decoded the competition test set to get a transcription for each utterance. They used the epitran tool ([14]) to make a dictionary for the language module. The multilingual VoxForge corpus contains various languages and is rich in phonetic variation. IPA makes it possible to show differences in pronunciation for every language from VoxForge data whereas epitran has support for

[12] `https://github.com/pettarin/ipapy`
[13] `https://lowresource-lang-eval.github.io/content/data/index_data_asr_en.html`
[14] `http://sox.sourceforge.net`
[15] `http://www.voxforge.org`

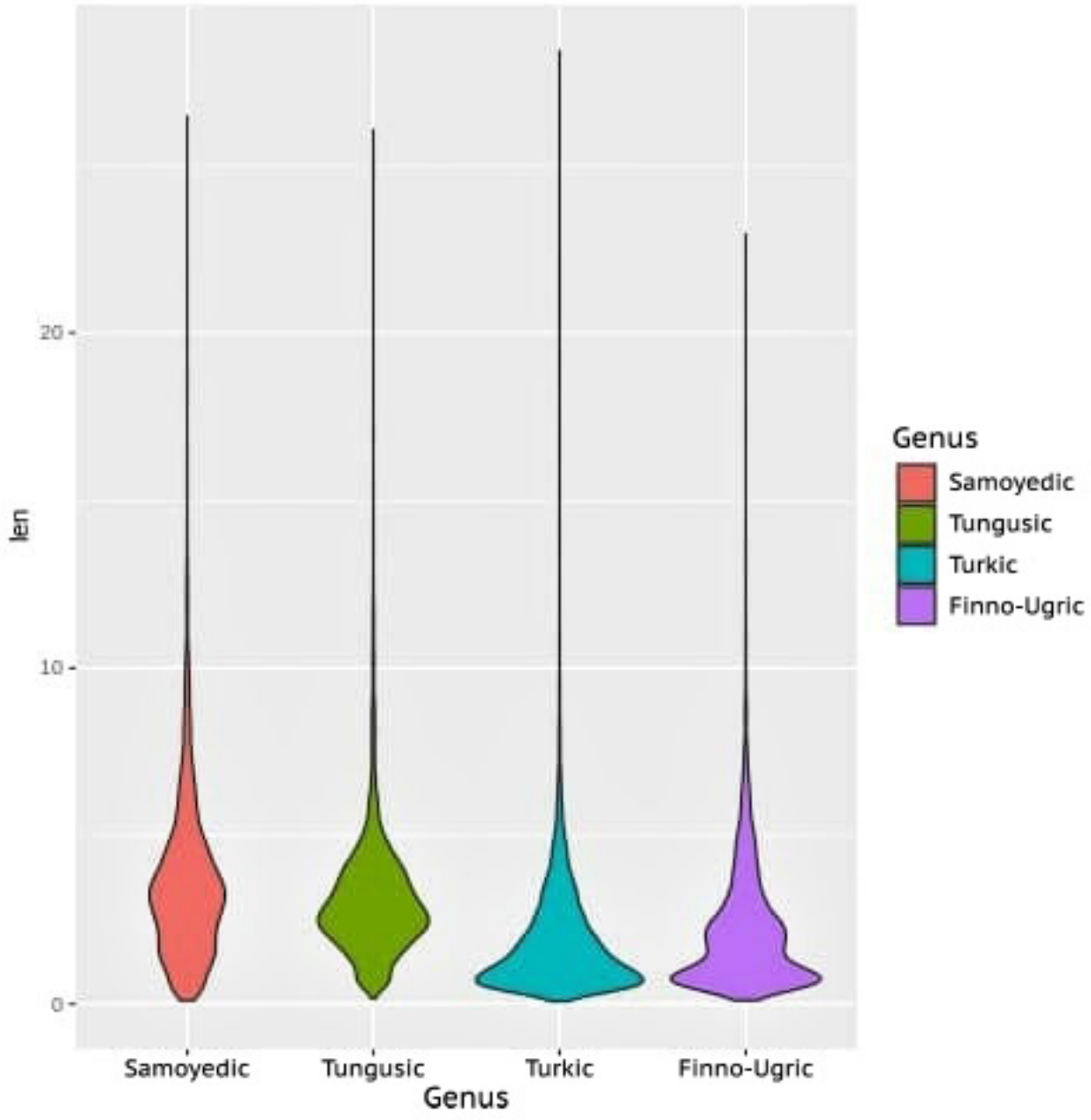


Figure 1: Recordings lenghts (in seconds) distribution across language groups

almost every language of this corpus. The result of the decoding process were lattices which helped get the phoneme sequences of each utterance. The algorithm was as follows:
- compute the best path through lattices and write them out as FSTs;
- store the lattices in archives containing transcriptions, alignments, and acoustic and LM costs;
- convert model-level alignments to phoneme sequences
- convert the phoneme sequences from the X-SAMPA format ([20]) to IPA format with an open-source converter xsampa [16]

4. Creating a new version of the multilingual model. The new transcriptions and the new words were added to the existing dictionary and lexicon files. Afterwards, they created a new language module and a new multilingual model with Kaldi tools.
5. Processing the test set. At this stage, they decoded the test set, getting phoneme sequences, which they then converted to the IPA format.

Some of the audio samples were not recognized by the model, there were $4,2\%$ of lines completed with «none» in final .tsv file which was used for evaluation.

5.1.3 Team DN

Daria Nosenko, one of the authors of this paper, contributed to this model. Team DN solution is founded on an end-to-end neural model for speech recognition QuartzNet([17]) based on Jasper([8]). QuartzNet model uses separable convolutions and is smaller than all other competing models. Team DN used TensorFlow-based NVIDIA OpenSeq2Seq toolkit[17] for their experiments with QuartzNet. The main features of this framework are modular architecture that allows assembling of new models from available components and fast Horovod-based distributed training supporting both multi-GPU and multi-node modes. They chose the multilingual VoxForge corpus for model training. The model training was performed using Horovod on 3 GPUs. The experiment included the following stages:

1. Train dataset preprocessing. All audio files with their annotations were combined into a single multilingual dataset. The .csv file for model training was generated using that dataset. Its rows have the following format: audio file name, audio file size, annotation. The file name contains its absolute path, the speaker folder name and the audio file name, which contains the language tag. If the file name did not contain the language tag, then it was added. The vocabulary for model training was generated from annotations of the entire dataset using Python ”set” function. Then VoxForge audio files were converted into mono and set to the same sample rate (16000 kHz) using SoundFile[18]. The multilingual corpus was split into train, validation and test in 80:10:10 ratio in such a way that subsets did not overlap by speakers (i.e., one speaker should not be included in any two subsets at the same time).
2. Training QuartzNet model on the VoxForge dataset. The model was trained on 70 epochs.
3. Predicting annotations for the Dialog test dataset using the model from the previous step.

Finally, there were 412 audio files that were not recognized by the model ($3,9\%$ of the total number).

6 System results

6.1 Language Identification task

The only team to submit the LId task results was the NTR team. Their submission accuracy is outlined in the table 2 along with the random baseline scores. Overall submission confusion matrix is attached in the Appendix A .

While for frequent languages the accuracy is slightly better, there is no significant correlation between how much a language is represented in the data in the data and its identification accuracy. While showing some maybe interesting granular patterns, the confusion matrix is hard to typologically analyze. Unseen language identification is shown to be especially hard.

[16] https://github.com/dohliam/xsampa
[17] https://nvidia.github.io/OpenSeq2Seq/html/index.html
[18] `https://pysoundfile.readthedocs.io/en/latest`

Team	LId	GId	FId
NTR	**0.06**	**0.34**	0.61
baseline	0.01	0.22	**0.82**

Table 2: Results for Language Identification task which consisted of Language Identification (**LId**), language Group Identification (**GId**), language Family Identification (**FId**)

Team	Total test set (files)	Not recognized (files)	normalized CER
DG	10445	438	1.0828
DN	10445	412	1.572267

Table 3: Results for Task2

6.2 ASR task

Despite beating the baseline system, both submissions tend to provide much longer sequences of phonemes than expected. A closely-read analysis discovers that systems prediction performance drops while going from the beginning to the end of the recording with first phonemes usually being nearly guessed. The results of the both teams are summarized in the table below (3).

While the task was formulated using the IPA alphabet, both submissions' alphabets were different due to system design. This complicated the analysis of systems and resulted in significant loss of the evaluation metric.

7 Conclusion

In this paper, we present the results of the first shared task on ASR and speech-based language identification and categorization for the languages of Russia. As a result of the shared task, we prepared several datasets for language classification, transcription, and speaker number detection, for the first time for the languages in question. The participating teams experimented in performing various speech processing tasks for the languages which lack modern ASR technology tools, using state-of-the-art models. When analyzing the results, we also explored the limitations of the systems, which can help improve them

Acknowledgements

We would like to thank the authors of Lingvodoc for allowing us to use their priceless data, and the participants of the shared task for their comment and suggestions.

The work of Elena Klyachko was supported by a grant of the Russian Science Foundation, Project 20-012-00520 (Dynamics of the development of the language situation in local groups of indigenous peoples of Siberia and Russian Far East based on linguistic biographies).

References

[1] Anlirika: An LSTM–CNN Flow Twister for Spoken Language Identification / Andreas Scherbakov, Liam Whittle, Ritesh Kumar et al. // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — 2021. — P. 145–148.

[2] Bedyakin Roman, Mikhaylovskiy Nikolay. Language ID Prediction from Speech Using Self-Attentive Pooling and 1D-Convolutions // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — 2021.

[3] Bedyakin Roman, Mikhaylovskiy Nikolay. Low-Resource Spoken Language Identification Using Self-Attentive Pooling and Deep 1D Time-Channel Separable Convolutions // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021", Moscow, June 16–19, 2021. — 2021.

[4] Black Alan W. CMU Wilderness Multilingual Speech Dataset // ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — 2019. — P. 5971–5975.

[5] Celano Giuseppe GA. A ResNet-50-based Convolutional Neural Network Model for Language ID Identification from Speech Recordings // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — 2021.

[6] Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages / Brij Mohan Lal Srivastava, Sunayana Sitaram, Kalika Bali et al. // SLTU. — 2018. — August.

[7] J.V. Normanskaya O.D. Borisenko. Dictionaries on Samoyedic languages and LingvoDoc software system for collaborative work on dictionaries and online publishing // NORDSCI 2018 Conference Proceeedings. — Vol. 1. — 2018. — P. 313–337.

[8] Jasper: An End-to-End Convolutional Neural Acoustic Model / Jason Li, Vitaly Lavrukhin, Boris Ginsburg et al. // Interspeech 2019. — 2019. — Sep. — Access mode: http://dx.doi.org/10.21437/interspeech.2019-1819.

[9] The Kaldi Speech Recognition Toolkit / Daniel Povey, Arnab Ghoshal, Gilles Boulianne et al. // IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. — Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society, 2011. — Dec. — IEEE Catalog No.: CFP11SRW-USB.

[10] Levow Gina-Anne. Promoting Language Technology for Endangered Languages with Shared Tasks // Proceedings of the 1st International Conference on Language Technologies for All. — Paris, France : European Language Resources Association (ELRA), 2019. — December. — P. 116–119. — Access mode: https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.30.pdf.

[11] Levow Gina-Anne, Ahn Emily P, Bender Emily M. Developing a Shared Task for Speech Processing on Endangered Languages // Proceedings of the Workshop on Computational Methods for Endangered Languages. — Vol. 1. — 2021. — P. 96–106.

[12] Macaire Cécile. Alignement temporel entre transcriptions et audio de données de langue japhug // 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT) / CNRS. — 2020. — P. 9–22. — Access mode: https://hal.archives-ouvertes.fr/hal-03066031/document#page=15.

[13] Michel Plüss Lukas Neukom Manfred Vogel. GermEval 2020 Task 4: Low-Resource Speech-to-Text // CEUR-WS.org. — 2020. — Access mode: http://ceur-ws.org/Vol-2624/germeval-task4-paper1.pdf.

[14] Mortensen David R, Dalmia Siddharth, Littell Patrick. Epitran: Precision G2P for many languages // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018.

[15] Overview of the Interspeech TLT2020 Shared Task on ASR for Non-Native Children's Speech / Roberto Gretter, Marco Matassoni, Daniele Falavigna et al. // Proc. Interspeech 2020. — 2020. — P. 245–249.

[16] Partanen Niko, Hämäläinen Mika, Klooster Tiina. Speech Recognition for Endangered and Extinct Samoyedic languages // arXiv preprint arXiv:2012.05331. — 2020. — Access mode: https://arxiv.org/ftp/arxiv/papers/2012/2012.05331.pdf.

[17] Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions / Samuel Kriman, Stanislav Beliaev, Boris Ginsburg et al. // ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — 2020. — May. — Access mode: http://dx.doi.org/10.1109/ICASSP40776.2020.9053889.

[18] SIGTYP 2021 shared task: Robust spoken language identification / Elizabeth Salesky, Badr M Abdullah, Sabrina Mielke et al. // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — 2021. — P. 122–129.

[19] User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis / Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer et al. // Proceedings of the 4th Workshop on Computational Methods for Endangered Languages. — 2021.

[20] Wells John C. Computer-coding the IPA: a proposed extension of SAMPA // Revised draft. — 1995. — Vol. 4, no. 28. — P. 1995.

[21] Wisniewski Guillaume, Michaud Alexis, Guillaume Séverine. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? // 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop / European Language Resources Association (ELRA). — 2020. — P. 306–315. — Access mode: `https://arxiv.org/ftp/arxiv/papers/2012/2012.05331.pdf`.

A Appendix

Family	Group	Language	Count
Altaic	Manchu-Tungus	evn	2529
Altaic	Manchu-Tungus	oac	1011
Altaic	Manchu-Tungus	ude	696
Altaic	Manchu-Tungus	ulc	576
Altaic	Turkic	alt-tub	2519
Altaic	Turkic	bak	1044
Altaic	Turkic	sah	388
Altaic	Turkic	tat	273
Altaic	Turkic	tyv	5961
Uralic	Samoyedic	enh	930
Uralic	Samoyedic	nio	725
Uralic	Samoyedic	sel	3022
Uralic	Samoyedic	yrk-for	76
Uralic	Finno-Ugric	kca	661
Uralic	Finno-Ugric	koi-yzv	461
Uralic	Finno-Ugric	kom	318
Uralic	Finno-Ugric	krl	771
Uralic	Finno-Ugric	mdf	103
Uralic	Finno-Ugric	mhr	93
Uralic	Finno-Ugric	mns	775
Uralic	Finno-Ugric	mrj	330
Uralic	Finno-Ugric	sjd	220
Uralic	Finno-Ugric	sms	549

Table 1: **Train** subsample statistics, number of utterances for each language is counted

Family	Group	Language	Count
Altaic	Manchu-Tungus	evn	19
Altaic	Manchu-Tungus	gld	390
Altaic	Manchu-Tungus	neg	360
Altaic	Manchu-Tungus	ude	2060
Altaic	Manchu-Tungus	ulc	17
Altaic	Turkic	alt	671
Altaic	Turkic	alt-tel	21
Altaic	Turkic	alt-tlg	16
Altaic	Turkic	atv-c	12
Altaic	Turkic	bak	296
Altaic	Turkic	chv	1826
Altaic	Turkic	cjs	25
Altaic	Turkic	clw	364
Altaic	Turkic	dlg	19
Altaic	Turkic	kim	229
Altaic	Turkic	kum	3
Altaic	Turkic	sah	483
Altaic	Turkic	tat	34
Altaic	Turkic	tyv	1269
Altaic	Turkic	uig	531
Uralic	Samoyedic	enf	442
Uralic	Samoyedic	enh	46
Uralic	Samoyedic	nio	47
Uralic	Samoyedic	yrk-ntu	93
Uralic	Finno-Ugric	fin	97
Uralic	Finno-Ugric	koi	265
Uralic	Finno-Ugric	kom	9
Uralic	Finno-Ugric	krl	264
Uralic	Finno-Ugric	mhr	15
Uralic	Finno-Ugric	mrj	139
Uralic	Finno-Ugric	myv	15
Uralic	Finno-Ugric	sms	4
Uralic	Finno-Ugric	udm	132
Uralic	Finno-Ugric	vot	232

Table 2: **Test** subsample statistics, number of utterances for each language is counted

Family	Group	Language	Count
Altaic	Manchu-Tungus	evn	2548
Altaic	Manchu-Tungus	gld	390
Altaic	Manchu-Tungus	neg	360
Altaic	Manchu-Tungus	oac	1011
Altaic	Manchu-Tungus	ude	2756
Altaic	Manchu-Tungus	ulc	593
Altaic	Turkic	alt	671
Altaic	Turkic	alt-tel	21
Altaic	Turkic	alt-tlg	16
Altaic	Turkic	alt-tub	2519
Altaic	Turkic	atv-c	12
Altaic	Turkic	bak	1340
Altaic	Turkic	chv	1826
Altaic	Turkic	cjs	25
Altaic	Turkic	clw	364
Altaic	Turkic	dlg	19
Altaic	Turkic	kim	229
Altaic	Turkic	kum	3
Altaic	Turkic	sah	871
Altaic	Turkic	tat	307
Altaic	Turkic	tyv	7230
Altaic	Turkic	uig	531
Uralic	Samoyedic	enf	442
Uralic	Samoyedic	enh	976
Uralic	Samoyedic	nio	772
Uralic	Samoyedic	sel	3022
Uralic	Samoyedic	yrk-for	76
Uralic	Samoyedic	yrk-ntu	93
Uralic	Finno-Ugric	fin	97
Uralic	Finno-Ugric	kca	661
Uralic	Finno-Ugric	koi	265
Uralic	Finno-Ugric	koi-yzv	461
Uralic	Finno-Ugric	kom	327
Uralic	Finno-Ugric	krl	1035
Uralic	Finno-Ugric	mdf	103
Uralic	Finno-Ugric	mhr	108
Uralic	Finno-Ugric	mns	775
Uralic	Finno-Ugric	mrj	469
Uralic	Finno-Ugric	myv	15
Uralic	Finno-Ugric	sjd	220
Uralic	Finno-Ugric	sms	553
Uralic	Finno-Ugric	udm	132
Uralic	Finno-Ugric	vot	232

Table 3: **Overall** dataset statistics, number of utterances for each language is counted

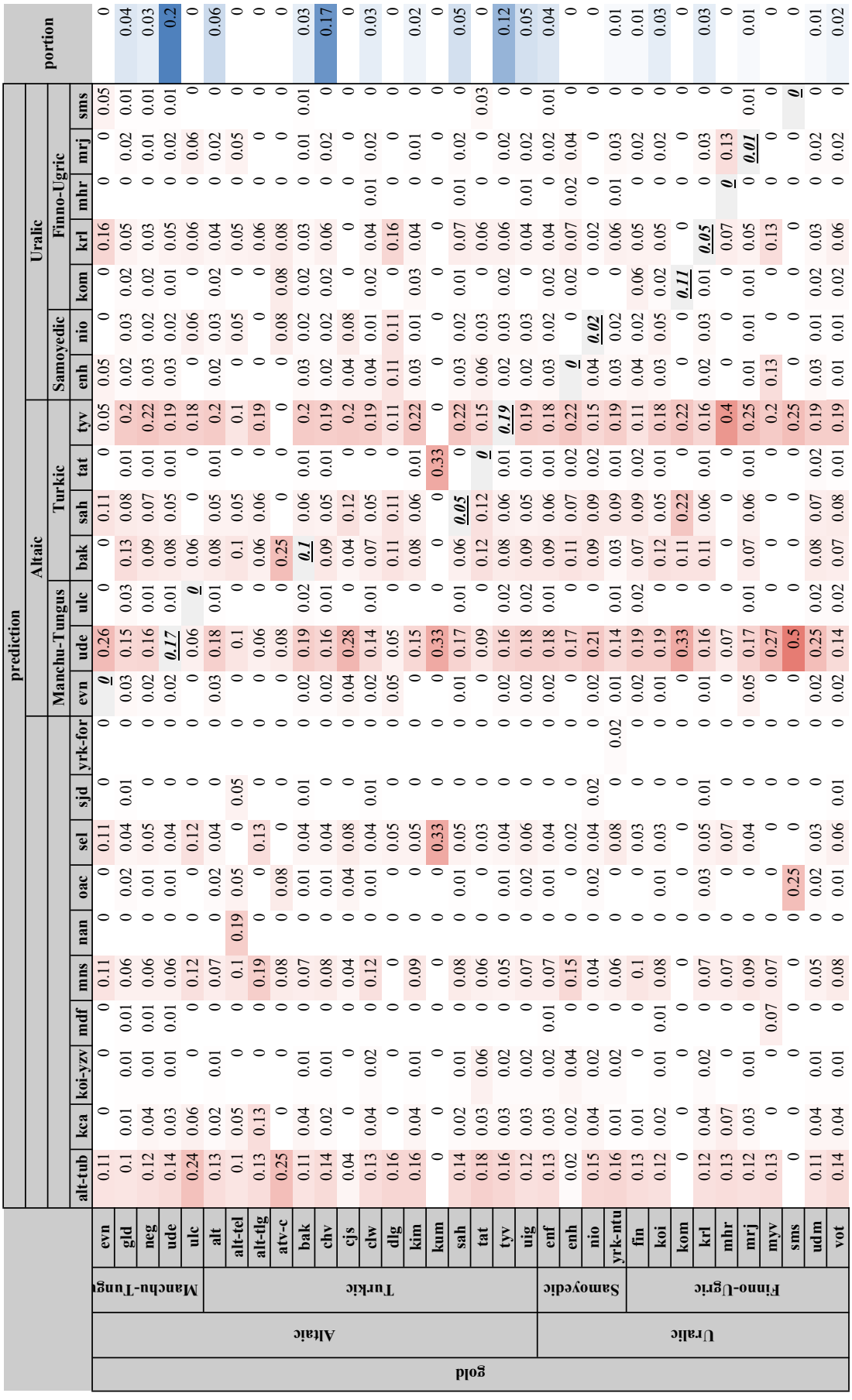


Figure 1: confusion matrix of the team **NTR** submission

The intonation of *yes* and *no* in an archaic Russian dialect

Knyazev S. V.
Vinogradov Russian Language Institute,
Russian Academy of Sciences,
Moscow, Russia
svknia@gmail.com

Pronina M. K.
Universitat Pompeu Fabra,
Barcelona, Spain
mariia.pronina@upf.edu

Abstract

The present paper analyzes the intonation of pragmatic particles *da* "yes" and *net* "no" found in the spontaneous dialogue speech corpus of a Northern Russian dialect, in which each word bears a pitch accent. Intonation that marks such particles sounds unusual for speakers of Standard Russian and is perceived by them as blunt and impolite. The main aim was to find a consistent pattern explaining the distribution of falling and rising pitch accents on such particles in a dialect of Vaduga (Arkhangelsk region). We tested three hypotheses that can account for this distribution: (a) semantic explanation (the type of pitch accent depends on the semantics of the very particle); (b) communicative explanation (it depends on the communicative function of the preceding utterance, that is, whether it is a question or not); (c) phonetic explanation (it depends on the pitch accent of the preceding utterance). A total of 240 utterances from 3 speakers were analyzed. Results showed that the semantics of the particle is not a relevant factor, while the communicative type and the pitch accent of the preceding utterance are significant predictors of the pitch accent that marks the particle, with the latter better explained the data. We propose that when analyzing the intonation of a dialect, semantic interpretation of the intonational constructions of the standard dialect should not be taken into account. Moreover, we suggest that a new approach of collecting prosodic data with elderly people while controlling for pragmatic context is needed

Keywords: Russian language, dialect, intonation, dialogue, politeness, *yes* and *no*
DOI: 10.28995/2075-7182-2021-20-403-412

Интонация *да* и *нет* в архаическом говоре с пословным тональным оформлением

Князев С. В.
Институт русского языка
им. В. В. Виноградова РАН,
Москва, Россия
svknia@gmail.com

Пронина М. К.
Университет Помпеу Фабра,
Барселона, Испания
mariia.pronina@upf.edu

1 Введение: русские говоры с пословным тональным оформлением

В процессе коммуникации значительная часть информации транслируется говорящим при помощи фразовой просодии, в некоторых случаях даже бóльшая, чем та часть, что передается при помощи лексических значений слов. Так, например, высказывание *Закройте дверь!*, оформленное при помощи восходящего акцента на слове *закройте* является (и воспринимается) в современном русском литературном языке (СРЛЯ) гораздо более вежливым, чем высказывание *Закройте, пожалуйста, дверь!*, оформленное при помощи нисходящего акцента на том же слове. Однако система интонационных значений в разных языках и в разных диалектах одного языка может существенно различаться. Данное исследование посвящено анализу тональных акцентов в одном из архаических русских говоров.

В июле 1987 г. один из авторов в составе диалектологической экспедиции МГУ им. М. В. Ломоносова и Института русского языка им. В. В. Виноградова РАН посетил с целью сбора

диалектного материала деревню Вадюга Верхнетоемского района Архангельской области в верховье реки Пинеги[1].

Одной из самых ярких диалектных особенностей верхнепинежских говоров является их специфическая интонация [10], которая в значительной степени связана с особым типом использования тональных просодических средств, так называемым **пословным тональным оформлением высказывания**: «фраза состоит из ряда отрезков с восходящей интонацией, последний же отрезок характеризуется восходяще-нисходящей интонацией с более быстрым падением чем восхождением» [7: 14], «почти каждое слово во фразе получает свое мелодическое оформление» [9: 64], так что для этих говоров «характерно пословное оформление интонационного контура» [9: 78]. Тональную обособленность каждого слова в этих говорах фиксировала и Е. А. Брызгунова [3: 247, 262]. Примеры разных типов тонального оформления приведены на рис. 1 и 2: в фонетической синтагме СРЛЯ лишь одно слово выделено изменением частоты основного тона (ЧОТ), в то время как в говоре тональное изменение наблюдается на каждом фонетическом слове.

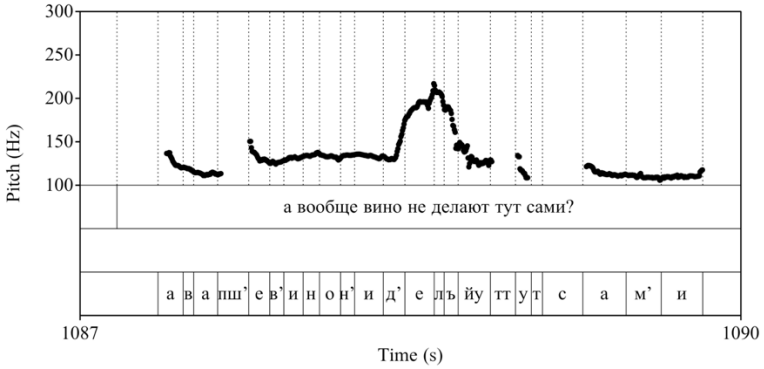


Рисунок 1: Кривая ЧОТ фразы *А вообще вино не делают тут сами?* (СРЛЯ)

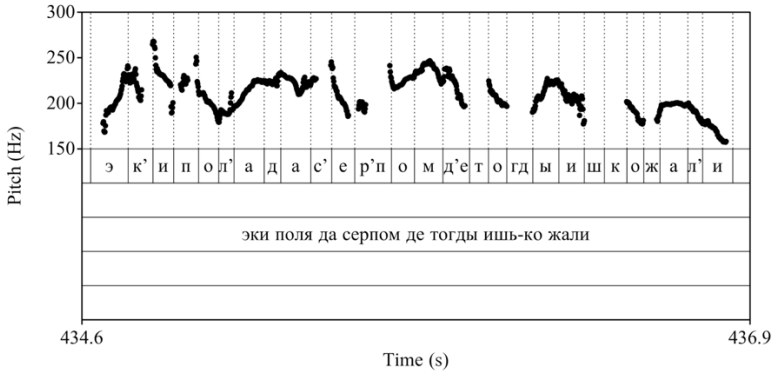


Рисунок 2: Кривая ЧОТ фразы *Эки поля да серпом де тогды ишь-ко жали* (д. Вадюга)

Кардинальные отличия диалектных систем рассматриваемого типа в области интонационного оформления высказывания от литературного языка приводят к тому, что взаимопонимание между носителями ЛЯ и диалекта может быть существенно затруднено – так, в ходе нашей экспедиции ее участники долгое время испытывали серьезные проблемы с восприятием диалектной речи, что, в целом, не свойственно ситуации нахождения носителя ЛЯ в диалектном окружении.

Другой проблемой в общении с носителями вадюжского диалекта на начальном этапе работы было, по отзывам участников, постоянное ощущение того, что носители говора в высшей степени неприветливы и недоброжелательны – основной причиной такого восприятия диалога являлась, в том числе, специфическая интонация ответных реплик носителей диалекта, близкая к

[1] В 1928-1929 гг., этот говор был исследован П. С. Кузнецовым [7]; описание его состояния в конце 1980-х гг. в сопоставлении с данными П. С. Кузнецова см. в [5].

той, что в СРЛЯ описывается как четвертая интонационная конструкция (ИК-4)[2] [4: 115]. В ИК-4 предцентровая часть произносится на среднем тоне; если центр ИК находится в конце, тон начинается с более низкой точки по сравнению со средним, затем в пределах слога ровно повышается; если есть заударная часть, то ударная произносится с понижением, а заударная – с ровным повышением тона [2: 41]. Использование ИК-4 в диалектной речи отмечалось не раз [3], в частности, – в утвердительных предложениях [10: 60-61, 63] и в ответных репликах в диалоге [13: 109]. Е. А. Брызгунова, описывая один из пинежских диалектов, отмечает: «Одной из ярких особенностей, "экзотикой" интонации говора д. Ваймуша Архангельской обл. является модальная реализация ИК-4, отмеченная в ситуации, когда говорящий что-либо доказывает, спорит, возражает» [3: 241]. В литературном языке «ИК-4 употребляется наряду с ИК-1, подчеркивая при этом противопоставление, категоричность утверждения, удивление, вызов» [4: 115].

Итак, постоянно слыша в диалогах с информантами интонацию, близкую к ИК-4 литературного языка, мы воспринимали ее как показатель ненейтрального отношения говорящего – раздражения, вызова, категоричности утверждения и даже нежелания поддерживать разговор. Однако данный тип просодического оформления высказывания встречается в говоре настолько часто, что это заставляет задуматься о том, действительно ли соответствующий интонационный контур не является эмоционально нейтральным.

2 Цели, задачи, материал и процедура исследования

Целью данного исследования, таким образом, был анализ интонационного оформления ответных диалогических реплик, основной **задачей** – поиск закономерностей в распределении нисходящего и восходящего тонального акцента в просодическом оформлении утвердительной и отрицательной частиц *да* и *нет*. Утвердительная и отрицательная частицы были выбраны в качестве материала исследования в силу того, что они являются наиболее частыми словами в существующих текстах и обладают достаточно очевидной семантикой.

Запись материала в ходе экспедиции 1987 г. производилась автором данной статьи на кассетный стереомагнитофон «Соната 213С» с использованием выносных динамических микрофонов, впоследствии записи были оцифрованы А. В. Архиповым. В настоящее время продолжается работа по созданию корпуса звучащих текстов говоров Верхнетоемского района Архангельской области, куда эти записи будут включены. В большинстве случаев тексты представляют собой беседы интервьюеров с носителями диалектов, в редких случаях – диалоги носителей говора.

Материалом для настоящего исследования служили записи, полученные от трех носителей анализируемого говора. В ходе основного эксперимента были проанализированы записи информанта ПЕМ, женщины 70 лет, который был выбран на основании того, что, во-первых, он является одним из наиболее типичных представителей анализируемого говора (в том числе и в отношении фразовой просодии), во-вторых, относится к старшему его поколению, сохраняющему диалектную основу в максимальной степени, в-третьих, от него имеются записи максимальной продолжительности (90 минут звучания, 141 тестовое слово). Дополнительная верификация полученных данных была проведена на материале реплик еще двух информантов женского пола – МИР (87 лет, 59 слов) и АПН (62 года, 40 слов).

На первом этапе исследования была проведена сплошная аннотация 2741 реплики[3] основного информанта (ПЕМ), затем были выделены контексты с частицами *да* и *нет* (общее количество – 141) и с помощью программы PRAAT осуществлен **анализ их тонального оформления**.

Полученные данные свидетельствуют о том, что тональное оформление частиц *да* и *нет* в исследуемом говоре достаточно вариативно: они могут быть оформлены следующими мелодическими контурами: восходящее движение тона [4] (см. рис. 3), ровное + восходящее (см. рис. 4), нисходяще-восходящее (см. рис. 5), восходяще-нисходящее (см. рис. 6), нисходящее (см. рис. 7), ровное + нисходящее (см. рис. 8).

[2] В действительности, этот тип интонационного оформления отличается от ИК-4 литературного языка, но в настоящей работе мы не будем останавливаться на этих различиях, данному вопросу посвящено отдельное исследование.

[3] Репликами в данном случае считаются звучащие отрезки между физическими (в том числе – дыхательными) паузами, их продолжительность колеблется в диапазоне от 0,1 до 8,5 секунд.

[4] Восходящее движение в чистом виде фиксируется только в тех случаях, когда частица включена в более широкий просодический контекст, то есть, не является отдельной фонетической синтагмой.

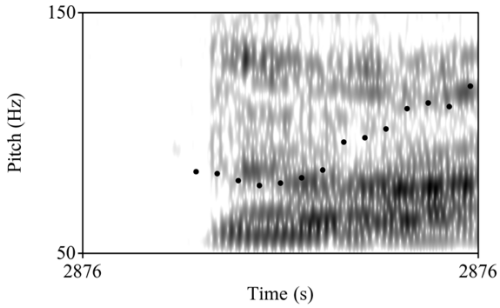


Рисунок 3: Динамическая спектрограмма и кривая ЧОТ (*Да*, восходящее движение ЧОТ)[5]

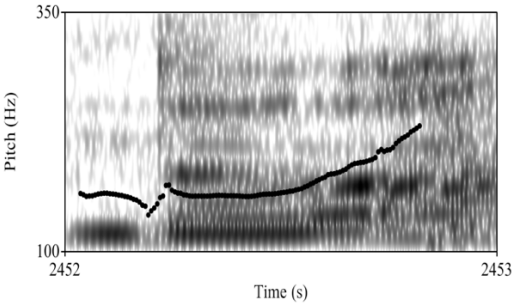


Рисунок 4: Динамическая спектрограмма и кривая ЧОТ (*Да*, ровное + восходящее движение)

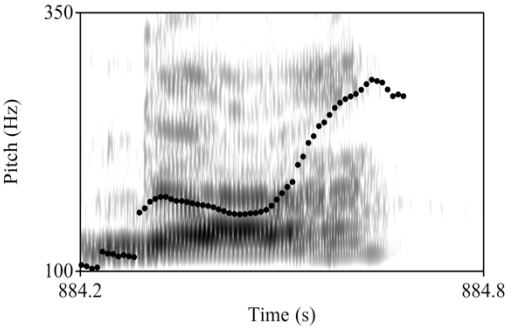


Рисунок 5: Динамическая спектрограмма и кривая ЧОТ
(*Да*, нисходяще-восходящее движение)

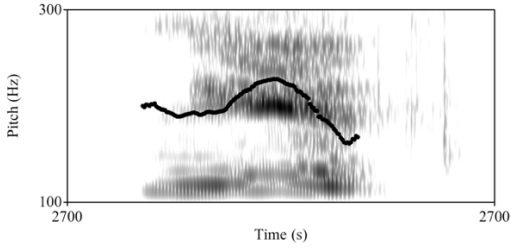


Рисунок 6: Динамическая спектрограмма и кривая ЧОТ
(*Нет*, восходящее + нисходящее движение)

[5] Незначительное понижение тона в начале частицы является в данном случае микропросодическим: это автоматическое изменение ЧОТ на звонком взрывном [д].

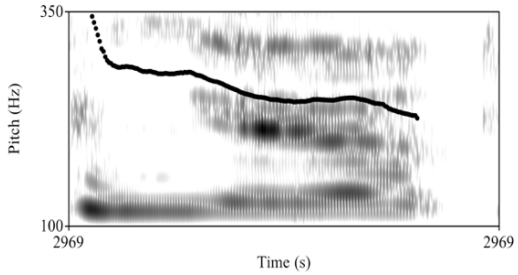


Рисунок 7: Динамическая спектрограмма и кривая ЧОТ (*Нет*, нисходящее движение)

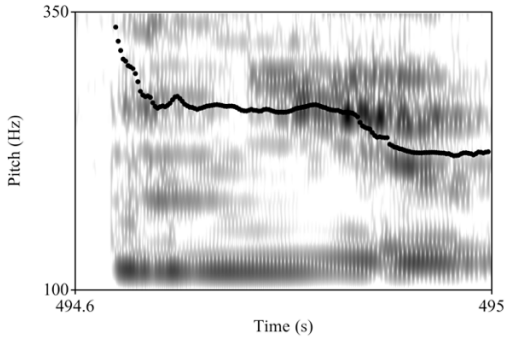


Рисунок 8: Динамическая спектрограмма и кривая ЧОТ
(*Нет*, ровное + нисходящее движение)

В целом, всё это разнообразие может быть сведено, однако, к противопоставлению восходящего акцента, сходного с ИК-4 литературного языка (рис. 3 – 5), и нисходящего (с возможным ровным или восходящим участком в начале), сходного с ИК-1 и ИК-2 (рис. 6 – 8).

3 Результаты и обсуждение

Одной из рабочих гипотез исследования было предположение о том, что в исследуемом говоре частицы *да* и *нет* интонируются по-разному: утвердительная – преимущественно с восходящим акцентом, отрицательная – с нисходящим (см. выше рис. 3 – 8, на которых все вхождения *да* оформлены восходящим движением ЧОТ, а все вхождения *нет* содержат нисходящее движение).

Результаты исследования представлены в таблице 1. Всего в проанализированных текстах встретилось 141 вхождение *да* и *нет* (104 и 37 соответственно)[6], в трех случаях (2 *да* и 1 *нет*) определить характер изменения ЧОТ оказалось невозможным вследствие дефектов записи, таким образом, всего было проанализировано 138 случаев (102 + 36).

	да	***нет***	**всего**
всего проанализировано	104	37	141
получены данные	102	36	138
нисходящий тон (кол-во)	22	5	27
нисходящий тон (% от общего числа)	22%	14%	20%

Таблица 1: Нисходящий тон на частицах *да* и *нет*: количество случаев, % от общего числа

На основании представленных в таблице данных можно заключить, что нисходящее (реже в сочетании с предшествующим ровным) движение тона в репликах информанта зафиксировано в 20% всех случаев (27 вхождений): 5 на *нет*; 22 на *да*. Уже этот факт позволяет усомниться в том,

[6] В четырех случаях частицы представлены в одной реплике дважды, в одном – трижды.

что восходящее движение тона на утвердительной и отрицательной частицах *да* и *нет* маркирует эмоции информанта («акцент вызова» по Е. А. Брызгуновой): сложно представить, что носитель диалекта лишь в одном случае из пяти использует нейтральное просодическое оформление *да* и *нет*. Тем не менее, для проверки рабочей гипотезы о том, что интонация зависит от семантики самих слов (утвердительная частица оформляется восходящим акцентом, а отрицательная – нисходящим) была построена бинарная логистическая регрессия в среде R, версия 4.0.2 (R Core Team, 2020), с использованием функции glm. В качестве предиктора был задан типа семантики (да vs. нет), а зависимой переменной был тип интонационного контура (восходящий vs. нисходящий). Информационный критерий Акаике (AIC) модели достиг 137.45. Результаты показали, что тип семантики не является значимым предиктором ($\beta = 0.457$, $p = 0.399$).

В связи с полученными результатами была предпринята попытка поиска иных факторов, объясняющих распределение восходящего и нисходящего акцентов на одних и тех же словах в исследуемом говоре. Анализ реплик, оформленных понижением тона, позволят предположить, что таким фактором является тип левого контекста – характер предшествующей реплики[7]: нисходящий акцент на частицах *да* и *нет* встречается преимущественно в тех случаях, когда предшествующая реплика является утверждением (в том числе – утверждением самого информанта), а не вопросом, см. ниже таблицу 2. Таких случаев, не вызывающих сомнений, в нашем материале зафиксировано 21 из общего числа в 26 реплик с нисходящим тоном (81%); исключениями являются 5 реплик, в которых нисходящий тон отмечен на частицах, которые могут быть интерпретированы как ответы на вопрос. Наоборот, восходящий акцент на частицах *да* и *нет* в подавляющем большинстве случаев отмечен после вопроса (в том числе – самого информанта). Несомненных случаев такого рода зафиксировано 101 из общего числа в 111 (91%); случаев, которые могут быть интерпретированы как исключения из данной закономерности, зафиксировано 10.

Предшествующая реплика	**Восходящий тон**	**Нисходящий тон**	**Всего**
Вопрос	101 (91%)	5 (19%)	106
Утверждение	10 (9%)	21 (81%)	31

Таблица 2: Нисходящий или восходящий тон на частицах *да* и *нет* в зависимости от семантики предшествующей реплики: количество случаев, % от общего числа

Как представляется, для описания наблюдаемой в говоре картины распределения нисходящего и восходящего акцентов на частицах *да* и *нет* можно предложить два разных объяснения – коммуникативное, и фонетическое. Первое могло бы быть связано со структурой диалога, распределением в нем ролей и способами передачи реплики собеседнику (turn allocation). Одним из способов такой передачи в русском языке как раз и является восходящий тон [8], и ИК-4 часто служит в СРЛЯ для маркирования того факта, что говорящий заинтересован в продолжении разговора – ср. интонационное оформление отрывка из одного из рассказов С.Довлатова:

(1) *Девушка-экскурсовод ела мороженое в тени. Я шагнул к ней:*
– Давайте познакомимся.
– Авро[1]ра, – сказала она, протягивая липкую руку.

(2) *Девушка-экскурсовод ела мороженое в тени. Я шагнул к ней:*
– Давайте познакомимся.
– Авро[4]ра, – сказала она, протягивая липкую руку.

В (1) ответ является нейтральным, и интонация ИК-1 никак не свидетельствует о желании или нежелании одного из участников продолжать диалог, в то время как ИК-4 в (2) является сигналом заинтересованности в продолжении разговора. Можно предположить, что в анализируемом говоре при подтверждении (*да*) или отрицании (*нет*) утверждения или состояния дел используется нисходящий акцент, поскольку дальнейший диалог в этом случае не предполагается. Наоборот, в

[7] В противоположность этому, в литературном русском языке *да* «как правило, несет на себе интонацию другой реплики, которая за ней следует» [6: 330]. Укажем в этой связи, что в нашем материале почти в половине всех случаев (69 из 141, 49%) частицы *да* и *нет* представляли собой законченные высказывания (без дальнейших реплик того же участника диалога), а в значительной части оставшихся частицы были включены в общий тональный контур реплики, что не позволило проанализировать влияние этого параметра на характер тонального оформления частиц.

случае подтверждения (*да*) или отрицания (*нет*) информации, заключенной в общем вопросе, имеет место восходящий тон (сходный с ИК-4), свидетельствующий о готовности к продолжению разговора и маркирующий готовность к передаче реплики собеседнику, что как раз является более вежливым просодическим оформлением в подобной ситуации, нежели нисходящий акцент, близкий ИК-1. Впрочем, есть и сильные аргументы против этой гипотезы. Во-первых, восходящий акцент в говоре используется и в ответах на вопрос, содержащийся в собственной реплике. Во-вторых, в ответах на реплики, являющиеся по своей семантике вопросами, но оформленные нисходящим движением тона, фиксируется нисходящий тон; наоборот, в реакциях на реплики, не являющиеся по своей семантике вопросами, но оформленные восходящим движением тона, имеет место восходящий акцент (всего 7 примеров). На этом основании можно сформулировать фонетическое объяснение: характер просодического оформления частиц *да* и *нет* зависит в говоре от типа акцента на предыдущей реплике.

Для проверки коммуникативной и фонетической гипотезы были построены две бинарные логистические регрессии. В обеих моделях тип интонационного контура на частице (восходящий vs. нисходящий) был задан в качестве зависимой переменной, но в первой модели предиктором выступал характер предшествующей реплики (утверждение vs. вопрос), в то время как во второй – характер изменения ЧОТ на предшествующей реплике (нисходящий vs. восходящий). Результаты показали, что как характер предшествующей реплики, так и направление движения тона на ней являются значимыми предикторами типа интонационного контура на частице (β = -3.733, p < .000 и β = -6.062, p < .000, соответственно). Критерий AIC первой модели равен 83.45, второй – 46.09. Критерий AIC применяется для выбора из нескольких статистических моделей: чем ниже критерий, тем лучше модель описывает данные. Из этого можно сделать вывод, что "фонетическая" модель работает лучше. Соответствующие данные приведены ниже в таблице 3, а примеры, иллюстрирующие это положение, на рис. 9 – 12, где хорошо видно полное совпадение тонального оформления ответной реплики с предшествующей ей вне зависимости от коммуникативного типа реплики-стимула.

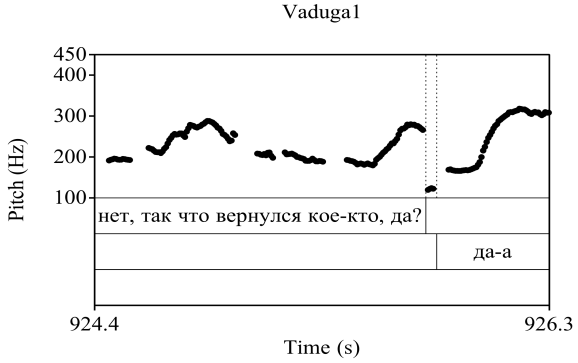


Рисунок 9: Интонограмма реплик – *Нет, так что вернулся кое-кто, да? – Да-а.* (ровное + восходящее движение тона на обеих репликах)

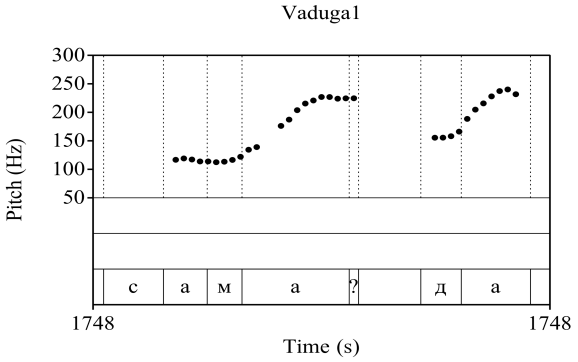


Рисунок 10: Интонограмма реплик – *Сама? – Да-а.*
(ровное + восходящее движение тона на обеих репликах)

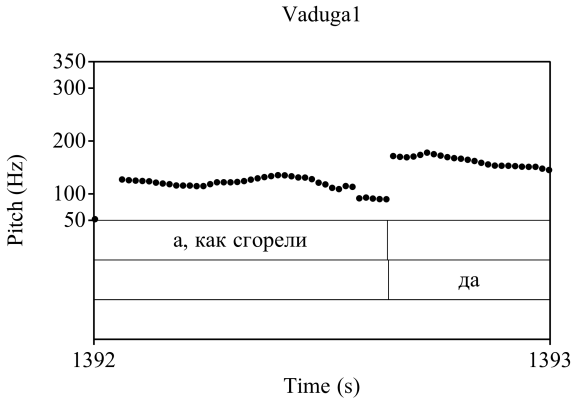


Рисунок 11: Интонограмма реплик – *А, как сгорели. – Да.*
(нисходящее движение тона на обеих репликах)

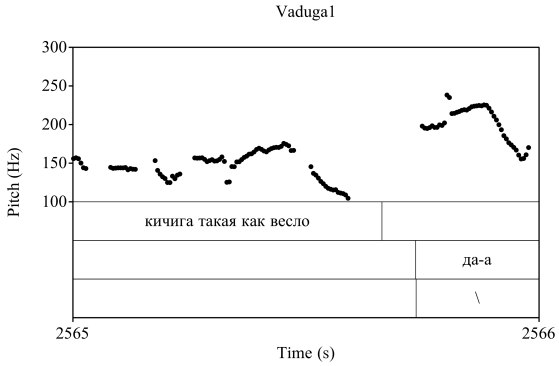


Рисунок 12: Интонограмма реплик – *Кичига такая, как весло. – Да-а.*
(восходяще-нисходящее движение тона на обеих репликах)

На основании приведенных данных можно заключить, что выбор нисходящего или восходящего тона на частицах *да* и *нет* в ответных репликах обусловлен фонетически и копирует тон предшествующей реплики, то есть, верным является фонетическое объяснение, которое позволяет существенно уменьшить число тех реплик, которые иначе пришлось бы считать исключениями. Их список может быть сокращен не только на 7 примеров, описанных выше и не противоречащих фонетической гипотезе, но и доведен до минимума, т.к. 1) нисходящий тон в двух репликах обусловлен не ответом на вопрос собеседника, а низким тоном предшествующей реплики другого информанта (при этом в одной из реплик вопрос интервьюера задан так тихо, что не мог быть расслышан информантом) – таким образом, для нисходящего тона на *да* и *нет* исключений не остается вовсе; 2) в одном из случаев ответ относится к первой части реплики, оформленной восходящим тоном, а не ко второй – с нисходящим; 3) в двух случаях частица включена в общий просодический контур высказывания, поэтому движение тона на ней может не быть показательным. Тем самым, настоящих «исключений» (восходящий тон после реплики с нисходящим) оказывается только 3 из 138 (2%); возможно, как раз они представляют собой эмоциальные реакции.

Предшествующая реплика	**Восходящий тон**	**Нисходящий тон**	**Всего**
Восходящий тон	108 (98%)	0 (0%)	108
Нисходящий тон	3 (2%)	26 (100%)	29

Таблица 3: Нисходящий или восходящий тон на частицах *да* и *нет* в зависимости от характера изменения ЧОТ на предшествующей реплике: количество случаев, % от общего числа

Для **дополнительной верификации** результатов проведенного эксперимента были проанализированы данные еще двух носителей верхнепинежского диалекта, собранные в ходе той же

экспедиции (МИР и АПН, см. выше раздел “Материал”). В общей сложности было выявлено еще 99 реплик с частицами *да* и *нет*. На основе этих данных были построены три бинарные логистические регрессии для проверки семантической, коммуникативной и фонетической гипотез. В спецификации семантической модели предиктором выступал тип семантики частицы (“да” / “нет”), коммуникативной – характер предшествующей реплики (утверждение / вопрос), фонетической – тип тона предшествующей реплики (нисходящий / восходящий). Во всех трех моделях зависимой переменной являлся тип интонационного контура на частице (восходящий vs. нисходящий). Полученные результаты подтвердили, что тип семантики не является значимым (β = -0.176, p = .805), в то время как характер и тип тона предшествующей реплики – значимые предикторы (β = -4.407, p < .000 и β = -5.730, p < .000, соответственно). Критерий AIC фонетической модели был выше аналогичного критерия коммуникативной модели (29.42 и 42.00, соответственно). Таким образом, как и в основном анализе, модели, построенные на дополнительных данных, показали, что тип тона на частице зависит от обоих факторов (характер предшествующей реплики и тип тона предшествующей реплики), однако фонетическая модель лучше объясняет данные.

4 Выводы

Один из основных выводов, который можно сформулировать на основании результатов настоящего исследования, заключается в том, что при анализе диалектной интонации не следует ориентироваться на семантическую интерпретацию интонационных конструкций СРЛЯ, поскольку диалект – это другая языковая система, в которой мелодические контуры, даже весьма сходные с литературными, могут передавать совсем другие значения. Очевидно, что вопреки первым впечатлениям о диалекте его носители не отвечают на вопросы собеседников невежливо – категорично, с вызовом или подчеркивая противопоставление [4: 115] – вполне возможно, что как раз наоборот, не будучи уверенными в том, какова именно семантика того или иного мелодического контура в реплике собеседника, говорящего на литературном языке, они из вежливости копируют в ответе основные тональные характеристики этого контура.

Полученные в ходе настоящего исследования данные свидетельствуют, на наш взгляд, еще и о том, что требуется дальнейшее изучение фразовой просодии русских диалектов, особенно в направлении не от формы к значению, поскольку точно оценить семантику, передающуюся в говоре при помощи просодии в таком случае оказывается очень сложно, а от содержания к форме: каким именно образом передаются те или иные коммуникативные значения в диалекте. Подавляющее большинство доступных в настоящий момент диалектных текстов представляют собой либо монологи носителей говора, либо диалоги, в которых вопросы задает диалектолог, поэтому большая часть коммуникативных типов в речи носителя либо не встречается вовсе, либо встречается крайне редко, а для тех, что встречаются, как показало данное исследование, представляется в высшей степени затруднительным оценить их семантику на основании наших знаний о литературной фразовой просодии, особенно при использовании системы ИК, в которой основным компонентом каждого мелодического контура является изначально заданная семантика. Поэтому, на наш взгляд, необходим иной подход к сбору диалектного материала для просодических исследований. Он должен быть организован таким образом, чтобы в текстах встречались все необходимые коммуникативные типы с известной заранее семантикой, желательно, на сопоставимом для всех исследуемых языковых систем и удобном для фонетического анализа звуковом материале; при этом он должен быть удобным для работы с информантами старшей возрастной категории, зачастую неграмотными или не способными по разным причинам работать с письменными инструкциями. Наиболее адекватным методом для решения этой задачи, как представляется, может служить *Discourse Completion Task* (DCT) – метод, адаптированный для просодических исследований из работ в области прагматики; он успешно используется для элиситации разных типов интонационных контуров и получил широкое распространение для описания интонации разных языков, прежде всего, романской группы [14]. При использовании DCT информантам предлагается дополнить (complete) каким-либо высказыванием с заданным коммуникативным значением краткий диалог или ситуацию из их повседневной жизни [1]. Применение этого метода позволит не только элиситировать предложения с заданной прагматикой, но и сравнить их просодическую реализацию в речи разных носителей, все это поможет получить надежные результаты о репертуаре просодических средств данного диалекта и их значениях.

References

[1] Barron, Anne. *Acquisition in interlanguage pragmatics: Learning how to do things with words in a study abroad context*. Amsterdam: John Benjamins, 2009.

[2] Bryzgunova E. A. *Zvuki i intonatsiya russkoy rechi* [Sounds and intonation of Russian speech]. Moscow, 1969 (In Russ.)

[3] Bryzgunova, E. A. Analiz russkoy dialektnoy intonacii [An analisys of Russian dialectal intonation] In: *Eksperimental'no-foneticheskie issledovaniya v oblasti russkoy dialektologii* [Experimental phonetic studies in the field of Russian dialectology]. Moscow, Nauka, 1977. P. 231-262 (In Russ.)

[4] Bryzgunova E. A. (*Intonatsiya* [Intonation]. In: Russkaya grammatika [Russian Grammar]. M.Yu.Svedova (ed.). Vol. I. Moscow, Nauka, 1980 (In Russ.)

[5] Knyazev S.V., Levina A.N., Pozharitskaya S.K. O govorax Verxney Pinegi i Vyi [On Upper Pinega and Vyya dialects]. In: *Russkie dialekty: istoriya i sovremennost'. Problemy russkogo yazykoznaniya* [Russian dialects: history and present state. Problems of Russian linguistics. VII]. Moscow, Moscow State University Publ., 1997. P. 198-220 (In Russ.)

[6] Kodzasov S. V. *Studies in Russian prosody*. Moscow, Yazyki Slavyanskikh Kul'tur, 2009 (In Russ.)

[7] Kuznetsov P. S. O govorax Verxney Pinegi i Verxney Toymy [On Upper Pinega and Upper Toyma dialects]. In: *Materialy i issledovaniya po russkoy dialektologii*. Tom 1 [Materials and studits in Russian dialectology. Vol. 1] Moscow – Leningrad, 1949. P. 5-44 (In Russ.)

[8] Paschen L. Boundary tones indicate turn allocation in Russian. In *Proceedings of ConSOLE XXII*, Leiden, 2015.

[9] Paufoshima, R. F. *Fonetika slova i frazy v severnorusskix govorax* [Phonetics of word and phrase in Northern Russian dialects]. Moscow, Nauka, 1983 (In Russ.)

[10] Paufoshima, R. F. Ob ispol'zovanii registrovyx razlichiy v russkoy frazovoy intonatsii (na materiale russkogo literaturnogo yazyka i severnorusskix govorov) [On the use of tone-level differencies in russian phrase intonation (based on data from Standard Russian an Northern Russian dialects)]. In: *Slavyanskoe i balkanskoe yazykoznanie. Prosodiya* [Slavic and Balkan linguistics. Prosody]. Moscow, Nauka, 1989 (In Russ.)

[11] R Core Team. R: A language and environment for statistical computing. Foundation for Statistical Computing, 2020. https://www.r-project.org/.

[12] Rodero, Emma/ Intonation and Emotion: Influence of Pitch Levels and Contour Type on Creating Emotions. *Journal of Voice*. Volume 25, Issue 1, January 2011, P. 25-34.

[13] Shchigel' E. V. Osobennosti intonacionnoy organizacii frazy v nekotoryx severnorusskix govorax [Peculiar properties of intonational structure in some Northern Russian dialects]. In: *Dialektografiya russkogo yazyka* [Dialectografy of Russian]. Moscow, Nauka, 1985. P. 102-111 (In Russ.)

[14] Vanrell, Maria del Mar, Feldhausen, Ingo & Lluïsa Astruc. The Discourse Completion Task in Romance prosody research: Status quo and outlook. In: Ingo Feldhausen, Jan Fliessbach & Maria del Mar Vanrell (eds.), *Methods in prosody: A Romance language perspective*, 191-227. Berlin: Language Science Press, 2018.

Parenthetical constructions in Russian spoken discourse: Basic types and prosodic features

Nikolay A. Korotaev
Russian State University for the Humanities
(RSUH)
n_korotaev@hotmail.com

Abstract

The paper discusses the notion of parentheticals in Russian spoken discourse. Using data from two prosodically annotated corpora — "Stories about presents and skiing" and "Russian Pear Chats & Stories" — I advocate for a discourse-oriented approach to parenthetical constructions. I define a parenthetical construction as consisting of three elements: the left context, the parenthetical unit, and the right context. Each element constitutes a separate discourse unit and is thus prosodically autonomous. I rely on the notion of projection [Auer 2005] to account for the discourse relationships between these three components. When the speaker pronounces the left context, she projects a continuation that is to be realized in the right context, while the parenthetical unit provides a digressive discourse step.

Typically (around 50% in my data), parentheticals are anchored to their left contexts and are pronounced with a falling or level pitch accent. Noted deviations from this prototype include free parentheticals, parenthetical uses of *vot*, and parentheticals pronounced with a rising pitch accent. Furthermore, I explore two prosodic features frequently associated with parentheticals, namely, increased articulation rate and pitch range narrowing. I show that, while both these tendencies are statistically significant, the latter has a larger effect size than the former.

Keywords: spoken discourse, Russian language, discourse structure, discourse prosody, parentheticals, intonation, speech rate, pitch range

DOI: 10.28995/2075-7182-2021-20-413-424

Конструкции с дискурсивными вставками в устной русской речи: базовые типы и просодические свойства

Николай Коротаев
Российский государственный гуманитарный
университет (РГГУ)
n_korotaev@hotmail.com

Ключевые слова: устная речь, русский язык, структура дискурса, дискурсивная просодия, парентеза, интонация, темп, тональный диапазон

1 Вводные замечания

Данная работа посвящена явлению дискурсивной вставки (парентезы) в неподготовленной устной речи. Под конструкцией со вставкой я, вслед за [Кибрик, Подлесская 2010], буду понимать последовательность из как минимум трех дискурсивных единиц: первая и последняя связаны между собой тем или иным содержательным отношением в рамках основной линии изложения, а между ними располагается фрагмент, выбивающийся из основной линии. С точки зрения динамического развертывания дискурса особенность таких конструкций заключается в том, что говорящий временно откладывает намеченное в первой единице продолжение, реализует побочное действие, а затем возвращается к изначальному плану. Так, в примере (1) говорящий, произнеся единицу E010 [1] , недвусмысленно указывает на планируемое продолжение: для этого

[1] Об оформлении примеров и используемых в них обозначениях см. раздел 2.

используются как синтаксические (деепричастная форма), так и интонационные (восходящий тональный акцент) средства.

(1) Pic-RUS_05-m_Ski-T

19.57	E010	/Выпив хорошенько,
20.83	p-007	(0.18)
21.02	E011	(\Непонятно,
21.71	p-008	(0.05)
21.76	E012	почему (0.05) кстати с \утра он пьёт.)
23.65	p-009	(0.78)
24.42	E013	/\он (0.42) решил покататься ещё на \лыжах,
26.98	E014	и /видимо нь= || выбрал не ту –↓горку.

Намеченное продолжение реализуется в строке E013, которая образует с E010 единый содержательный, синтаксический и интонационный комплекс. При этом между этими двумя единицами производится вставка: комментарий рассказчика не входит в основную линию изложения и — что еще более существенно — не может восприниматься как завершение ранее начатого высказывания. В транскрипте вставленные единицы заключены в скобки. На рис. 1 представлена тонограмма этого фрагмента: можно заметить, что единицы E011–E012 произносятся с ощутимо менее рельефными движениями частоты основного тона, чем единицы E010 и E013[2].

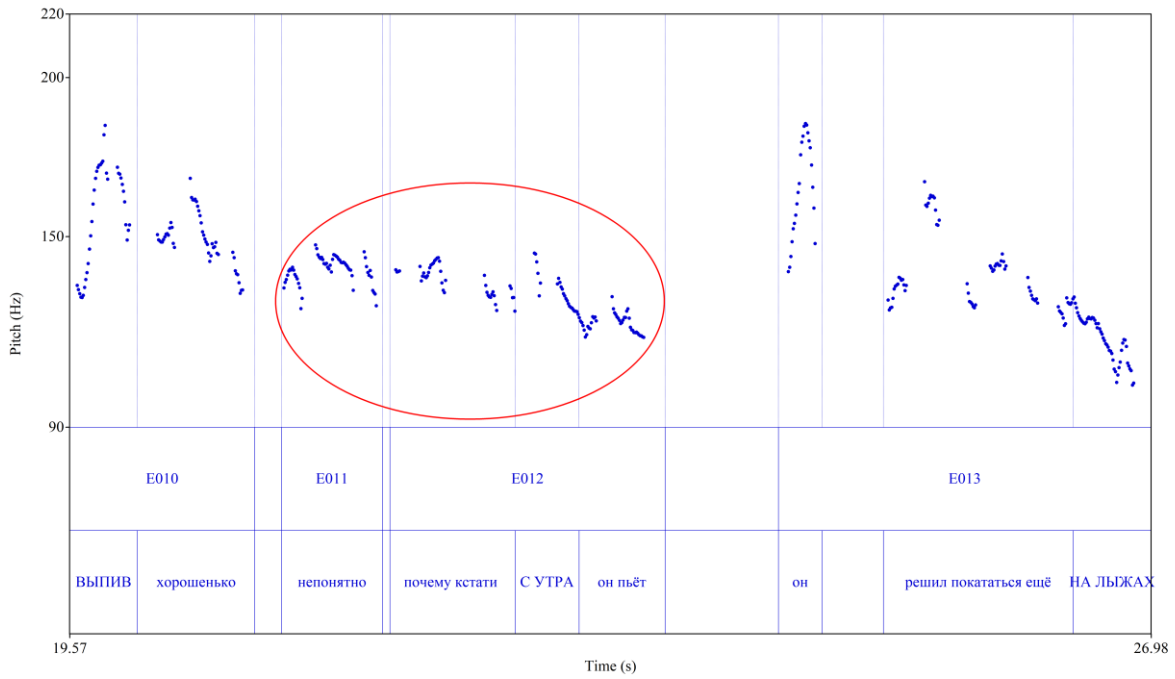


Рис. 1. Тонограмма примера (1). Заглавными буквами выделены словоформы, произносимые с главными акцентами; овалом — участок тонограммы, соответствующий вставке

Вставкам посвящена обширная литература, см., в частности, [Urmson 1952; McCawley 1982; Кобозева 1999; Гавриленко 2004; Dehé & Kavalova (eds.) 2007; Mazeland 2007; Падучева 2010: 321-334; Богданова 2011; Dehé 2014]. Сколько-либо подробное обсуждение этих источников выходит за рамки настоящей работы, однако необходимо отметить, что единого понимания парентезы, по всей видимости, не существует. Вставка — это многофакторное явление, при описании которого, в зависимости от исследовательских задач, на первый план выходят соображения семантико-прагматического, синтаксического, просодического, дискурсивного и

[2] Акустический анализ примеров выполнялся в программе Praat [Boersma & Weenink 2018].

иного характера. Для настоящей работы центральным является вопрос о том, каким образом вставки реализуются в структуре неподготовленного устного русского текста. Непосредственные задачи исследования сводятся к следующему:

- сформулировать рабочее определение конструкции со вставкой и продемонстрировать, какие классы конструкций выделяются в привлеченном устном материале;
- реализовать предварительный анализ акустических характеристик вставленных элементов.

Дальнейший текст статьи имеет следующую структуру. В разделе 2 будет описан материал исследования и кратко представлены базовые принципы используемой дискурсивной нотации. В разделе 3 будут рассмотрены интегральные свойства конструкций со вставками, на которые я предлагаю опираться при определении. В разделе 4 я покажу, какие базовые классы конструкций со вставками выделяются в корпусе, и рассмотрю их дифференцирующие свойства. В разделе 5 будут представлены предварительные результаты количественного анализа темповых и тональных характеристик вставок. Раздел 6 содержит заключительные замечания.

2 Материал

Материалом исследования послужили записи двух корпусов звучащей речи, снабженных дискурсивно-просодической разметкой:

- «Истории о подарках и катании на лыжах» (http://spokencorpora.ru/showcorpus.py?dir=03pands/rus) — 35 минут звучания, около 4 500 словоупотреблений;
- три записи корпуса «Рассказы и разговоры о грушах» (https://multidiscourse.ru/; [Кибрик 2018]) — 60 минут звучания, около 10 000 словоупотреблений[3].

Все записи содержат речь носителей русского языка (от 18 до 30 лет): исключительно монологического (нарративного) характера в первом корпусе и смешанного диалого-монологического — во втором. На начало проведения исследования все записи имели доступную разметку. Перечислю базовые особенности текстового компонента разметки (= дискурсивной транскрипции), существенные для понимания приводимых в статье примеров; более подробное описание принципов аннотации см. в [Kibrik et al. 2020].

- Транскрипты поделены на нумерованные строки. Строки, номер которых содержит литеру E, соответствуют элементарным дискурсивным единицам (ЭДЕ). ЭДЕ — это минимальные шаги в развитии устного дискурса, выделяемые на основании набора просодических критериев.
- В каждой ЭДЕ (за возможным исключением коротких оборванных единиц) содержится по крайней мере одна акцентированная, т.е. просодически выделенная, словоформа. Наличие акцентов отмечается в транскрипте при помощи слэшей и стрелок перед акцентированными словоформами; при этом направление слэшей указывает на характер движения частоты основного тона в ударном слоге, направление стрелок — на значимые внеударные движения. Например, словоформа */выпив* в ЭДЕ E010 примера (1) произносится с восходящим акцентом; а в словоформе – *↓ горку* (ЭДЕ E014) реализован ровно-нисходящий акцент, причем нисходящая часть расположена на заударном слоге.
- Подчеркиванием ударной гласной дополнительно отмечаются главные акценты в ЭДЕ. Направление движения тона в главном акценте является одним из центральных факторов, определяющих иллокутивно-фазовое значение ЭДЕ, которое, в свою очередь, кодируется при помощи пунктуационных знаков в конце строки. Выделяются две группы пунктуационные знаков: финальные (прежде всего, точка) и нефинальные (запятая и некоторые

[3] В заголовках примеров указывается кодовое имя записи. Коды, начинающиеся с Pic-RUS, соответствуют рассказам корпуса «Истории о подарках и катании на лыжах»; коды, начинающиеся с Pears, — записям корпуса «Рассказы и разговоры о грушах».

другие). Это разграничение связано с противопоставлением завершенности vs. незавершенности, подробно рассматриваемым, в частности, в работах [Янко 2008; Кодзасов 2009].

3 Интегральные свойства и рабочее определение

Как уже было отмечено в разделе 1, в данной работе за основу взято определение вставки из [Кибрик, Подлесская 2010]. Приведу фрагмент этого определения, также сформулированного на устном корпусном материале: «Вставка — это временное отклонение от маршрута изложения, выполнение некоторого побочного хода, и затем возврат к исходному маршруту. (…)_... последовательность из трех дискурсивных единиц X + Y + Z, в которой Y — уточнение к X, а основная линия повествования связывает X и Z». Ниже я несколько подробнее рассмотрю компоненты этого определения и укажу на свойства, общие для всех вставочных конструкций, обнаруженных в корпусе.

3.1 Трехчастная структура

Для квалификации конструкции как содержащей вставку необходимо наличие трех компонентов: вставка не просто соотносится с некоторым контекстом, но и обязательно разрывает его. При таком понимании к вставкам не относятся коммуникативно подавленные элементы левой / правой периферии клаузы: начальные употребления вводных оборотов вида *(как) кажется*, постпозитивные расширения и проч. Подобные элементы в этой статье не рассматриваются. Далее для единиц, обозначенных выше как X, Y и Z, я буду использовать рабочие термины «левый контекст», «вставка» и «правый контекст». Так, в примере (1) в строке E010 завершается левый контекст, строки E011 и E012 составляют вставку, а в строке E013 начинается правый контекст[4].

3.2 Проекция на продолжение

Для интерпретации того факта, что левый и правый контекст реализуют основную линию изложения, кажется удобным опираться на введенное П. Ауэром понятие проекции [Auer 2005]. Под проекцией понимается разделяемое участниками коммуникации представление о том, что одно речевое действие в известной степени влечет, или проецирует, другое. В конструкциях со вставками проекция возникает при произнесении левого контекста: реализуя вставку, говорящий подразумевает, что слушатель сохранит сформированные ранее ожидания продолжения до начала произнесения правого контекста.

Степень проекции может быть различной. В примере (1) был представлен случай сильной проекции, поскольку использование в левом контексте деепричастной клаузы с высокой степенью вероятности предсказывает дальнейшее появление главной клаузы. Соответственно, не получив ожидаемого структурного продолжения в строках E011-012, слушатель интерпретирует эти ЭДЕ как вставку и «переносит» свои ожидания на правый контекст. Однако столь явные указания на характер продолжения имеются далеко не всегда. Так, в примере (2) строка R-vE164 содержит интонационный сигнал незавершенности (восходящее движение тона в главном акценте), но какими структурными единицами эта незавершенность должна «разрешаться», не специфицируется. Это можно назвать слабой проекцией.

(2) Pears23R

792.12	R-vE163	для чего-то берёт последнюю /грушу,
793.72	R-vE164	снимает с себя /платок такой,
795.18	pR-136	(0.11)
795.29	R-vE165	(\красный вот,
795.96	pR-137	(0.20)
796.16	R-vE166	который на –шею вешают,)

[4] В [Кибрик, Подлесская 2010] также упоминаются случаи «односторонней парентезы», при которых единицы, изначально порождаемые как вставки, берут на себя функцию правого контекста. Такие примеры встречаются и в нашем материале, но в настоящей работе не рассматриваются.

797.26	R-vE167	протирает эту последнюю /грушу,
798.81	pR-138	(0.18)
798.98	R-vE168	и /–кладёт /–обратно в \корзину.

3.3 Автономность вставки

Вне зависимости от типа проекции вставленные единицы не могут интерпретироваться как полноценное продолжение структуры, проецируемой в левом контексте. Появление вставки сигнализирует о том, что ожидаемое продолжение откладывается. Вместе с тем, согласно приведенному определению, вставка все же представляет собой отдельный — пусть и побочный — дискурсивный шаг. Формально это выражается тем, что вставка имеет автономный интонационный контур и потому интерпретируется как ЭДЕ или группа ЭДЕ.

Отмечу, что аналогичный подход представлен в работах по анализу бытового диалога, в которых под парентезой понимаются отдельные репликообразующие единицы (*turn-constructional units*), также выделяемые на просодических основаниях [Mazeland 2007; Schegloff 2007: 237ff]. Напротив, представление о парентезе как о безакцентном элементе коммуникативной структуры, синтаксически не связанном с объемлющим контекстом [Янко 2001: 81], с предлагаемым определением не согласуется. Как видно из примеров выше и ниже по тексту, вставки обладают интонационной автономностью и при этом могут быть как синтаксически независимы от левого и правого контекста (см. примеры (1), (5)), так и вступать с ним в те или иные синтаксические отношения (2–4, 6).

3.4 Уровень локальной структуры

Отдельным является вопрос об уровне иерархической структуры дискурса, на котором происходит вставка. В примерах (1) и (2) речь шла о вставке между ЭДЕ. Случай другого рода представлен в примере (3): здесь вставка интонационно автономной единицы C-vE113 происходит между составляющими простой клаузы.

(3) Pears22C

374.87	C-vE111	И он с этими /\грушами возвращается к своим /друзьям,
377.08	C-vE112	↑\даёт им —
377.49	C-vE113	(\каждому,)
377.89	C-vE114	— по /груше,
378.37	C-vE115	и /они идут мимо /фермера,
379.66	C-vN036	(ц 0.38)
380.04	C-vE116	/весело жуя эти –груши.

Дистрибутивное местоимение *каждому* формирует отдельную ЭДЕ, поскольку произносится с ощутимым акцентом, выбивающимся из единого интонационного контура клаузы *даёт им по груше*. Если бы *каждому* произносилось безударно или же вовсе не было произнесено, последовательность строк C-vE112–114 составляла бы одну ЭДЕ. Таким образом, можно сказать, что вставка здесь происходит не между, а внутри ЭДЕ. В [Кибрик, Подлесская 2010] для таких случаев используется термин «сплит»; в транскрипте левый и правый контекст маркируется при помощи длинного тире[5]. Особенность сплита состоит в том, что говорящий решает осуществить побочный дискурсивный шаг, не дожидаясь завершения текущего шага. Возможно, с этим связана достаточно частотная особенность конструкций со сплитом: в отличие от случаев (1)-(2), в которых вставки обладают «подавленным» коммуникативным статусом, вставка, попадающая в сплит, нередко, напротив, информационно выделена.

Как бы то ни было, и в примерах (1)–(2), и в примере (3) вставка происходит на уровне, не выходящем за рамки одного «устного предложения», т.е. реализуется в ситуации интонационно выраженной в левом контексте дискурсивной незавершенности. По всей видимости, вставки

[5] Аналогичное противопоставление между вставками, происходящими между репликообразующими единицами vs. внутри одной репликообразующей единицы, рассматривается на материале диалогических данных в [Mazeland 2007].

возможны и на более высоком уровне дискурсивной структуры — между предложениями. Обособленный характер вставленного материала в таких случаях может быть, например, выражен лексически (*кстати, между прочим* и др.). Однако в общем случае, ввиду отсутствия интонационной проекции на продолжение после левого контекста, идентификация таких последовательностей как содержащих вставку значительно менее очевидна. В настоящей работе такие примеры не рассматриваются[6].

Итак, при разметке корпуса использовалось следующее понимание конструкции с дискурсивной вставкой:

- конструкция состоит из трех компонентов: левого контекста, собственно вставки и правого контекста;
- левый контекст характеризуется дискурсивной незавершенностью и проецирует продолжение, реализуемое не во вставке, а в правом контексте;
- вставка представляет собой отдельную ЭДЕ или группу ЭДЕ.

4 Типы конструкций в корпусе: дифференцирующие свойства

Всего в исследованном подкорпусе обнаружено 236 конструкций с дискурсивными вставками: 174 случая в «Рассказах и разговорах о грушах» и 62 — в «Историях о подарках и катании на лыжах». Выше, при обсуждении интегральных свойств вставочных конструкций, были также упомянуты некоторые параметры варьирования: наличие и характер синтаксического отношения между вставкой и контекстом, объем вставки, тип проекции и др. Ниже будут рассмотрены еще две группы свойств, связанных (а) с характером содержательного отношения вставки к контексту и (б) с типом содержательной и интонационной обособленности вставки. Опора на эти параметры позволяет определить базовые типы конструкций с дискурсивными вставками в проанализированном материале.

4.1 Ядерный тип

Ядро выявленных случаев (к ним, в частности, относятся приведенные выше примеры (1) – (3)) составляют конструкции, характеризуемые следующим набором прототипических свойств.

(i) Вставка содержательно соотносится с конкретным элементом синтаксической структуры объемлющего контекста: с целой клаузой (пример (1)), с глагольной группой (3), с именной вершиной (2) и др. В [Kavalova 2007] для подобных случаев предлагается термин *anchored parentheticals*. Также существенно, что в прототипической ситуации «якорь», к которому содержательно прикрепляется вставка, располагается в левом, а не в правом контексте.

(ii) Вставка обладает существенной степенью внутренней законченности. При максимальной выраженности этого признака во вставке реализуется самостоятельная иллокутивная функция, отдельная от иллокутивной функции объемлющего контекста, — см. пример (1), в котором во вставку помещен оценочный комментарий говорящего, внешний по отношению к миру рассказа. Типичным интонационным коррелятом иллокутивной независимости выступает нисходящий или ровный акцент в главном акценте. Если же вставка не обладает полноценной иллокутивной силой, наличие нисходящего или ровного акцента становится главным фактором, противопоставляющим вставку окружающему контексту. Нередко при этом падение частоты основного тона не достигает нижнего для данного говорящего уровня; в транскриптах (2) и (3) на это указывают запятые (а не точки) перед закрывающимися скобками.

Заметной интонационной характеристикой прототипической вставки также может выступать последовательное понижение частоты основного тона на протяжении всего вставленного фрагмента — см. тонограмму примера (4) на рис. 2. Вставка, состоящая из двух клауз (E015–E016), произносится в «сплющенном» тональном диапазоне и с практически равномерным, пологим падением. Такое оформление резко контрастирует с интонационными фигурами, реализованными в левом и правом контексте и включающими чередования восходящих и нисходящих тональных движений.

[6] Другой подход представлен, в частности, в [Богданова 2011], где прямо утверждается, что интонационные критерии не позволяют выделить вставки в контексте и что вставки не чувствительны к границам устных «предложений».

(4) Pic-RUS_01-f_Ski-T

30.15	E014	/Покатался /он (0.05) ↑не \очень –↑удачно,
32.14	p-013	(0.56)
32.70	E015	(так как был \пьяный,
33.71	p-014	(0.15)
33.85	E016	а за рулём как известно /\нельзя пить,)
35.55	p-015	(0.86)
36.41	E017	/и /попал в ↓–реанимацию!

Рис. 2. Тонограмма примера (4). Стрелкой выделено последовательное понижение частоты основного тона на протяжении двух вставленных ЭДЕ

Далеко не все вставки характеризуется столь же явной интонационной обособленностью, как в примере (4). Тем не менее в целом немногим менее половины всех обнаруженных в корпусе случаев удовлетворяют двум приведенным выше условиям (см. далее табл. 1). Оставшиеся примеры тем или иным образом отклоняются от указанного прототипа.

4.2 «Свободные» вставки

В отличие от вставок в примерах (1)–(3), «свободные» вставки либо содержательно соотносятся со всем объемлющим контекстом целиком (а не с его отдельными элементами), либо вовсе чужеродны контексту. В [Dehé 2014: 8ff] для таких случаев используются ярлыки *floating* и *detached parentheticals*. Пример вставки, полностью выпадающей из основной линии изложения, приведен в (5). В строке N-vE233 говорящая начинает отвечать на вопрос собеседницы о внешности персонажа обсуждаемого фильма, проецируя продолжение посредством стандартной восходящей интонации. В это же время альтернативную попытку ответить на тот же вопрос предпринимает третий участник записи. Не желая уступать право хода, говорящая приостанавливает ответ на вопрос, предлагает третьему участнику повременить с его версией (строки N-vE234–235), а уже затем, в строке N-vE236, приступает к ранее проецированному продолжению[7].

[7] В [Богданова 2011] аналогичные примеры определяются как металингвистические вставные конструкции и рассматриваются в ряду характерных для спонтанной речи типов. Как видно из табл. 1, в наших данных такие примеры относительно редки, что, разумеется, не отменяет их содержательной специфики.

(5) Pears22N

502.38	N-vE233	Он такой очень /смешной,	Обращается к Пересказчице
503.65	pN-054	(0.25)	
503.90	N-vE234	(/Можно я расскажу^{h}?	Обращается к Комментатору
504.93	N-vE235	А потом ты ко= || /откомментируешь.)	
506.57	*N-vL007*	{laugh 0.64}	
507.21	N-vE236	у него такая объёмная /шевелюра,	Обращается к Пересказчице
508.93	N-vN052	(ц 0.33)	
509.25	N-vE237	(\вот,)	
509.66	N-vE238	и /усы ==	

4.3 Парентетическое *вот*

В строке N-vE237 примера (5) представлен еще один тип вставки — парентетическое *вот*. Полноударная реализация дискурсивного маркера *вот* стандартно используется для указания на «завершение структурно значимого фрагмента текста и переход к следующему» [Дараган 2003]. Если *вот* следует за финальной ЭДЕ, считать его вставкой — согласно приведенному в разделе 3 определению — нет оснований. Если же подытоживаемый дискурсивный фрагмент произносится с интонацией незавершенности, то *вот* может иметь парентетическую природу. Именно так происходит в примере (5). Урегулировав коллизию с правом очередности, говорящая возвращается к описанию персонажа и произносит ЭДЕ N-vE236 — вторую после N-vE234 в серии нефинальных единиц основной линии. Прежде чем продолжить, она реализует вставочное *вот*, давая таким образом понять, что не планирует более говорить о прическе, а намеревается перейти к другим аспектам внешности обсуждаемого персонажа — что и делает в строке N-vE238. (Дальнейшее развитие прерывается по причинам, не имеющим отношения к рассматриваемым в работе явлениям.)

Парентетическое *вот* практически всегда произносится с нисходящим или ровным акцентом. При этом оно не соотносится с каким-либо отдельным элементом левого или правого контекста, а указывает на то, как говорящий в целом оценивает формируемую локальную структуру. По предварительным наблюдениям, частота использования подытоживающего *вот* (как в финальной, так и в парентетической позиции) обусловлена индивидуальной манерой речи: одни говорящие прибегают к этому средству структурирования речевого потока регулярно, другие — редко или практически никогда. Так, из 51 вхождения парентетического *вот* в трех записях корпуса «Рассказы и разговоры о грушах» 21 случай обнаруживается в речи одной участницы. Отмечу также, что парентетическую функцию способны выполнять и некоторые другие дискурсивные маркеры или их сочетания: *ну*, *ну вот*, *значит* и др. Однако в рассмотренном материале *вот* частотность таких случаев значительно уступает частотности парентетического *вот*.

4.4 Псевдопарентетические вставки

Наиболее заметное отклонение от рассмотренного в разделе 4.1 прототипа наблюдается при произнесении вставленных единиц с восходящим движением тона. В этом случае возникает противоречие между содержательной ролью вставки в локальной структуре и ее интонационным оформлением. Так, в примере (6) основная линия изложения связывает ЭДЕ N-vE173 и N-vE175 — два нерестриктивных придаточных определительных, сочиненных между собой и имеющих общее подлежащее. В первом из них при помощи нисходяще-восходящего акцента типа ИК-4 проецируется продолжение, реализуемое во втором. Располагающаяся между ними ЭДЕ N-vE174 демонстрирует стандартные структурные и содержательные характеристики вставки: в ней не содержится намеченного ранее продолжения основной линии, она уточняет элемент левого контекста и составляет побочный дискурсивный шаг. Однако интонационно эта строка

оформлена так же, как левый контекст — в ней тоже реализуется нисходяще-восходящее тональное движение, как если бы она была полноценным элементом основной линии. Как можно заметить, в транскриптах такие «псевдопарентетические» единицы не заключаются в скобки.

(6) Pears04N

287.59	N-vE172	(ˀ 0.32) /И-и (э 0.69) дальше мы ещё видим \↑ф̲ермераш,
290.92	pN-066	(0.69)
291.61	N-vE173	(в 0.11) который \↑спуска̲ешь,
293.21	pN-067	(0.38)
293.59	N-vE174	с \↑л̲естницы,
294.23	N-vE175	и видит что-о /ˀодной из корзин не \хват̲аетʰ.

Количественное распределение рассмотренных в разделе 4 типов представлено в табл. 1. Как видно, к ядерному типу относится немногим менее половины случаев, следующими по частотности оказались конструкции с парентетическим *вот*, далее — конструкции с псевдопарентетическими вставками. Также можно отметить, что три наиболее часто встречающихся типа по-разному распределены по значениям параметра «уровень дискурсивной структуры»: парентетические *вот* встречаются только между ЭДЕ; для ядерного типа расположение вставки внутри ЭДЕ возможно, но значительно менее частотно, чем позиция между ЭДЕ; псевдопарентетические вставки с равной частотой возникают между и внутри ЭДЕ.

Тип вставочной конструкции	**Уровень дискурсивной структуры**		
	Между ЭДЕ	Внутри ЭДЕ	***Всего***
Ядерный тип	100	14	114 (48.3%)
Со «свободными» вставками	8	2	10 (4.2%)
С парентетическим *вот*	68	0	68 (28.8%)
С псевдопарентетическими вставками	19	20	39 (16.5%)
Прочие	2	3	5 (2.1%)
ВСЕГО	197	39	236 (100%)

Табл. 1. Количественное распределение типов конструкций с дискурсивными вставками в проанализированном корпусе: общее и отдельные по уровням дискурсивной структуры

5 Темп и регистр во вставках: предварительный анализ

В литературе, посвященной парентезе, неоднократно отмечалось, что вставки характеризуются рядом просодических признаков: тихим и ускоренным произнесением, сниженным или сжатым тональным регистром и др. [Crystal 1969; Цеплитис 1974; Kutik et al. 1983; Bolinger 1989; Wichmann 2001; Гавриленко 2004; Кибрик, Подлесская 2010; Dehé 2014; и др.]. Для проверки этих утверждений на корпусном материале в рамках настоящего исследования был проведен дополнительный акустический анализ двух просодических характеристик вставок относительно левого и правого контекста: темпа и тонального регистра.

В качестве меры темпа была выбрана средняя продолжительность слога в миллисекундах, в качестве меры тонального регистра — ширина тонального диапазона, которая рассчитывалась как разность между 90-м и 10-м процентилем в ряду значений ЧОТ, определяемых в объекте Pitch программы Praat. Соответственно, под коэффициентом ускорения понималось отношение темпа во вставке к темпу в левом / правом контексте; под коэффициентом сжатия тонального диапазона — отношение ширины диапазона во вставке к ширине диапазона в левом / правом контексте. Так, в приведенном выше примере (4) коэффициент ускорения вставки *так как был пьяный, а за рулём как известно нельзя пить* по отношению как к левому, так и к правому контексту составил 1.10; сжатие тонального диапазона составило 3.13 по отношению к левому контексту и 3.37 по отношению к правому (ср. наблюдаемый на тонограмме контраст на рис. 2; указанная разница соответствует 8-9 полутонам).

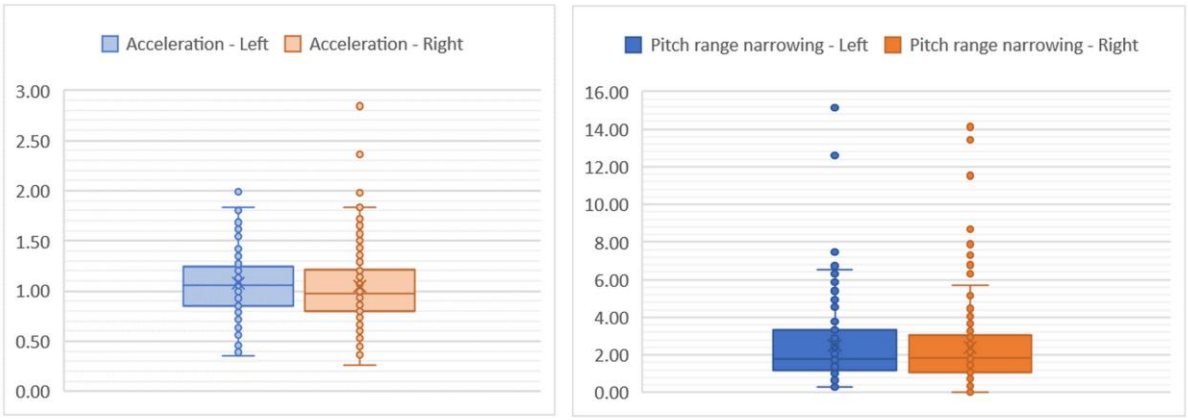


Рис. 3. Диаграммы размаха просодических характеристик вставок. Слева направо: коэффициент ускорения относительно левого контекста, коэффициент ускорения относительно правого контекста, коэффициент сжатия тонального диапазона относительно левого контекста, коэффициент сжатия тонального диапазона относительно правого контекста

Поскольку в конструкциях с парентетическим *вот* вставки состоят всего из одного слога, эти примеры были исключены из выборки. Для оставшихся 168 примеров были подсчитаны значения четырех коэффициентов; диаграммы размаха представлены на рис. 3. При проверке одновыборочным t-тестом статистическую значимость показали сжатие тонального диапазона относительно левого и правого контекста (M = 2.50 и 2.39 соответственно; $p < 0.0001$; умеренный размер эффекта по d Коэна), а также ускорение относительно левого контекста (M = 1.08; $p = 0.002$; малый размер эффекта). Значимых корреляций между темповыми и тональными параметрами не обнаружено, что можно интерпретировать как указание на независимость этих двух просодических характеристик.

Кроме того, были выявлены некоторые предварительные закономерности, связанные с параметрами классификации вставочных конструкций, рассмотренными выше в разделах 3 и 4:

- Сжатие тонального регистра относительно правого контекста статистически значимо для вставок ядерного типа, но не для псевдопарентетических вставок.
- Вставки внутри ЭДЕ (см. пример (3)) демонстрируют бóльшую тенденцию к ускорению относительно левого контекста, чем вставки между ЭДЕ. При этом данное различие статистически значимо только для псевдопарентетических вставок ($p = 0.02$), но не для вставок ядерного типа ($p = 0.27$).
- Для сжатия тонального диапазона наблюдается противоположная картина: эта характеристика более заметно реализуется во вставках между ЭДЕ, чем во вставках внутри ЭДЕ ($p = 0.02$).

Содержательная интерпретация полученных количественных результатов, по всей видимости, станет возможна при увеличении объема выборки. Предварительно можно заключить, что из двух рассмотренных просодических характеристик вставок на нашем материале более последовательно проявляется сжатие тонального диапазона. Несколько более подробный анализ просодических характеристик вставок ядерного типа содержится в [Коротаев 2021].

6 Заключение

В настоящей работе вставка, или парентеза, рассматривается как явление дискурсивного уровня. Вставочная природа языковой единицы — это характеристика конкретной дискурсивной ситуации, в которой вставка противопоставляется окружающему контексту. Соответственно, интегральные и дифференцирующие свойства парентетических конструкций обусловлены (а) внутренними свойствами объемлющего контекста; (б) внутренними свойствами вставленных единиц; (в) характером взаимоотношения вставки и контекста. При определении понятия на материале неподготовленной устной речи за основу было взято представление о вставке как об отдельном

дискурсивном шаге, в котором временно откладывается проецированное в левом контексте продолжение основной линии изложения. Таким образом, парентеза становится компонентом одной из возможных стратегий «разрешения» дискурсивной незавершенности, см. [Коротаев 2018].

Согласно собранным корпусным данным, прототипические вставки содержательно соотносятся с тем или иным элементом левого контекста, а также произносятся с нисходящим или ровным тональным акцентом. Отклонения от этого прототипа возможны по обоим параметрам, но наиболее заметным оказывается использование во вставке восходящей интонации — если она не связана с выражением самостоятельного иллокутивного значения. Просодические характеристики вставок, противопоставляющие их левому и правому контексту, включают в себя ускоренное произнесение и сжатие тонального диапазона, при этом второе свойство реализуется в наших данных более рельефно, чем первое.

Благодарности

Исследование выполнено при финансовой поддержке РНФ, проект № 17-18-01184.

Литература

[1] Auer, Peter. (2005). Projection in Interaction and Projection in Grammar, Text - Interdisciplinary Journal for the Study of Discourse, Vol. 25(1), pp. 7–36. https://doi.org/10.1515/text.2005.25.1.7.
[2] Boersma, Paul, Weenink, David (2018). Praat: Doing phonetics by computer (6.0.43) [Computer software]. Access mode: https://www.fon.hum.uva.nl/praat/.
[3] Bogdanova, Natalia V. (2011). Parenthetical constructions in spoken spontaneous monologues [Vstavnye konstrukcii v zvučaščem spontannom monologe], Studies in Speech Culture [Voprosy kul'tury reči], Iss. 10, pp. 204-212.
[4] Bolinger, Dwight (1989). Intonation and Its Uses: Melody in Grammar and Discourse. Stanford, Stanford University Press.
[5] Ceplītis, Laimodts (1974). Analysis of speech intonation [Analiz rečevoj intonacii]. Riga, Zinātne.
[6] Crystal, David (1969). Prosodic Systems and Intonation in English. Cambridge, Cambridge University Press.
[7] Daragan, Yulia V. (2003). Parasitism or symbiosis? How speakers deal with communication failures and verbal means they use to do that [Parazitizm ili simbioz: Mexanizm preodolenija kommunikativnyx sboev i obsluživajuščie ego verbal'nye sredstva], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" [Komp'juternaja lingvistika i intellektual'nye texnologii: po materialam ežegodnoj meždunarodnoj konferencii "Dialog"], pp. 166–178.
[8] Dehé, Nicole (2014). Parentheticals in Spoken English: The Syntax-Prosody Relation. Cambridge, Cambridge University Press. https://doi.org/10.1017/CBO9781139032391
[9] Dehé, Nicole, Kavalova, Yordnaka (Eds.) (2007). Parentheticals. John Benjamins. https://doi.org/10.1075/la.106
[10] Gavrilenko, Irina I. (2004). Supplementary information in academic texts: Semantic, syntactic, and prosodic features (based on Russian data) [Dopolnitel'naja informacija v naučnyx tekstax: Semantičeskie, sintaksičeskie i prosodičeskie osobennosti (na materiale russkogo jazyka)], PhD Dissertation Abstract. Moscow, Moscow State University.
[11] Kavalova, Yordanka (2007). And-parenthetical clauses, N. Dehé, Y. Kavalova (eds.) Parentheticals. John Benjamins, pp. 145–172. https://doi.org/10.1075/la.106.09kav
[12] Kibrik, Andrej A. (2018). Russian multichannel discourse. Part II. Corpus development and avenues of research [Russkij mul'tikanal'nyj diskurs. Čast' II. Razrabotka korpusa i napravlenija issledovanij], Psychological Journal [Psixologičeskij Žurnal], Vol 39(2), pp. 79–90. https://doi.org/10.7868/80205959218020083
[13] Kibrik, Andrej A., Korotaev, Nikolay A., Podlesskaya, Vera I. (2020). Russian spoken discourse: Local structure and prosody, S. Izre'el, H. Mello, A. Panunzi, T. Raso (eds.) In search of basic units of spoken language: A corpus-driven approach. John Benjamins, pp. 37–76. https://doi.org/10.1075/scl.94.01kib
[14] Kibrik, Andrej A., Podlesskaya, Vera I. (2010). Parenthetical construction in spoken discourse [Vstavočnye konstrukcii v ustnom diskurse], V. Z. Demyankov, V. Ja. Porxomovskij (eds.) In language and culture: Sound, sign, meaning. For the 70th birthday of Viktor A. Vinogradov [V prostranstve jazyka i kul'tury: Zvuk, znak, smysl. Sbornik statej v čest' 70-letija V. A. Vinogradova]. Moscow, LSC, pp. 87–99.
[15] Kobozeva, Irina M. (1999). On two types of introductory constructions with parenthetical verbs [O dvux tipax vvodnyx konstrukcij s parentetičeskim glagolom], E. V. Raxilina, Ja. G. Testelets (eds.) Typology and Lingustic theory: From description to explanation. For the 60th birthday of Aleksandr E. Kibrik [Tipologija i teorija jazyka: Ot opisanija k ob''jasneniju. K 60-letiju A. E. Kibrika]. Moscow, LSC, pp. 539–543.

[16] Kodzasov, Sandro V. (2009). Studies in Russian Prosody [Issledovanija v oblasti russkoj prosodii]. Moscow, LSC.
[17] Korotaev, Nikolay A. (2018). How intonation structure spoken narratives: Non-final phase contexts [Intonacionnaja struktura ustnogo rasskaza v kontekste nezaveršennosti], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" [Komp'juternaja lingvistika i intellektual'nye texnologii: po materialam ežegodnoj meždunarodnoj konferencii "Dialog"], Vol. 17(24), pp. 342–356.
[18] Korotaev, Nikolay A. (2021). Tempo and pitch properties of parenthetical constructions in spontaneous Russian discourse [Temp i tonal'nyj registr v konstrukcijax so vstavkami v nepodgotovlennom ustnom diskurse], Analysis of the Russian colloquial speech [Analiz razgovornoj russkoj reči], St. Petersburg. In print.
[19] Kutik, Elanah J., Cooper, William E., Boyce, Suzanne (1983). Declination of fundamental frequency in speakers' production of parenthetical and main clauses, The Journal of the Acoustical Society of America, Vol. 73(5), pp. 1731–1738. https://doi.org/10.1121/1.389397
[20] Mazeland, Harrie (2007). Parenthetical sequences, Journal of Pragmatics, Vol. 39(10), pp. 1816–1869. https://doi.org/10.1016/j.pragma.2007.05.005
[21] McCawley, James D. (1982). Parentheticals and discontinuous constituent structure, Linguistic Inquiry, Vol. 13(1), pp. 91–106.
[22] Paducheva, Elena V. (2010). Semantic studies: Semantics of tense and aspect in Russian; Semantics of the narrative [Semantičeskie issledovanija: Semantika vremeni i vida v russkom jazyke; Semantika narrativa] (2nd ed.). Moscow, LSC.
[23] Schegloff, Emanuel A. (2007). Sequence Organization in Interaction: A Primer in Conversation Analysis, Vol. 1. Cambridge, Cambridge University Press. https://doi.org/10.1017/CBO9780511791208
[24] Urmson, James O. (1952). Parenthetical verbs, Mind, Vol. 61(244), pp. 480–496.
[25] Wichmann, Anne (2001). Spoken parentheticals, K. Aijmer (Ed.), A Wealth of English: Studies in Honour of Goran Kjellmer. Gothenburg, Gothenburg University Press, pp. 177-193.
[26] Yanko, Tatiana E. (2001). Communicative strategies of Russian speech [Kommunikativnye strategii russkoj reči]. Moscow, LSC.
[27] Yanko, Tatiana E. (2008). Intonational strategies in spoken Russian from a comparative perspective [Intonacionnye strategii russkoj rechi v tipologicheskom aspekte]. Moscow, LSC.

Audio and Text-Driven approach for Conversational Gestures Generation

Korzun V. A.
MIPT
Moscow, Russia
korzun@phystech.edu

Dimov I. N.
MSU
Moscow, Russia
iliyadimov@icloud.com

Zharkov A. A.
MIPT
Moscow, Russia
andrey.zharkov@phystech.edu

Abstract

This paper describes FineMotion's gesture generating system entry for the GENEA Challange 2020. We start by using simple baselines and expand them by using context and combining both audio and textual features. Among the participating systems, our entry attained the highest median score in the human-likeness evaluation and second highest median score in appropriateness.

Keywords: embodied agents, neural networks, gesture generation, social robotics, deep learning, word embeddings

DOI: 10.28995/2075-7182-2021-20-425-432

Генерация разговорных жестов на основе речи и текста

Корзун В. А.
МФТИ
Москва, Россия
korzun@phystech.edu

Димов И. Н.
МГУ
Москва, Россия
iliyadimov@icloud.com

Жарков А. А.
МФТИ
Москва, Россия
andrey.zharkov@phystech.edu

1 Introduction

Gestures are often underrated in human communication. They may contribute a lot to a speech going as far as to change what is being said to the opposite: a simple shrug can make audience question the credibility of the speech. Humans actively use co-speech gestures to convey their emotions or visualize their attitude [5, 9].

The task of generating conversational motions can be used for social robots [16], conversational agents, and even automatic animation of virtual characters. Both rule-based and deep learning approaches have been employed to varying degrees of success. In this work, we propose several models to solve this problem as well as analyze what makes movement seem appropriate and indistinguishable from humans and which features are essential for such a task.

The GENEA Challenge [17] was conducted to explore what kind of models can produce human-like behavior for motion generation. The challenge organizers shared a 3.5-hour long dataset of audio, transcripts, and corresponding motions for body movement as well as several strong baselines. They also conducted a human evaluation of generated motions, consisting of 250 experts.

Our systems were initially built upon baselines [1, 12] provided by organizers. We made several architectural adjustments, but conceptually the core of our systems was not dissimilar from aforementioned models. Our main contributions are adding contextual information and combining both textual and audio information in one model.

Our paper is organized in the following way: section 2 describes related work; section 3 describes data preprocessing, which is shared between all experiments; section 4 describes our models; section 5 contains the discussion of our results; and section 6 contains the conclusion. Our code is publicly available[1] to help other researchers reproduce our results. Our repository also contains a link to trained weights and videos of generated motions.

The dataset used in all experiments is described in [3]. A complete task description along with evaluation of systems proposed for the workshop is described in [17]. Our team was labeled as SD for anonymization purposes.

2 Related work

In [12] authors consider motion generation problem as a mapping of sequence of words to a sequence of human poses. To solve this problem they used sequence to sequence model [15] with soft attention mechanism. The encoder processes input sequence of words which then transmitted to the decoder to generate gesture motions. Word-level features are represented by GloVe [10] embeddings. Gestures are represented by 10 principal components converted from OpenPose [11] features by Principal Component Analysis (PCA). Their sequence to sequence model also has several modifications: decoder hidden state is initialized by hidden state from previous sequence to make series of poses continuous. They also use modified loss

$$\mathcal{L} = \mathcal{L}_{mse} + \alpha \cdot \mathcal{L}_{continuity} + \beta \cdot \mathcal{L}_{variance}$$

where $\mathcal{L}_{mse}$ is a mean squared error, $\mathcal{L}_{continuity}$ is defined as

$$\mathcal{L}_{continuity} = \frac{\sum_{t=2}^{m} \|p_t - p_{t-1}\|}{m-1}$$

where p_t is a pose at time step t. $\mathcal{L}_{variance}$ is defined as negative of the variance o p_t.

In [1] authors consider a slightly different problem: given a sequence of speech features $s = [s_t]_{t=1:T}$ extracted from frames of speech audio at regular intervals t, the task is to generate a corresponding gesture sequence $\hat{g} = [\hat{g}_t]_{t=1:T}$. They use MFCC [2] features to represent audio and features learned by Denosing Auto Encoder to represent gestures. Authors use a recurrent neural network to encode a window of audio features, then this representation of the window used to generate a single frame of gestures. Savitsky-Golay filter [13] is used for smoothing the final predictions.

In the research paper [6] authors propose a GAN approach to gesture generation. They use a 1D UNet for MFCC to motion translation. The discriminator is used to avoid regressing to a mean pose.

3 Data preprocessing

The challenge organizers provided 23 recordings with an overall length of 3 hours and 40 minutes for training. Each recording consists of an audio file with speech recording, text transcripts, and BVH (bounding volume hierarchy) file with the motion data. The initial motion was captured by 60 frames per second; the generated motions for evaluation were rendered at 20 frames per second. The motion skeleton contained 71 joints, but we used only 15 points corresponding to the upper body without hands and fingers.

We split the dataset for training and validation in the following way: the first recording *Recording_001* was used for validation (12 minutes), while the rest of the recordings were used for training (3 hours 28 minutes total). As the evaluation process is rather long, we used only 1 minute of *Recording_001* for human evaluation, and the remaining part of the sample was used to calculate mean squared error on joints as a sanity check.

[1] https://github.com/FineMotion/GENEA_2020

For all our models we used the same audio and motion data preparation pipeline provided in one of the baselines [1]. For audio representation we used MFCC. We then averaged every five consequent Mel features to align audio features with motions (so that they have 20 FPS each). We represent motion data by 3 dimensional axis-angle rotation vectors for 15 joints. Thus each motion frame has 45 float features. This values are normalized over the mean value on train dataset. All aforementioned transformations of data result in input audio feature matrices to have size (N, 26) and output motion matrices to have size (N, 45), where *N* represents the number of frames in the sample.

We use the term "context window". The context window consists of 61 frames centered around a certain point in time, represented by a frame. We also use a "mean pose" calculated from the training dataset to use it as a starting value in recurrent models.

For paddings we used the MFCCs of silence recording. In text-based models we also used text features in form of GloVe embedding for words in context window.

For all proposed models we smoothed generated motions by applying the Savitzky-Golay filter to them. The length of the filter window and the order of the polynomial are 9 and 3, respectively. We did not use any external data.

4 Proposed models

The task of generating a motion can be summarized in the following way: given a set of audio features $A = (A_1, A_2, ..., A_n)$ and words $W = (W_1, W_2, ..., W_k)$ predict a corresponding set of motions $M = (M_1, M_2, ..., M_n)$.

$$f(W, A) = M \tag{1}$$

During training we minimize MSE loss and use it to assess model convergence. Our final loss functions are modified by additional terms which are described in corresponding model sections.

4.1 Sequence to sequence model

Our first described model is a sequence to sequence model, which is a reimplementation of [12] on sound-based features. The model in the aforementioned paper used words to generate corresponding motions. The competition dataset provides audio, motions and words. Three seconds of speech correspond to 60 poses and usually contain less than 10 words. We have decided to build our system on audio features and use textual information to further improve the quality of the models. Aside from difference in density between the two sets of features, speech obviously conveys more information like emotions, pauses, voice crackling, which are usually lost in text-to-speech systems.

As motions and audio features are mapped on a one-to-one basis, our first model is a simple seq2seq [15] consisting of GRU [7] encoder and decoder over audio and motions. This baseline system is illustrated on Figure 1. The encoder takes several audio features (MFCC) $A_{i-k}...A_i$ from corresponding frames, encodes them into a higher dimensional space represented by $AE_{i-k}..AE_i$ and passes it to the decoder, which predicts the following motions labeled $M_{i+1}...M_m$. Decoder's final layer combines decoder hidden state and encoder-decoder dot-product attention [8] to make a motion prediction. As the decoder requires a pose as the first input, we supply it a previously predicted pose or the "mean pose" if no previous poses are available.

We tried to further improve this model by adding a word encoder, which is illustrated in the dotted box in Figure 1. The simultaneous use of words and audio features has the problem of alignment. The GENEA 2020 challenge dataset had a transcript with words and corresponding time regions. Such markup is hard to annotate. Moreover while it helps to map words to various time windows there is still a problem of combining just a few words and multiple audio features. We use attention mechanism between the two representations for automatic alignment.

Words are embedded using GloVe [10] and are passed to another GRU. The hidden state of word-level encoder is not directly passed to the decoder, but a second encoder-decoder attention vector is calculated, which is supplied to the final layer of the decoder, to make prediction based both on audio, previous poses and words. The words are taken from a 2-second window.

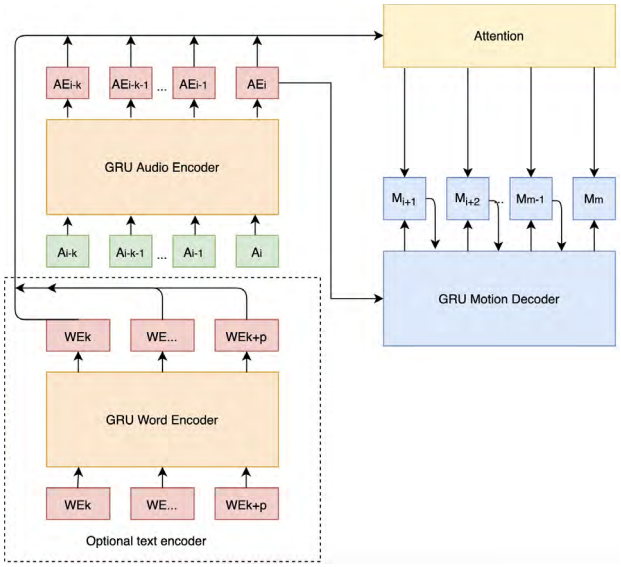


Figure 1: Scheme of baseline seq2seq model on audio features with optional word-level encoder.

We tuned several hyperparameters and training strategies. As the authors in [12] we employed continuity loss and variance loss to make the generated motions more fluid and natural. The addition of variance loss significantly improved co-speech gesture quality. We trained model with learning rate of 0.001 using Adam optimizer; audio encoder was a 2-layered bidirectional GRU with the hidden dimension of 150 units; word encoder was a single-layered GRU, both input and output dimensions were set to 100 units; decoder was a single-layered GRU with hidden dimension of 150 units. The model was trained for 100 epochs with a batch size of 512, where each sample contained 10 previous poses and 20 poses for prediction.

We also explored various combinations of windows sizes for encoder and decoder. We did not find larger windows to be beneficial to the quality of our predictions and we kept the same window sizes as in the original paper: we use 10 previous frames to predict the following 20 frames.

Another strategy we tried to employ is a variation of scheduled sampling [14]. During training our autoregressive decoder models a following function:

$$f(h_{i-1}, \tilde{m}_{i-1}, E) = m_i, \tag{2}$$

where h_{i-1} is decoder's hidden state, $\tilde{m}_{i-1}$ is the true motion on a previous time step and E corresponds to encoder states. During teacher forcing we replace the real motion $\tilde{m}_{i-1}$ with previously generated motion m_{i-1} with a probability of 0.5. The main idea behind it is to help the model to explore the error space and become more robust. In the end we found out that not supplying real poses at all was the best option and the rest of our models are using their own predictions during training, just as it would happen during inference. This may be attributed to variance loss: the model was rewarded for making different poses, which likely resulted in a pretty constant deviation from true poses.

We also do not save hidden state of encoder and decoder between batches during training. Each training sample is processed individually without knowledge of previous time period, but during inference the model always supplies it's state for the next segment. This may be the reason behind choppiness in predicted movement. We used smoothing to eliminate this shortcoming.

We'd like to state that our evaluation of hyperparameters is rather subjective: all the changes were judged by a small group of people on a one-minute sample from the validation recording. It is quite possible that we misjudged some of our experiments because of an unsuitable time sector or a simple human error.

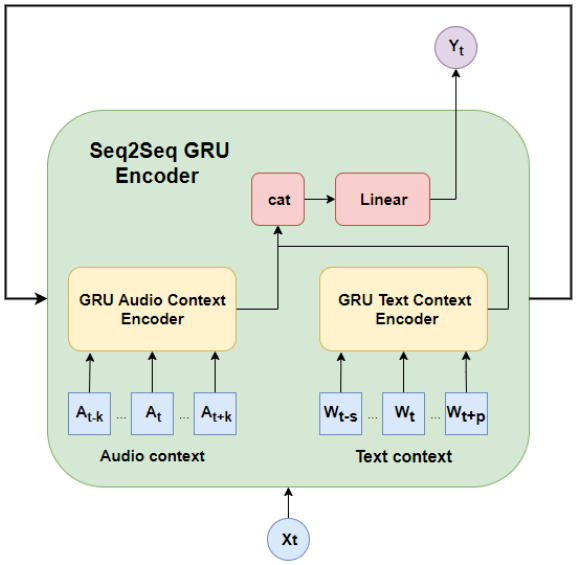


Figure 2: Scheme of contextual encoder.

4.2 Contextual encoder

The second model is inspired by [1]. We have decided to keep sequence to sequence model and enhance it with contextual representations. In our basic sequence to sequence encoder each input corresponds to a single frame.

We decided to represent each frame as a 3-second window around it, which resulted in 61 frames. We used two additional GRU encoders to encode the audio and textual context window as displayed on Figure 2. The audio encoder consists of 3 linear layers with batch normalization and RELu activation. Those layers are used to project audio features for the one-layer one-directional GRU. All audio encoder layers have hidden size 150. The textual encoder is bidirectional one-layer GRU over GloVe embeddings and hidden size of encoder is similar to the embeddings size which is 100. The outputs of both context encoders are concatenated and projected to be passed as inputs to the seq2seq encoder with hidden dimension of 150 units. The rest of the model is a simple sequence to sequence architecture with attention, which was described earlier.

We train this model with Adam optimizer with the learning rate of 0.001 and the batch size of 50. The final model was trained for the 100 epochs, however the target loss stabilized after 80th epoch. Furthermore, motions generated after 80th and 100th epochs were virtually identical.

4.3 Adversarial training

Even a single speaker has a significant variation of his movements even in extremely similar situations, same phrases and contexts. However, so far we described only models which tried to recreate the same movements as the ground truth, even if it was not the only correct behaviour, but one of the many possible motions. To try to overcome this problem we used adversarial training (as done, i.e. in [6, 4]).

The generator model produces motions from audio, while discriminator model tries to classify real and generated motions. The generator loss is

$$L_G = L_{base}(G) + \lambda L_{adv}(G, D), \tag{3}$$

where L_{base} contains whatever non-adversarial components of generator loss and L_{adv} represents adversarial loss with weight λ. In all of our experiments we used non-saturating GAN loss.

We tried several discriminator models based on blocks of (1D convolution, 1D batch normalization, LeakyRelu(0.2)). After series of that blocks we flatten the outputs of convolutional block and apply two

more linear layers. We varied total number of blocks from 2 to 6 with at least two of them reducing spatial dimension (stride > 1).

Unfortunately, the training with adversarial loss was not stable (especially for relatively high λ values around 10.0). Sometimes we got interesting and diverse results (mostly for small λ values around 0.1), however the quality was still lacking in comparison with our best model so in the final system adversarial training was not used.

5 Results and discussion

The challenge organizers used two human-evaluation metrics for evaluation:

- **Human-likeness** - the generated motion should be realistic for human. The evaluation participants should score the motion file without audio by this criterion.
- **Appropriateness** - the generated motion should match the corresponding audio. So participants score motion with audio.

Summary statistics (sample median and sample mean) were provided in [17] and are listed in table 1. The challenge organizers provided results for all participating systems (with label SX), baselines (BA[1] and BT[12]), natural (N) and mismatched (M) motion capture. Our system is labeled as SD.

	Human-likeness		Appropriateness	
ID	Median	Mean	Median	Mean
N	$72 \in [70, 75]$	67.6 ± 1.8	$81 \in [79, 83]$	73.8 ± 1.8
M	"	"	$56 \in [53, 59]$	53.3 ± 2.0
BA	$46 \in [44, 49]$	46.2 ± 1.7	$40 \in [38, 41]$	40.4 ± 1.8
BT	$55 \in [53, 58]$	54.6 ± 1.8	$38 \in [35, 40]$	38.5 ± 1.9
SA	$38 \in [35, 41]$	40.1 ± 1.9	$35 \in [31, 37]$	36.4 ± 1.9
SB	$52 \in [50, 55]$	52.8 ± 1.9	$43 \in [40, 45]$	43.3 ± 2.0
SC	$57 \in [55, 60]$	55.8 ± 1.9	$50 \in [48, 52]$	50.6 ± 1.9
SD	$60 \in [57, 61]$	58.8 ± 1.7	$49 \in [46, 50]$	48.1 ± 1.9
SE	$49 \in [47, 51]$	49.6 ± 1.8	$47 \in [44, 49]$	45.9 ± 1.8

Table 1: Summary statistics of user-study ratings

Our systems were built upon baselines, which allows us to estimate the importance of proposed modifications. Our final submission has the highest median score among the participating systems and baselines. It also has the second highest score by appropriateness. Although our system shows strong improvement over baselines, it is still far behind human generated motions.

Challenge organizers used a special set of mismatched audios and human motions during the human evaluation. This approach was not surpassed by any system. That means that our synthetic generated motions are significantly less appropriate than random human movement.

To select the best model we compared them on validation data using human evaluation among the members of our team. The seq2seq model with contextual encoder was unanimously chosen as the best model, however seq2seq with attention over text and audio was a close second.

We found out that our team was looking for specific sorts of movements during the motion evaluation: we generally were looking for correspondence between motions and verbal pauses. We were more inclined to vivid movements, even if they were choppy, and last but not least - we were always looking for fast and sharp movements coinciding with loud and aggressive speech patterns.

Our humble human evaluation has come to a conclusion, that the approach with context encoder helps to make generated motions smoother, because is uses more information, especially for the last frames in a sequence, while basic seq2seq heavily relies on smoothing.

6 Conclusion

In this paper we proposed several modifications for existing approaches in co-speech gesture generation. In our approach we combined text and audio features and thus were able to outperform text- and

audio-only baselines. Our models were rated highest for Human-likeness metric and second highest for appropriateness, however compared with the real data (human gestures) there is a striking gap in our system's performance and real motions, meaning that there is still a lot to be improved.

Our team also did not explore various sound preprocessing techniques, which could result in a more high-dimensional vector input representation, which would allow models to extract a more rich set of features.

We believe that future research should focus on multimodal representations. We also believe that the quality of generated motions will increase with the expansion of the dataset, which will enable the researchers to train more sophisticated models, like GANs or transformers.

7 Acknowledgements

The reported study was funded by RFBR according to the research project № 20-31-90051

References

[1] Analyzing input and output representations for speech-driven gesture generation / Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter et al. // Proceedings of the ACM International Conference on Intelligent Virtual Agents. — IVA '19. — 2019. — P. 97–104.

[2] Davis Steven, Mermelstein Paul. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE transactions on acoustics, speech, and signal processing. — 1980. — Vol. 28, no. 4. — P. 357–366.

[3] Ferstl Ylva, McDonnell Rachel. Investigating the use of recurrent motion modelling for speech gesture generation // IVA '18 Proceedings of the 18th International Conference on Intelligent Virtual Agents. — 2018. — Nov. — Access mode: `https://trinityspeechgesture.scss.tcd.ie`.

[4] Generative adversarial nets / Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza et al. // Advances in neural information processing systems. — 2014. — P. 2672–2680.

[5] Knapp Mark L, Hall Judith A, Horgan Terrence G. Nonverbal communication in human interaction. — Cengage Learning, 2013.

[6] Learning individual styles of conversational gesture / Shiry Ginosar, Amir Bar, Gefen Kohavi et al. // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2019. — P. 3497–3506.

[7] Learning phrase representations using RNN encoder-decoder for statistical machine translation / Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre et al. // arXiv preprint arXiv:1406.1078. — 2014.

[8] Luong Minh-Thang, Pham Hieu, Manning Christopher D. Effective approaches to attention-based neural machine translation // arXiv preprint arXiv:1508.04025. — 2015.

[9] Matsumoto David, Frank Mark G, Hwang Hyi Sung. Nonverbal communication: Science and applications. — Sage Publications, 2012.

[10] Pennington Jeffrey, Socher Richard, Manning Christopher D. Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — P. 1532–1543.

[11] Realtime multi-person 2d pose estimation using part affinity fields / Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 7291–7299.

[12] Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots / Young-woo Yoon, Woo-Ri Ko, Minsu Jang et al. // Proceedings of the IEEE International Conference on Robotics and Automation. — ICRA '19. — 2019. — P. 4303–4309.

[13] Savitzky Abraham, Golay Marcel JE. Smoothing and differentiation of data by simplified least squares procedures. // Analytical chemistry. — 1964. — Vol. 36, no. 8. — P. 1627–1639.

[14] Scheduled sampling for sequence prediction with recurrent neural networks / Samy Bengio, Oriol Vinyals, Navdeep Jaitly, Noam Shazeer // Advances in Neural Information Processing Systems. — 2015. — P. 1171–1179.

[15] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to sequence learning with neural networks // Advances in neural information processing systems. — 2014. — P. 3104–3112.

[16] To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability / Maha Salem, Friederike Eyssel, Katharina Rohlfing et al. // International Journal of Social Robotics. — 2013. — Vol. 5, no. 3. — P. 313–323.

[17] A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020 / Taras Kucherenko, Patrik Jonell, Youngwoo Yoon et al. // 26th International Conference on Intelligent User Interfaces. — 2021. — P. 11–21.

Current Landscape of the Russian Sentiment Corpora

Kotelnikov E. V.
Vyatka State University, Kirov, Russia;
ITMO University, Saint Petersburg, Russia
kotelnikov.ev@gmail.com

Abstract

Currently, there are more than a dozen Russian-language corpora for sentiment analysis, differing in the source of the texts, domain, size, number and ratio of sentiment classes, and annotation method. This work examines publicly available Russian-language corpora, presents their qualitative and quantitative characteristics, which make it possible to get an idea of the current landscape of the corpora for sentiment analysis. The ranking of corpora by annotation quality is proposed, which can be useful when choosing corpora for training and testing. The influence of the training dataset on the performance of sentiment analysis is investigated based on the use of the deep neural network model BERT. The experiments with review corpora allow us to conclude that on average the quality of models increases with an increase in the number of training corpora. For the first time, quality scores were obtained for the corpus of reviews of ROMIP seminars based on the BERT model. Also, the study proposes the task of the building a universal model for sentiment analysis.

Keywords: sentiment analysis, text corpora, deep learning, BERT
DOI: 10.28995/2075-7182-2021-20-433-444

Текущее состояние русскоязычных корпусов для анализа тональности текстов

Котельников Е. В.
Вятский государственный университет, Киров, Россия;
Университет ИТМО, Санкт-Петербург, Россия
kotelnikov.ev@gmail.com

Аннотация

В настоящее время существует более десятка русскоязычных корпусов для анализа тональности, отличающихся источником текстов, предметной областью, размерами, количеством и соотношением классов тональности, способом разметки. В работе рассматриваются общедоступные русскоязычные корпуса, приводятся их качественные и количественные характеристики, позволяющие составить представление о текущем состоянии корпусов для анализа тональности. Предлагается ранжирование корпусов по качеству разметки, которое может быть полезно при выборе корпусов для обучения и тестирования. Исследуется влияние обучающей выборки на качество анализа тональности на основе применения глубокой нейросетевой модели BERT. Эксперименты с корпусами отзывов позволяют сделать вывод о том, что при увеличении количества обучающих корпусов качество моделей в среднем повышается. Впервые получены оценки качества для корпусов отзывов семинаров РОМИП на основе модели BERT. Также ставится задача построения универсальной модели для анализа тональности.

Ключевые слова: анализ тональности, текстовые корпуса, глубокое обучение, BERT

1 Introduction

Currently, the text sentiment analysis is still an urgent problem. Despite the fact that modern deep neural network models allow for some datasets to reach an accuracy close to 100% in the case of binary classification (positive/negative) of the SST-2 English movie review corpus [8], with the number of classes more than two the accuracy does not exceed 60% [21].

The most important factor in the construction of sentiment analysis systems is the availability of a variety of high-quality text corpora. The English corpora for sentiment analysis have been fairly well researched [25]. The first Russian-language text corpora devoted to sentiment analysis appeared in 2011. These are three corpora of reviews for books, movies and cameras prepared for the ROMIP (Russian Information Retrieval Evaluation Seminar) competition [4]. Over the past 10 years, more than a dozen Russian-language corpora have been annotated by sentiment and made available for public access.

Russian-language corpora, as opposed to English-language, despite some recent works [7, 20], have not been sufficiently researched yet. In particular, there are no works devoted to the analysis of the quality of corpora, as well as studies of the dependence of the models' quality on the training corpora. There are also no performance scores of the modern deep neural network models for the review corpora of the ROMIP competitions.

The most important characteristics of the corpora intended for sentiment analysis are the source of texts, the domain, the size of the corpus, the size of the texts, the number and ratio of sentiment classes, the annotation method, the presence of a split into training and test parts. This paper examines the existing Russian-language publicly available corpora, annotated by sentiment.

The contribution of this work is as follows:

- an overview of all publicly available Russian-language corpora with detailed characteristics is provided;
- the ranking of corpora by annotation quality is proposed;
- new quality ratings have been obtained for the existing Russian-language corpora of reviews;
- the influence of the training dataset on the performance of the sentiment analysis of reviews was investigated.

The study also proposes the task of the building a universal model for sentiment analysis.

The rest of the paper is organized as follows. The second section reveals the characteristics of existing corpora. The third section is devoted to the materials and methods used in the experimental study. In the fourth section, the results of the experiments are presented and discussed. The fifth section provides an overview of previous works on Russian-language corpora for sentiment analysis. In the final section the conclusions are drawn and directions for further research are indicated.

2 The Russian text corpora for sentiment analysis

2.1 Characteristics of corpora

This section discusses existing Russian-language corpora for sentiment analysis. As noted in the Introduction, the main characteristics of the corpora are the source of the texts, the domain, the size of the corpus, the size of the texts, the number and ratio of sentiment classes, the annotation method, the presence of a split into training and test parts.

Sources of the texts. All corpora can be divided into four sources of the texts: 1) reviews of products, works of art and organizations; 2) tweets; 3) posts on social networks; 4) news articles. The source of the texts determines the domain, style and size of the texts.

The domain defines the topic of the texts, for example, restaurant reviews or political news. There are corpora without an explicitly defined domain (for example, RuTweetCorp).

Corpus sizes vary from several dozen (RuSentRel) to hundreds of thousands of texts (RuTweetCorp). *The size of the texts* is related to the source and ranges from a few words (for tweets) to several thousand words (for reviews, news and social media posts).

Sentiment classes. A one-dimensional scale is often used to indicate sentiment (Figure 1a). There are 6 main classes: positive, weakly positive, negative, weakly negative, neutral and contradictory. However, in the case of the one-dimensional scale, uncertainty arises with an intermediate (zero) value, which can have the meaning of a neutral sentiment (the absence of sentiment) or contradictory (presence of both positive and negative sentiments). Therefore, it is more convenient to represent the sentiment classes on a plane (the boundaries between the classes are shown conditionally; the symmetry of the positive and negative sentiment is assumed) (Figure 1b).

The minimum number of classes in existing corpora is two – positive and negative (hereinafter referred to as "+" and "–"). In the case of three-class annotation, the third class is considered either

contradictory (referred to as "±") or neutral (referred to as "0"), and these two classes are not always separated in the corpora. There are also cases with a five-point rating (from 1 to 5), where the value "3" can mean either contradictory or neutral class.

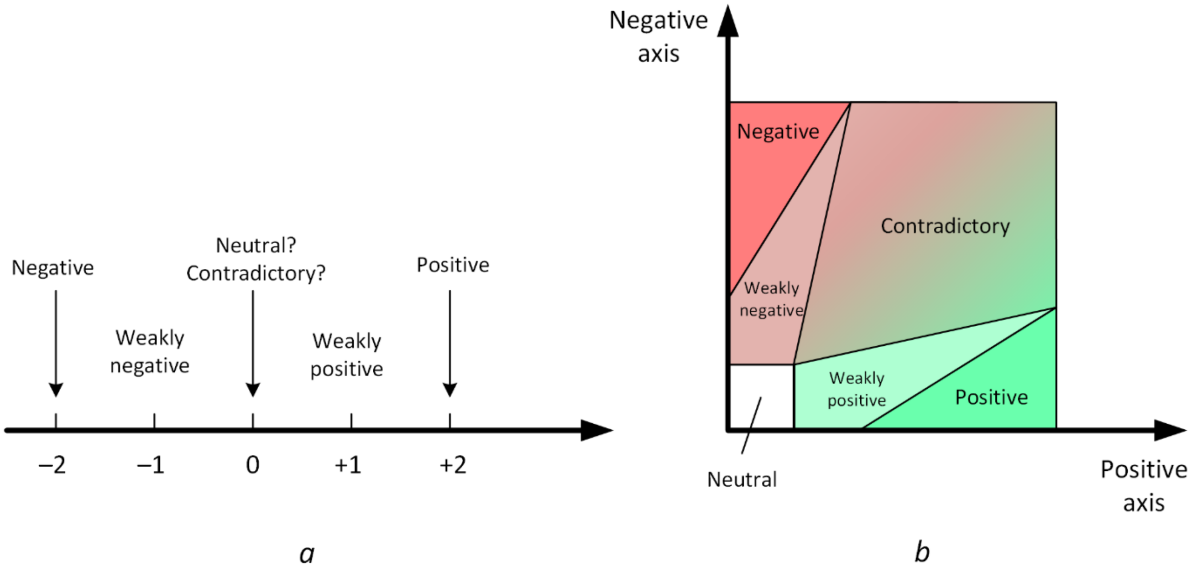


Figure 1: Sentiment spaces: a – 1D space, b – 2D space

The ratio of the sentiment classes. In many domains, there is a natural imbalance of texts according to the sentiment classes, which turns (if no special measures were taken) into corpora. For example, reviews about products, works of art and organizations are characterized by a significant predominance of positive texts, while tweets that mention banks and telecommunications companies tend to be more negative. It is known that class imbalance degrades the performance of classification [2], but there are methods that can level this aspect [23].

Annotation method. Existing corpora are annotated using three main approaches – manual annotation, automatic annotation, and use of the author's annotation. Manual annotation, in turn, is divided into expert annotation, when texts are marked up by a small number of qualified and motivated annotators, and crowdsourced annotation, in which a large number of crowdworkers are involved for annotation on a paid or free basis using specialized web platforms. Automatic annotation uses indirect sentiment information available in the text, for example, emoticons. In the third approach, the sentiment class of texts (usually reviews) is indicated in accordance with the score provided by the author of the text.

An important issue is the degree of confidence in the quality of the annotation. We propose to distinguish the following confidence levels, depending on the number of annotators and the approach to annotation:

1. High level (L1) – the annotation was carried out by at least two annotators, including on the basis of a crowdsourcing approach;
2. Middle level (L2) – the annotation was carried out by only one annotator;
3. Lower middle level (L3) – the annotation was based on the author's score;
4. Low level (L4) – the annotation was carried out automatically.

Train/test split. In many existing corpora there is a split into training and test parts – this is important for the reproducibility of experimental results.

2.2 Existing corpora

Table 1 shows the qualitative characteristics of existing corpora for sentiment analysis; Table 2 presents the quantitative characteristics of these corpora.

ROMIP-2011. For the competition of sentiment analysis systems within the ROMIP-2011 seminar, three corpora were created – reviews of books, movies and cameras [4]. Reviews were collected using queries on Yandex's Blog Search. Each corpus included two parts: a training part, which was marked up

using the author's score, and a test part, which was marked up by two annotators. The author's scores are given on a scale of [1..10] for reviews about books and movies and [1..5] for reviews about cameras. Test annotation was performed for three scales: binary (positive/negative), three-class (adding a contradictory sentiment) and five-class [1..5]. In the original paper [4], the results of systems for test reviews were evaluated according to the AND scheme (the system's score coincides with the score of both annotators) and OR scheme (the system score coincides with the score of at least one annotator). In this paper the AND scheme is used.

Corpus	**Source**	**Domain**	**Annotation**	**Confidence level**	**Number of classes: labels**
ROMIP-2011	Reviews	Book, movie and camera reviews	Train: author Test: 2 annotators	Train: L3 Test: L1	2: {+, –} 3: {+, –, ±} 5: {1, 2, 3, 4, 5}
ROMIP-2012 (reviews)	Reviews	Book, movie and camera reviews	1 annotator	Test: L2	2: {+, –} 3: {+, –, ±} 5: {1, 2, 3, 4, 5}
ROMIP-2012 (quotes)	News	–	1 annotator	Train: L2 Test: L2	Train, 4: {+, –, ±, 0} Test, 3: {+, –, 0}
SentiRuEval-2015 (reviews)	Reviews	Car and restaurant reviews	1 annotator (+checking)	Train: L2 Test: L2	4: {+, –, ±, 0}
SentiRuEval-2015 (tweets)	Twitter	Banks, telecom companies	3 annotators	Train: L1 Test: L1	Train, 4: {+, –, ±, 0} Test, 3: {+, –, 0}
SentiRuEval-2016	Twitter	Banks, telecom companies	Crowdsourcing	Test: L1	3: {+, –, 0}
SemEval-2016	Reviews	Restaurant reviews	2 annotators	L1	4: {+, –, ±, 0}
LinisCrowd	Social media posts	–	Crowdsourcing: 1 annotator, >1 annotator	1 annotator: L2 >1 annotator: L1	5: {–2, –1, 0, 1, 2}
Russian Hotel Reviews	Reviews	Hotel reviews	Author	Train: L3 Test: L3	5: {1, 2, 3, 4, 5}
RuSentiment	Social media posts	–	3 annotators	Train: L1 Test: L1	3: {+, –, 0}
RuSentRel	News	International politics	2 annotators	Train: L1 Test: L1	2: {+, –}
RuReviews	Reviews	Woman clothes and accessories reviews	Author	L3	3: {+, –, 0}
RuTweetCorp	Twitter	–	Automatic	L4	2: {+, –}
Kaggle Russian News Dataset	News	Kazakh news	?	?	3: {+, –, 0}
Twitter Sentiment for 15 European Languages	Twitter	–	1 annotator	L2	3: {+, –, 0}

Table 1: Qualitative characteristics of text corpora

Corpus	Domain	Number of texts	Training set	Test set	pos/neg/third class, %	Mean number of words (± Std Dev)
ROMIP-2011 (2 classes)	Books	19,946	19,680	266	89.6/10.4/0.0	49.3±101.6
	Movies	12,653	12,341	312	84.7/15.3/0.0	77.9±161.4
	Cameras	8,873	8,618	255	88.2/11.8/0.0	52.0±76.8
ROMIP-2012 (reviews, 2 classes)	Books	129	–	129	86.8/13.2/0.0	199.9±319.9
	Movies	408	–	408	80.9/19.1/0.0	298.5±545.0
	Cameras	411	–	411	96.6/3.4/0.0	57.0±70.6
ROMIP-2012 (quotes, 2 classes)	–	8,833	4,260	4,573	29.0/42.5/28.5	35.3±24.8
SentiRuEval-2015 (reviews)	Cars	403	203	200	52.9/13.9/33.3	116.6±68.2
	Restaurants	403	200	203	70.0/13.4/16.6	132.6±44.2
SentiRuEval-2015 (tweets)	Banks	9,417	4,883	4,534	7.4/18.2/74.4	9.6±4.9
	Telecom	8,613	4,839	3,774	14.5/28.2/57.3	12.2±5.5
SentiRuEval-2016	Banks	3,302	–	3,302	9.1/23.1/67.8	12.3±4.9
	Telecom	2,198	–	2,198	8.3/45.9/45.9	14.4±5.4
SemEval-2016	Restaurants	405	302	103	72.1/13.1/14.8	133.5±44.7
LinisCrowd (1 annotator)	–	28,853	–	–	7.7/42.5/49.8	148.6±103.6
LinisCrowd (>1 annotator)	–	10,566	–	–	6.9/40.4/52.7	139.8±75.9
Russian Hotel Reviews	Hotels	57,204	50,328	6,876	82.8/6.1/11.1	92.6±103.4
RuSentiment	–	26,745	24,124	2,621	37.8/14.6/47.6	12.6±16.9
RuReviews	Woman clothes and accessories	89,999	–	–	33.3/33.3/33.3	20.2±19.9
RuTweetCorp	–	226,834	–	–	50.7/49.3/0.0	12.2±4.9
Kaggle Russian News Dataset	Kazakhstan news	8,263	8,263	–	33.8/17.4/48.8	520.2±1192.2

Table 2: Quantitative characteristics of text corpora
("third class" – neutral and/or contradictory class; empty texts are excluded)

ROMIP-2012. The ROMIP-2012 competition was held in 2012 [5]. The corpora of ROMIP-2011 were used as training datasets. To obtain test data, new corpora of reviews about books, movies and cameras were marked up with a single annotator. The same scales were used for annotation as in ROMIP-2011 – 2-, 3- and 5-class. In addition to reviews, training and test corpora of quotes from news were also prepared for the competition. The scale {+, –, ±, 0} was used for annotation of the training corpus; there was no contradictory sentiment in the annotation of the test corpus.

SentiRuEval-2015. In 2015 the next sentiment analysis systems competition took place, which was aimed at two tasks: aspect-based sentiment analysis of reviews and object-oriented sentiment analysis of tweets [11]. For the first task, training and test corpora of car reviews were prepared, annotated by the aspects of Drivability, Reliability, Safety, Appearance, Comfort, Costs and General, and reviews about restaurants, for which the aspects of Food, Service, Interior, Price and General were highlighted. The annotation on the scale {+, –, ±, 0} was carried out by one annotator, but then a check was carried out. Table 2 provides data on the General aspect.

For the second task, training and test corpora of tweets about eight banks and seven telecommunications companies were annotated. The markup was done by three annotators; a voting scheme was used to obtain the final score. The scales {+, –, ±, 0} and {+, –, 0} were used for the training and test data respectively.

SentiRuEval-2016. The SentiRuEval-2016 competition also analyzed tweets in relation to banks and telecommunications companies [12]. The training corpora were built by combining the training and test

corpora of the SentiRuEval-2015. Crowdsourcing was used to annotate the test data on the {+, –, 0} scale.

SemEval-2016. In 2016 within the international competition SemEval-2016, the subtask of aspect-based sentiment analysis, including Russian-language reviews of restaurants, was singled out [15]. The training corpus was built on the basis of the corresponding SentiRuEval-2015 corpus and more than half overlaps with it. The test corpus was built from scratch. The annotation on the scale {+, –, ±, 0} was carried out by two annotators.

LinisCrowd. Within the Linis Crowd project, posts and comments of Top-2000 bloggers on LiveJournal were offered for crowdsourcing annotation [9]. The scale was {–2, –1, 0, 1, 2}. For each text a different number of scores were obtained – from 1 to 57. We divided the corpus into two parts: texts annotated by only one user (middle confidence level – L2) and texts annotated by several users (high confidence level – L1). In texts with several scores the final sentiment score was chosen according to the majority of scores.

Russian Hotel Reviews. Rybakov and Malafeev [18] offered a corpus of hotel reviews collected from tripadvisor.ru. The scores of reviews' authors were used. These scores correspond to a five-point scale for the aspects of Price-quality ratio, Location, Room, Cleanliness, Service, Quality of sleep and General. The corpus is divided into training and test parts.

RuSentiment. The corpus of posts on the social network VKontakte was presented in [16]. The corpus was marked by three annotators on the {+, –, 0} scale (the positive subcategory Speech Act was also highlighted, which we included in the positive class).

RuSentRel. Loukachevitch and Rusnachenko [13] presented a corpus of news articles on international politics from the inosmi.ru website. This corpus is annotated in relation to the named entities mentioned in the texts. The annotation was carried out on a {+, –} scale by two annotators; the third annotator resolved the contradictions. An overall sentiment score of the text was not made, therefore, information on the corpus is not provided in Table 2.

RuReviews. Smetanin and Komarov [19] presented a corpus of reviews about women's clothes and accessories, collected from some major e-commerce site. The original five-point scale was transformed by the authors into a three-point scale {+, –, 0}.

RuTweetCorp. Rubtsova [17] presented the largest corpus of Russian-language tweets for sentiment analysis to date. The annotation on the {+, –} scale was carried out automatically based on emoticons in the tweets.

Kaggle Russian News Dataset. The Kaggle[1] website presents a corpus of Kazakhstan news in Russian, annotated on a {+, –, 0} scale. The source of the texts and the method of annotation are unknown. The training and test parts of the corpus are available on the Kaggle website, but the sentiment scores are given only for the training part.

Twitter Sentiment for 15 European Languages. Mozetic et al. [14] consider tweet corpora for 15 European languages, including Russian. 93,321 messages were annotated with one annotator on the {+, –, 0} scale. However, the tweets themselves are not available (only their IDs are available), so Table 2 does not provide information about the corpus.

3 Materials and methods

3.1 BERT

To classify texts by sentiment, we use a deep neural network language model BERT (Bidirectional Encoder Representations from Transformers) [6], which showed the best results for the sentiment analysis in Russian [7, 20].

BERT is a Transformer encoder [22], which includes multiple layers; each layer contains a self-attention mechanism. Devlin et al. [6] presented two versions of BERT – $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$. In the base version the number of layers is 12, in the large version – 24. Work with BERT, as a rule, involves two stages. At the first stage, a language model is built by training on the tasks of predicting masked words and the next sentence using large text corpora (for example, Wikipedia). At the second stage, the pre-trained language model is fine-tuned to a specific task, for example, sentiment analysis. The

[1] https://www.kaggle.com/c/sentiment-analysis-in-russian.

impressive BERT results are based on the deep bi-directionality of the model, that is, considering left and right word contexts across all layers.

BERT uses subword tokenization to represent input texts [24]. The maximum input size for BERT is 512 tokens (subwords). In addition to word tokens, special tokens are used, for example, [CLS], which is always placed first and represents the text as a whole.

To classify texts in BERT a linear layer with a SoftMax function is used. The weights of this layer are randomly initialized before fine-tuning. This layer receives as the input the output vector corresponding to the special token [CLS].

In our work as a pre-trained language model we have used the Russian-language version of BERT – RuBERT, proposed by Kuratov and Arkhipov [10]. To train this model, a multilingual version of the $\text{BERT}_{\text{BASE}}$ was taken (12 layers, hidden size 768, feed-forward hidden size 3,072, and 12 self-attention heads). This version was retrained on the Russian-language Wikipedia and news corpus.

3.2 Corpora

One of the main goals of our work is to study the dependence of the performance of sentiment analysis on training data. To do this, we took 7 corpora of reviews as training datasets: three train parts of the ROMIP-2011 corpora (referred to as *R11_book_tr*, *R11_mov_tr* and *R11_cam_tr*), two train parts of the SentiRuEval-2015 corpora (*SRV15_car_tr* and *SRV15_rest_tr*), RuReviews corpus (*RuReviews*) and train part of the Russian Hotel Reviews (*Hotel_tr*). As test datasets we selected 9 review corpora: three test parts of the ROMIP-2011 corpora (*R11_book_te*, *R11_mov_te* and *R11_cam_te*), three test parts of the ROMIP-2012 corpora (*R12_book_te*, *R12_mov_te* and *R12_cam_te*), two test parts of the SentiRuEval-2015 corpora (*SRV15_car_te* and *SRV15_rest_te*) and test part of the Russian Hotel Reviews (*Hotel_te*).

Training corpora have the confidence levels L2 (2 corpora of the SentiRuEval-2015) and L3 (the remaining 5 corpora); test corpora – L1 (ROMIP-2011), L2 (ROMIP-2012, SentiRuEval-2015) and L3 (Russian Hotel Reviews).

Versions with binary scores (positive/negative) have been taken for all the corpora, that is, the task of two-class classification was solved. The characteristics for all of these corpora are shown in Tables 1 and 2.

The following preprocessing procedures were applied to the texts:

- URLs, e-mails and phone numbers were replaced with special tokens;
- characters that were repeated more than two times were replaced with a sequence of two such characters;
- joyful and sad emoticons were replaced with “joy” and “sadness” tokens.

The input size of the BERT model is limited to 512 tokens; the length of reviews often exceeds this size (see Table 2). To work with long reviews, the following strategy was used: half of the input tokens were taken from the beginning of the text, the other half – from the end of the text. This strategy is based on the fact that the main opinion in a review is often given either at the beginning or at the end.

4 Results and discussion

4.1 Experimental design

To select hyperparameters a preliminary series of experiments was carried out using only training corpora. As a result, the following hyperparameters values were selected, which were used in all subsequent experiments: the number of epochs is 5, the batch size is 8, and the learning rate is 2e–5.

The experiments were carried out using the Google Colab Pro platform, which provides graphics cards Tesla V100-SXM2-16GB or Tesla P100-PCIE-16GB.

Since the BERT training process is stochastic and depends on the random initialization of the weights of the output linear classification layer, three training runs were carried out for each experiment. As a result, we give the mean with the standard deviation.

The total training time (without preliminary experiments) was 80 hours. One run of the training process for the model with the largest amount of training data (all seven training corpora) was 7 hours 20 minutes.

The class imbalance of the review corpora (see Table 2) was the reason that the macro-averaged F1-score was used as the main performance metric, which equally took into account the metrics for all classes, regardless of the number of texts. In addition, this metric was used in other papers exploring these corpora.

To simulate the expansion of the training dataset, two series of experiments have been carried out.

In **the first series**, the increase in the number of corpora was as follows:

- at the first stage, two training corpora of the SentiRuEval-2015 (*SRV15_car_tr* and *SRV15_rest_tr*) were used as a combined training corpus;
- at the second stage, they were joined by RuReviews and train part of the Russian Hotel Reviews (*RuReviews* and *Hotel_tr*);
- at the third stage, three training corpora ROMIP-2011 (*R11_book_tr*, *R11_mov_tr* and *R11_cam_tr*) were added and as a result, all seven training corpora were used;

In **the second series**, the extension of the training dataset was carried out as follows:

- at the first stage, three training corpora of the ROMIP-2011 (*R11_book_tr*, *R11_mov_tr* and *R11_cam_tr*) were used as a combined training corpus;
- at the second stage, two training corpora of the SentiRuEval-2015 (*SRV15_car_tr* and *SRV15_rest_tr*) were added to them;
- at the third stage, they were joined by RuReviews and training part of the Russian Hotel Reviews (*RuReviews* and *Hotel_tr*) and as a result, all seven training corpora were used.

The third stage of both series is the same experiment (all seven training corpora).

At each stage of each series, the RuBERT model was fine-tuned with the above mentioned hyperparameters. The fine-tuned model was tested on all nine test corpora.

As a baseline, we used the results of fine-tuning of the RuBERT for a situation where the training dataset was the related training corpus for the test corpus. For example, for test corpora of book reviews ROMIP-2011 (*R11_book_te*) and ROMIP-2012 (*R12_book_te*) as the training dataset we used the training corpus of book reviews ROMIP-2011 (*R11_book_tr*), and for the test corpus of car reviews SentiRuEval-2015 (*SRV15_car_te*) the training dataset was the training corpus of car reviews SentiRuEval-2015 (*SRV15_car_tr*).

4.2 Results

The results of the experiments are shown in Tables 3 and 4 and in Figure 2.

Set of corpora	Number of corpora	R11_book_te	R11_mov_te	R11_cam_te	R12_book_te	R12_mov_te	R12_cam_te	SRV15_car_te	SRV15_rest_te	Hotel_te	Average
Baseline	1	.745 ±.013	**.762 ±.015**	.885 ±.030	.648 ±.031	**.698 ±.023**	**.672 ±.024**	.786 ±.050	.885 ±.021	.912 ±.005	.777 ±.092
SRV15_car_tr, SRV15_rest_tr	2	.525 ±.016	.577 ±.025	.728 ±.024	.602 ±.022	.622 ±.009	.516 ±.016	.867 ±.020	**.916 ±.023**	.773 ±.020	.681 ±.017
SRV15_car_tr, SRV15_rest_tr, RuReviews, Hotel_tr	4	.522 ±.018	.608 ±.026	.884 ±.025	.613 ±.023	.661 ±.010	.642 ±.007	**.887 ±.010**	.910 ±.012	**.915 ±.005**	.738 ±.005
All corpora	7	**.841 ±.020**	.756 ±.022	**.915 ±.011**	**.724 ±.034**	.684 ±.024	.665 ±.030	.869 ±.018	.890 ±.014	.914 ±.003	**.806 ±.011**

Table 3: First series of experiments, macro-averaged F_1-score (Mean ± Std Dev)

Set of corpora	Number of corpora	R11_book_te	R11_mov_te	R11_cam_te	R12_book_te	R12_mov_te	R12_cam_te	SRV15_car_te	SRV15_rest_te	Hotel_te	Average
Baseline	1	.745 ±.013	**.762 ±.015**	.885 ±.030	.648 ±.031	**.698 ±.023**	**.672 ±.024**	.786 ±.000	.885 ±.021	.912 ±.005	.777 ±.092
R11_book_tr, R11_mov_tr, R11_cam_tr	3	.804 ±.004	.755 ±.013	.885 ±.013	.608 ±.041	.693 ±.015	.653 ±.012	.787 ±.041	.759 ±.010	.810 ±.016	.751 ±.010
R11_book_tr, R11_mov_tr, R11_cam_tr, SRV15_car, SRV15_rest	5	.808 ±.048	.750 ±.011	.895 ±.021	.614 ±.044	.676 ±.020	.658 ±.018	.840 ±.008	**.890 ±.044**	.822 ±.009	.773 ±.007
All corpora	7	**.841 ±.020**	.756 ±.022	**.915 ±.011**	**.724 ±.034**	.684 ±.024	.665 ±.030	**.869 ±.018**	**.890 ±.014**	**.914 ±.003**	**.806 ±.011**

Table 4: Second series of experiments, macro-averaged F_1-score (Mean ± Std Dev)

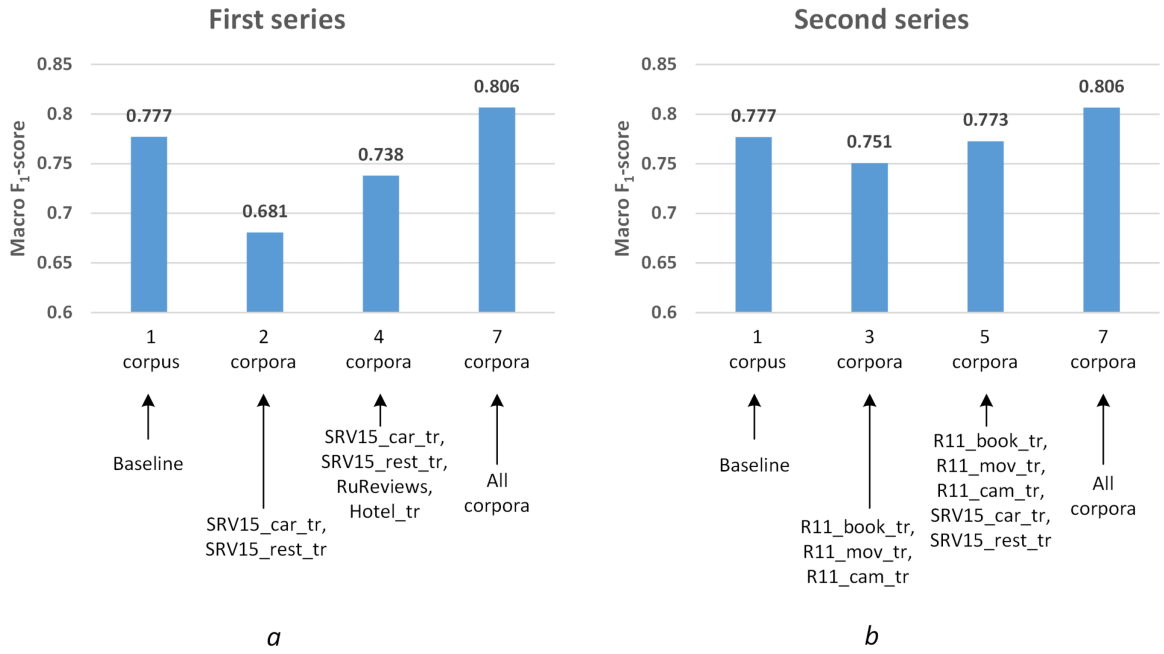


Figure 2: Macro-averaged F_1-scores for different numbers of training text corpora: *a* – first experimental series, *b* – second experimental series

4.3 Discussion

For three corpora out of nine (*R11_book_te*, *R11_cam_te* and *R12_book_te*), as well as on average (F_1-score=0.806), the model trained on all seven training corpora shows the best result. It should be noted that this result has been obtained by one model, in contrast to the result in the second place (baseline: F_1-score=0.777), obtained by averaging the results of different models.

Models trained on small SentiRuEval-2015 corpora (second row of Table 3) show low results, except for the related test corpora (*SRV15_car_te* and *SRV15_rest_te*) – training data is clearly not enough for high-quality training. Adding training corpora RuReviews and Russian Hotel Reviews (third row of Table 3) significantly increases the average F_1-score (from 0.681 to 0.738), despite the fact that the

domains of these two corpora do not correspond to the test corpora (with the exception of *Hotel_te*; but even excluding *Hotel_te* from consideration still gives an increase of the average F_1-score: from 0.681 to 0.716).

When training on all seven corpora (fourth row of Table 3) the performance scores for the SentiRuEval-2015 and Russian Hotel Reviews test corpora either do not decrease (for *SRV15_car_te* and *Hotel_te*), or decrease slightly (for *SRV15_rest_te*). Thus, the addition of ROMIP training corpora practically does not impair the learning process for these test corpora.

Models built on three ROMIP training corpora (second row of Table 4) for three of the six ROMIP test corpora (*R11_mov_te*, *R11_cam_te* and *R12_mov_te*) do not change the performance scores in comparison with training on the related corpora (baseline) (within 0.01), for two corpora reduce the performance scores (*R12_cam_te* – by 0.02 and *R12_book_te* – by 0.04) and for one corpus increase the performance score (*R11_book_te* – by 0.06).

The addition of SentiRuEval-2015 training corpora (third row of Table 4) significantly improves the performance for two SentiRuEval-2015 test corpora and practically does not change it for six ROMIP corpora.

Finally, the addition of RuReviews and Russian Hotel Reviews training corpora (fourth row of Table 4) significantly improves the performance for book, car and hotel corpora.

Figure 2 shows that in both series of experiments with an increase in the number of corpora (in the first series: 2 → 4 → 7; in the second series: 3 → 5 → 7) F_1-score on average increases. Thus, it can be concluded that expanding the training dataset has a positive effect on performance. In addition, the use of all available review corpora allows to obtain the best performance on average (F_1-score = 0.806). This circumstance allows us to look with cautious optimism at the possibility of building a universal neural network model for text sentiment analysis.

4.4 Comparison with previous results

Comparison of the obtained results with the results of other papers is possible only for the ROMIP-2011 and ROMIP-2012 test corpora. For SentiRuEval-2015, the performance scores are known for only three classes [11]; for Russian Hotel Reviews in [18] performance scores are given only for three aspects, but not for the review as a whole.

Table 5 shows the best results for the ROMIP-2011 test corpora from [4] and ROMIP-2012 from [5], as well as the results for these corpora obtained in our work: the results of the models trained on the related training corpora (our baseline – the first row in Tables 3 and 4) and the results of the model trained on all seven corpora (the last row in Tables 3 and 4). The best result for *R11_cam_te* in accordance with [4] was obtained using linear SVM; other two best models for the ROMIP-2011 test corpora were left unknown. The best result for *R12_book_te* in accordance with [5] was obtained by maximum entropy classifier, for *R12_mov_te* – rule-based classifier, and for *R12_cam_te* – linear SVM.

Neural network models for four corpora out of six have shown better results than the participants in the ROMIP competition. For the remaining two corpora, the results of neural network models differ from the previous results by less than 0.01. A significant advantage for the ROMIP-2011 book review corpus (0.841 vs 0.723) is due to the fact that the test corpus is highly imbalanced (244 positive reviews and 22 negative reviews – 91.7% and 8.3%), and the neural network model received a high macro precision (0.873 vs 0.698) due to accurate recognition of negative reviews ($precision_{pos}$=0.778).

Model	R11_book_te	R11_mov_te	R11_cam_te	R12_book_te	R12_mov_te	R12_cam_te
The best models from [4]	0.723	**0.770**	**0.921**	–	–	–
The best models from [5]	–	–	–	0.715	0.669	0.707
The model trained on related training corpus (baseline)	0.745	0.762	0.909	0.648	**0.698**	**0.723**
The model trained on all the corpora	**0.841**	0.756	0.915	**0.724**	0.684	0.665

Table 5: Results for test corpora of ROMIP-2011 and ROMIP-2012 (macro-averaged F_1-score)

5 Related work

Recently, several interesting papers have appeared in which the existing Russian-language corpora for the sentiment analysis have been investigated.

Zvonarev and Bilyi [26] used classifiers based on Logistic regression, XGBoost and Convolutional Neural Network for sentiment analysis of the RuTweetCorp. Baymurzina et al. [1] explored fastText and ELMo embeddings for sentiment analysis of the RuSentiment corpus.

At the end of 2018, the BERT neural network model [6] was presented based on the Transformer architecture [22], which showed State-of-the-Art results in several natural language processing tasks at once. After that, in several papers, the BERT model was applied for sentiment analysis in Russian.

Kuratov amd Arkhipov [10] built the RuBERT model – a Russian-language version of the BERT model based on the original multilingual version. RuBERT was used, inter alia, for sentiment analysis of the RuSentiment corpus.

Golubev and Loukachevitch [7] tested neural network models CNN, LSTM, BiLSTM and several variants of the BERT in the sentiment analysis task on the corpus of quotes ROMIP-2012[2], as well as on the SentiRuEval-2015 and SentiRuEval-2016 corpora.

Smetanin and Komarov [20] explored different versions of the BERT and Universal Sentence Encoder [3] for SentiRuEval-2015 (tweets), SentiRuEval-2016, RuTweetCorp, RuSentiment, Linis Crowd, Kaggle Russian News Dataset and RuReviews corpora.

In the above studies the state-of-the-art results for mentioned corpora were obtained on the basis of the BERT.

In our work, in contrast to those considered, we have investigated the effect of expanding the training dataset on the performance of sentiment analysis based on the BERT. In addition, in comparison with [20], our review includes the ROMIP-2011 and ROMIP-2012 corpora, and it was the first time that the performance scores of the BERT model have been obtained for them.

6 Conclusion

Currently, there are more than a dozen Russian-language text corpora, annotated by sentiment. These corpora differ significantly in sources, domains, sizes, quality of annotation and sentiment scales. Most of the corpora have a strong imbalance by classes, which reflects the distribution of texts in reality, but makes it difficult to train classifiers.

A variety of corpora can be used to build better models, which is confirmed by our experiments – the performance of the models increases on average with an increase in the number of training corpora. Also, information about the confidence level of the annotation quality can be used when choosing corpora for training and testing.

An important task is to study the possibility of constructing a universal sentiment analysis model that would find application in the fields where text analysis is required for many domains. In our work, it is shown that the performance is (obviously) strongly influenced by the presence of a corpus in a given domain. Less obvious was the fact that adding corpora in other domains, as a rule, either does not worsen the performance, or improves it.

Thus, the direction of further research is the possibility of building a universal model that is robust in relation to the domain.

References

[1] Baymurzina D., Kuznetsov D., Burtsev M. (2019), Language model embeddings improve sentiment analysis in Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2019», No. 18 (24), pp. 53–63.

[2] Buda M., Maki A., Mazurowski M.A. (2018), A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks, Vol. 106, pp. 249–259.

[3] Cer D., Yang Y., Kong S.-Y., Hua N., Limtiaco N, John R.S. et al. (2018), Computing Research Repository, arXiv:1803.11175.

[2] There is some uncertainty in the literature with the designation of the two ROMIP seminars held in 2011–2013: the first seminar is designated as ROMIP-2011 or ROMIP-2012, the second workshop – ROMIP-2012 or ROMIP-2013. This uncertainty stems from the fact that the first seminar was held in 2011, and the corresponding paper was published at the "Dialogue" conference in 2012. A similar situation took place with the second seminar.

[4] Chetviorkin I., Braslavskiy P., Loukachevitch N. (2012), Sentiment Analysis Track at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2012», No. 11 (18), Vol. 2., pp. 1–14.
[5] Chetviorkin I., Loukachevitch N. (2013), Sentiment Analysis Track at ROMIP 2012, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2013», No. 12 (19), Vol. 2, pp. 40–50.
[6] Devlin J., Chang M-W., Lee K., Toutanova K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 4171–4186.
[7] Golubev A., Loukachevitch N. (2020), Improving Results on Russian Sentiment Datasets, Proceedings of the Artificial Intelligence and Natural Language Conference (AINL-2020), pp. 109–121.
[8] Jiang H., He P., Chen W., Liu X., Gao J., Zhao T. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization (2020), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2177–2190.
[9] Koltsova O., Alexeeva S., Kolcov S. (2016), An opinion word lexicon and a training dataset for Russian sentiment analysis of social media, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2016», No. 15 (21), pp. 277–287.
[10] Kuratov Y., Arkhipov M. (2019), Adaptation of deep bidirectional multilingual transformers for Russian language, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2019», No. 18 (24), pp. 333–340.
[11] Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. (2015), SentiRuEval: Testing object-oriented sentiment analysis systems in Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2015», No. 14 (20), Vol. 2, pp. 3–13.
[12] Loukachevitch N., Rubtsova Y. (2016), SentiRuEval-2016: overcoming time gap and data sparsity in tweet sentiment analysis, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2016», No. 15 (21), pp. 416–426.
[13] Loukachevitch N., Rusnachenko N. (2018), Extracting Sentiment Attitudes from Analytical Texts, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2018», No. 17 (23), pp. 459–468.
[14] Mozetič I., Grčar M., Smailović J. (2016), Multilingual Twitter sentiment classification: The role of human annotators, PLoS ONE, Vol. 11, No. 5, e0155036.
[15] Pontiki M., Galanis D., Papageorgiou H., Androutsopoulos I., Manandhar S., AL-Smadi M. et al. (2016), SemEval-2016 Task 5: Aspect Based Sentiment Analysis, Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 19–30.
[16] Rogers A., Romanov A., Rumshisky A., Volkova S., Gronas M., Gribov A. (2018), RuSentiment: an enriched sentiment analysis dataset for social media in Russian, Proceedings of the 27th International Conference on Computational Linguistics, pp. 755–763.
[17] Rubtsova Y. (2014), Automatic Term Extraction for Sentiment Classification of Dynamically Updated Text Collections into Three Classes, Proceedings of the International Conference on Knowledge Engineering and the Semantic Web (KESW-2014), pp. 140–149.
[18] Rybakov V., Malafeev A. (2018), Aspect-based sentiment analysis of Russian hotel reviews, Proceedings of the 7th International Conference on Analysis of Images, Social Networks and Texts (AIST-2018), pp. 75–84.
[19] Smetanin S., Komarov M. (2019), Sentiment analysis of product reviews in Russian using convolutional neural networks, 2019 IEEE 21st conference on business informatics, Vol. 1, pp. 482–486.
[20] Smetanin S., Komarov M. (2021), Deep transfer learning baselines for sentiment analysis in Russian, Information Processing and Management, Vol. 58, 102484.
[21] Sun Z., Fan C., Han Q., Sun X., Meng Y., Wu F., Li J. (2020), Self-Explaining Structures Improve NLP Models, Computing Research Repository, arXiv:2012.01786.
[22] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N. et al. (2017), Attention is all you need, Advances in Neural Information Processing Systems, Vol. 30, pp. 5998–6008.
[23] Wei J., Zou K. (2019), EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6382–6388.
[24] Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., Macherey W. et al. (2016), Google's neural machine translation system: Bridging the gap between human and machine translation, Computing Research Repository, arXiv: 1609.08144.
[25] Yadav A., Vishwakarma D.K. (2020), Sentiment analysis using deep learning architectures: a review, Artificial Intelligence Review, Vol. 53, pp. 4335–4385.
[26] Zvonarev A., Bilyi A. (2019), A comparison of machine learning methods of sentiment analysis based on Russian language Twitter data, Proceedings of the 11th Majorov international conference on software engineering and computer systems. Saint Petersburg, Russia: ITMO University.

‘No way!’ Discourse formulae of disagreement in Russian and English: a comparative study

Evgenia Koziuk
HSE University
Moscow, Russia
zhenya.yuryevna@gmail.com

Yulia Badryzlova
HSE University
Moscow, Russia
yuliya.badryzlova@gmail.com

Abstract

The study explores the discourse formulae (DFs) of disagreement in Russian and English belonging to the subclasses of refusal and prohibition. Starting with a subset of six Russian target DFs, we establish their English equivalents using corpus analysis. We also define the typical speech acts to which the DFs in both languages react, and design model contexts that exemplify these types of speech acts. We use the model contexts as stimuli in our Russian and English surveys where we look at the preferences of native speakers in choice of DFs across the speech acts. We use the data of the surveys to establish the pragmatic function of each DF, (i.e. refusal or prohibition, or both), and their potential in each subclass (strong, medium, or weak). For each DF, we also identify the types of speech acts to which they react most readily. We compare the results of our analysis to the lexicographic description of the target DFs as presented in the Russian-English Dictionary of Idioms.

Keywords: discourse formulae, disagreement, refusal, prohibition
DOI: 10.28995/2075-7182-2021-20-445-455

«Ещё чего!» Дискурсивные формулы несогласия в русском и английском языках: сравнительное исследование

Козюк Е. Ю.
Национальный исследовательский университет
«Высшая школа экономики»
Москва, Россия
zhenya.yuryevna@gmail.com

Бадрызлова Ю. Г.
Национальный исследовательский университет
«Высшая школа экономики»
Москва, Россия
yuliya.badryzlova@gmail.com

Аннотация

В статье анализируются дискурсивные формулы (ДФ) несогласия в русском и английском языках, принадлежащие к подклассам отказа и запрета. Для шести русских ДФ мы устанавливаем их переводные эквиваленты в английском языке, используя анализ корпусных данных. Мы также выделяем типы речевых актов, которые являются наиболее характерными для рассматриваемых ДФ в обоих языках, и предлагаем набор модельных контекстов, иллюстрирующих каждый их этих типов. Полученные модельные контексты используются нами в качестве стимулов при опросе носителей русского и английского языка, в котором изучаются предпочтения в выборе той или иной ДФ в зависимости от типа речевого акта. На основе данных, собранных в ходе опроса, мы определяем прагматическую функцию каждой ДФ (т.е. принадлежность к подклассу отказа либо запрета, либо к обоим из них). Кроме того, полученные данные позволяют нам оценить потенциал ДФ в каждом из подклассов (высокий, средний или низкий). Мы также определяем, какие из типов речевых актов являются наиболее характерными для той или иной ДФ. Результаты исследования мы сопоставляем с лексикографическим описанием рассматриваемых ДФ в Русско-английском словаре идиоматических выражений.

Ключевые слова: дискурсивные формулы; несогласие; отказ; запрет

1 Introduction

The present work on discourse formulae is a spin-off of Russian Constructicon, a joint project by the National Research University (Moscow) and the Arctic University of Norway [1,2]. Russian Constructicon[1] is an online database of more than 2,000 Russian constructions accompanied by descriptions of their semantics, illustrative examples, and translations into English and Norwegian. According to Fillmore [3], constructions are primary units of language; they contain slots which are filled by variables, and carry syntactic, lexical, semantic, and pragmatic information. Constructions are non-compositional: the meaning of a construction is not equal to the aggregate of its components and their syntactic relations. However, when working on Russian Constructicon, its team observed constructions which are somewhat different in their nature: containing no empty slots, they constitute a completed utterance, and are used as separate sentences. We called them discourse formulae [4], e.g.:

(1) *V otpusk tak xočetsja! – I ne govori.*
'I so much wish I could go on vacation now! – You can say that again.'

Thus, the semantics of discourse formulae (DFs) is largely defined by their pragmatics; consequently, establishing cross-lingual equivalence of discourse formulae is a non-trivial task which requires application of corpus and survey methods.

The work on Russian DFs has branched off into a standalone project called Pragmaticon, currently in progress[2]. The database of the state-of-the-art Pragmaticon contains approximately 800 DFs; they are divided into seven classes: agreement (*ne vopros!* 'no problem!'); confirmation (*a kak že!* 'of course!'); disagreement, consisting of three subclasses: refusal (*ni za čto!* 'no way!'), prohibition (*ni v koem slučae!* 'out of the question!'), and negation (*kak by ne tak!* 'nothing of the kind!'); amazement (*ničego sebe!* 'holy cow!'); echo-questions (*èto kak?* 'how's that?'); devalorizaiton (*s kem ne byvaet!* 'things happen'), and indifference (*kak skažeš'* 'if you say so')[3]. Beside semantic descriptions, Pragmaticon also features information on accompanying gestures and intonation, and translation equivalents in English, Slovenian, German, and Chinese.

Discourse formulae should be differentiated from other functionally similar phenomena described below.

Discourse words (e.g. Russian *edva* 'hardly', *voobšče* 'in fact', *prjamo* 'really') are extensively represented in dictionaries. Baranov et al. [5] differentiate several classes of discourse words; each class is characterized by a common semantic component, cf. the discourse words of incompleteness (e.g. *edva*, *ele*, *čut'* 'barely', *s trudom* 'with an effort' *nemnogo* 'a little', and *počti* 'almost').

Discourse markers (*so*, *like*, *well*, *see*, etc.) [6] function as constituent parts of utterances. They contribute to cohesion of the discourse, serving as linking devices, e.g.:

(2) *Well, see, I guess what it is is the- um people get... t'be fifty eight, sixty, they move out of the houses, they move into an apartment.* [7].

Unlike discourse markers, discourse formulae constitute a completed utterance and can be used autonomously.

Communicatives [8] are used in dialogues as reactions to interlocutor's utterances for stereotypical expression of evaluations, opinions, and emotions (e.g. Russian *Net už* 'No way', *Kakoe tam!* 'Nothing of the kind!', *Obladet'*! 'Oh boy!', *Na zdorov'e!* 'You are welcome!'). While communicatives can be one- or multi-word units, discourse formulae are always comprised of more than one word.

Speech formulae are defined by Baranov and Dobrovol'skij [9] as "…idiomatic expressions of various structural types (predominantly completed utterances) either possessing fixed illocutionary power or defining the illocutionary characteristics of an utterance ". The class of discourse formulae is narrower than that of speech formulae: the former are always used in response to a stimulus utterance (a certain speech act); therefore, DFs are identified in the context of a specific speech act to which they react.

Our study focuses on two subclasses of DFs of disagreement – refusal and prohibition. We examine a subset of Russian formulae of refusal and prohibition which are most representative of these subclasses. We conduct a corpus study in order to establish the English equivalents of the Russian DFs, as well as the types of speech acts that are most typical for the target DFs in both languages. We define

[1] The new website of Russian Constructicon is currently under development, available at https://constructicon.github.io/russian/. Should the location change, notification will be published on the project's website (https://site.uit.no/russian-constructicon/).

[2] Pragmaticon will soon be available for access at https://pragmaticon.ruscorpora.ru/

[3] Since Pragmaticon is an ongoing project, the number and type of classes is subject to change.

model contexts for each type of speech act and design Russian and English questionnaires in order to collect data on preferences of native speakers when choosing this or that DF for a particular speech act. We use the collected data to identify the pragmatic function of each Russian and English DF, i.e. to define whether they belong to the subclass of refusal or prohibition (or both). The survey data also allows us to assess the potential of each DF in its subclass (strong, medium, or weak), and the strength of association between each DF and specific types of speech acts. Finally, we compare our results with the descriptions of the target DFs as presented in the Russian-English Dictionary of Idioms [10].

2 Present study

DFs of refusal and prohibition serve to express directive negation; their main function is to preclude some future situation from happening. Specifically, the formulae of refusal are used when the speaker refuses to commit a certain action; as for the formulae of prohibition, it is the speaker who issues the prohibition on the interlocutor to commit a certain action. In both cases, the result is that a realistic future event becomes unreal due to the will of the speaker.

Russian Pragmaticon currently contains about 250 DFs of disagreement (in the three subclasses – refusal, prohibition, and negation). For our study we have selected the following six DFs of refusal and prohibition: *eščë čego*, *vot eščë*, *ni za čto*, *ni v koem slučae*, *i reči byt' ne možet*, and *tol'ko ne èto*. The selection was made on the basis of frequency (i.e. DFs with the highest frequencies in corpora), lexical diversity (i.e. DFs expressed by diverse lexical means), and functional diversity (i.e. DFs reacting to a wide range of speech acts).

2.1 Corpus analysis

The English equivalents of the Russian target DFs were established using the parallel corpus Context Reverso[4] and the online dictionary Multitran[5] . The candidate equivalents were verified by examining their occurrences in the Movie subcorpus of COCA (the Corpus of Contemporary American English)[6] and consulting with native experts; additionally, the candidates were filtered by their frequency in the COCA. In the selection process we aimed at diversity, so as to avoid resembling formulae (cf. *not for anything (in the world/on earth)* and *not for (all) the world*); we also excluded DFs affiliated with the formal register and thus requiring specific contexts (e.g. *by no means*, *on no account*, *nothing of the kind*). The preliminary list of English equivalents and their distribution across the Russian formulae are shown in Table 1; at the subsequent stages of the present study this list will be put to test using corpus and survey data.

	not a chance	**out of the question**	**no way**	**not on your life**	**under no circum-stances**	**anything but that**	**you wish**	**not again**
eščë čego	+	+	+				+	
vot eščë	+							
ni za čto	+		+	+				
ni v koem slučae	+	+	+		+			
i reči byt' ne možet		+	+					
tol'ko ne èto						+		+

Table 1: Preliminary distribution of English equivalents across Russian DFs.

[4] https://context.reverso.net/translation/
[5] https://www.multitran.com/
[6] https://www.english-corpora.org/coca/

In order to establish the classes of speech acts in which the Russian target formulae typically occur, we analyzed their occurrences in the Russian National Corpus (RNC)[7] and defined the respective typical speech acts [see 11]. The English contexts were either translated from Russian, or selected from COCA. Nine classes of speech acts were identified (the gaps indicate the position of the DF):

(1) Offering help:
(Russian) *Provodit' vas, tovarišč general? – _____. Mogu dvigat'sja bez postoronnej pomošči.*
'Shall I accompany you, comrade General? – _____.I am capable of walking by myself.'
(English) *Let me help you – _____. I'm not a little girl. I can do it myself.*

(2) Command:
(Russian) *Poguljaj s sobakoj. – _____. Počemu vsegda ja?*
'Go walk the dog. – _____. Why is it always me?'
(English) *Help your sister do her homework. – _____. I have my own homework to do.*

(3) Advice:
(Russian) *Uxodi v otstavku, uezžaj kuda-nibud'. – _____. Ty sam ponimaeš', čto èto bylo by begstvo.*
'Resign from office, leave the town!' – _____. You know perfectly well this would mean retreat.'
(English) *You should visit the Tate Modern when you're in London. – _____. I hate modern art.*

(4) Asking for advice:
(Russian) *Možet, mne ne stoit publikovat' knigu? – _____. Vse uže rešeno.*
'The book may not be worth publishing, may it? – _____. It is a settled matter now.'
(English) *Should we invite Steve to the party? – _____. He was rude to me last week.*

(5) Suggesting joint activity:
(Russian) *Davaj letom za granicu poedem? – _____. Tol'ko na daču!*
'Let's go abroad this summer?' – _____. Only our country house!'
(English) *Let's fly business class. – _____ . I'm not wasting my money.*

(6) Request:
(Russian) *Požalujsta, otdaj mne svoi starye krossovki. – _____. Oni mne samomu nužny!*
'Will you let me have your old sneakers? – _____. I need them myself!'
(English) *Can you recite your poetry? – _____. I am a bad poet. I don't recite my poems to anyone.*

(7) Asking for permission:
(Russian) *Možno ja druga provožu? – _____. On ne malen'kij, sam dorogu znaet.*
'May I see my friend off? – _____. He is not a little boy, he knows the way.'
(English) *Can I see my friend off? – _____. He is not a little boy. He knows the way.*

(8) Commissive:
(Russian) *Vot čto značit zimoj bez šapki xodit'. Ja sejčas že vzyvaju vrača! – _____. Ja prosto vyp'ju čaja s limonom, i vse projdet.*
'This is what you get for walking around bareheaded in winter! I am calling the doctor immediately! – _____. I will just drink tea with lemon, and it all will go.'
(English) *Thanks for the meal. I'll pay. – _____. I invited you!*

(9) Question on future intension:
(Russian) *Ty dumaeš' stixi pisat'? – _____. Ja ubedilsja, čto u menja ničego ne vyxodit.*
'Are you thinking of writing poetry? – _____. I have tried and seen I am none of a poet.'
(English) *Are you planning to become a lawyer? – _____. I'm studying international relations.*

The nine types of target speech acts are divided between the two subclasses (refusal and prohibition) as follows:

- **Refusal**: command, request, suggesting joint activity, advice, and question on future intension.
- **Prohibition**: asking for permission, asking for advice, offering help, and commissive.

[7] https://ruscorpora.ru/new/

We expect that some of the target formulae will gravitate towards one of the subclasses, while others may be used in both of them. Noticeably, as seen in Table 1, the English *no way* and *not a chance* correspond to the majority of the Russian DFs, and therefore are expected to be dominant, i.e. to fit most of the stimulus contexts (1-9).

2.2 The survey

The contexts of the prototypical speech acts (1-9 above) were compiled into two questionnaires (Russian and English, respectively), which were offered to respondents[8]. The respondents could choose any number of DFs to fill in the gaps, without ranking them. The questionnaires were administered via Google Forms. The Russian questionnaire was filled by 34 native speakers of Russian, age 18-55, women 70.5%. The respondents of the English survey were 40 native speakers of English, age 20-66, women 56.1%. In regard to the variety of English spoken, 61% identified as speakers of American, 17.1% of British, 19.5% of Canadian, and 2.4% of Australian English. All respondents in both samples confirmed that they are not professional linguists; every respondent gave their informed consent to participate in the experiment.

3 Results

The results of the survey are visualized in Figures 1-3. The heatmap in Figure 1 shows how the DFs are distributed across the speech acts, and how the speech acts align with the two subclasses (refusal and prohibition); thus, Figure 1 highlights the dominant speech acts for each DF. Figure 2 contains averaged numbers across the speech acts in each of the two subclasses; it allows us to judge to which of the subclasses (refusal or prohibition) each DF belongs. Figure 3 demonstrates the results of correspondence analysis[9] [12–14] between DFs and speech acts; it allows up to make judgements about the centrality of each formula in the class of disagreement: the nearer to the center, the more dominant the DF is; the farther off the center, the more peripheral it is. (For convenience of presentation, we do not show the speech acts in this plot.)

3.1 Analysis of Russian DFs

The formulae *eščë čego* and *vot eščë* belong to both subclasses – prohibition and refusal, and react to the majority of speech acts. Their dominant speech acts are: command, request, suggesting joint activity, asking for permission, and offering help. Besides, *vot eščë* can be used in the speech act of devalorizaiton (which is not discussed in this paper), i.e. for negative evaluation of situation.

The DFs *ni v koem slučae* and *i reči byt' ne možet* belong to the subclass of prohibition. The dominant speech acts of *ni v koem slučae* are advice and commissive; slightly less dominant is the act of asking for help. The dominant speech acts of *i reči byt' ne možet* are advice and asking for advice, while asking for permission is also dominant, but to a lesser degree. When reacting to advice (i.e. expressing refusal), the refusal expressed by these two DFs is somewhat different from what happens when the speaker uses the two dominant formulae of refusal – *eščë čego* and *vot eščë*. In the latter case, the speaker expresses reluctance to follow the advice, whereas in the former the speaker informs the interlocutor about the impossibility of following the advice, cf. (10) and (11):

(10) *Da ty by na nego načal'stvu požalovalas'! – Vot ešče / ešče čego, iz-za takix pustjakov načal'stvo bespokoit'.*
'You should complain to the authorities about him! – No way! Bothering the authorities about such trifles!

(11) *Da ty by na nego načal'stvu požalovalas'! –Ni v koem slučae / i reči byt' ne možet, iz-za takix pustjakov načal'stvo bespokoit'.*
'You should complain to the authorities about him! – That's out of the question! Bothering the authorities about such trifles!

The DF *ni za čto* belongs to the subclass of refusal. Its dominant speech acts are suggesting joint activity, request, and advice. However, it can also act as prohibition in reaction to offer of help and commissive.

[8] The full version of the questionnaires is available at https://docs.google.com/document/d/1_ZQIOuyfjIvexgzRM-n_0fcLg5V0g1yDRk-z3vUHIjA/edit?usp=sharing
[9] Prince, the Python factor analysis library: https://pypi.org/project/prince/

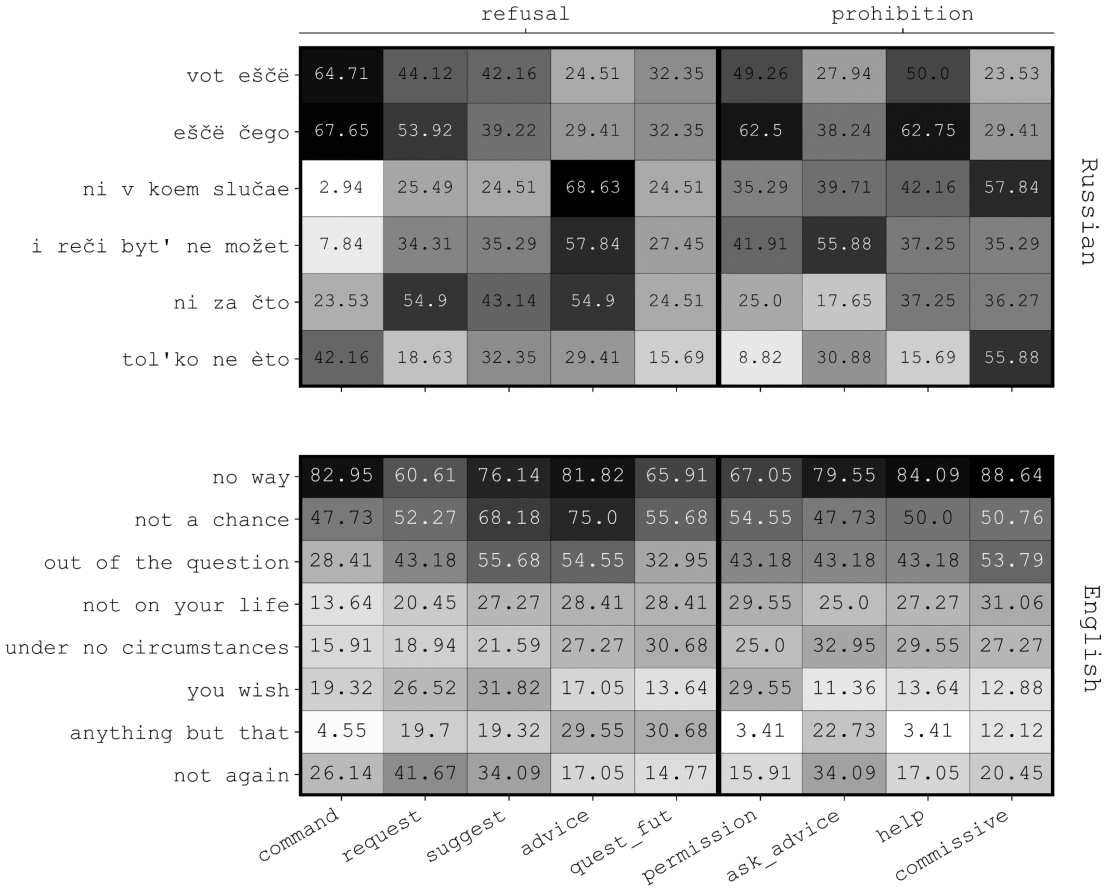


Figure 1: Distribution of DFs across speech acts, % (`suggest` = suggesting joint activity, `quest_fut` = question on future intention, `permission` = asking for permission, `ask_advice` = asking for advice, `help` = offering help).

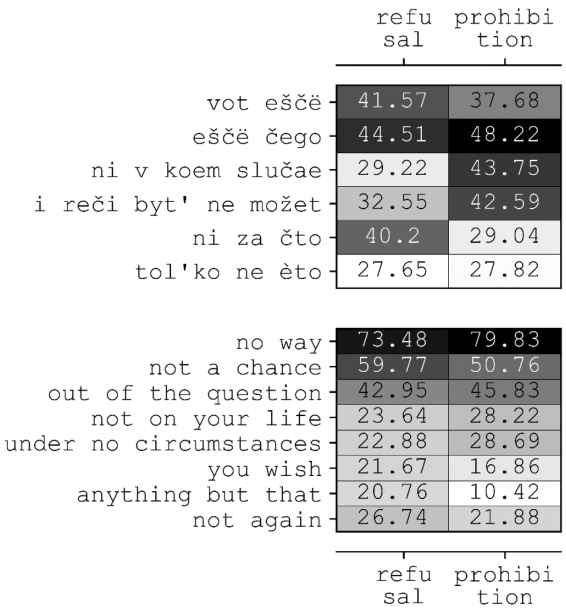


Figure 2: Distribution of DFs between the subclasses of refusal and prohibition, %.

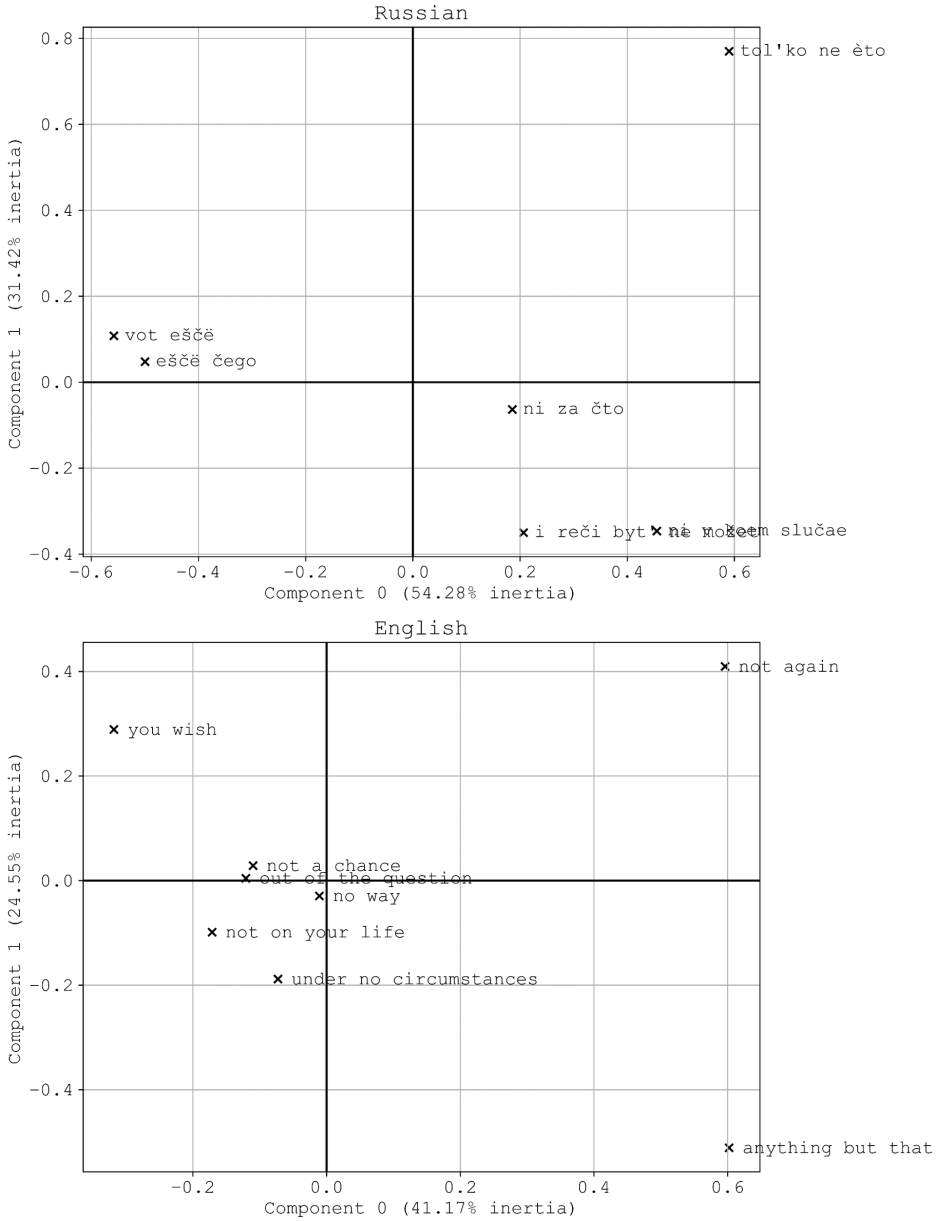


Figure 3: Correspondence analysis between DFs and speech acts.

Yet, as compared to *ni v koem slučae* and *i reči byt' ne možet*, *ni za čto* is chosen less frequently for expression of prohibition.

The formula *tol'ko ne èto* is peripheral in the class of disagreement.

3.2 Analysis of English DFs

No way and *not a chance* are the dominant DFs in both classes, refusal and prohibition. Still, *no way* is relatively more dominant than its counterpart: *no way* is strongly preferred in all the contexts of refusal and prohibition, whereas *not a chance* is comparatively weakly preferred in the contexts of command and asking for advice. Besides, the refusal and prohibition conveyed by *no way* are more emotionally charged.

Out of the question also belongs to both of the subclasses, refusal and prohibition, although is a less dominant formula than *no way* and *not a chance*. Its most dominant speech acts are advice, commissive, and suggesting joint activity.

Not on your life and *under no circumstances* show moderate reaction to the speech acts of prohibition, and somewhat weaker reaction to the speech acts of refusal – in particular, to command and request. The

speech act of refusal to which they react most readily is in response to question on future intensions, e.g.:

(12) *Would you share a room? – Not on your life / under no circumstances. I really don't like her personality.*

*You wis*h contains a negative evaluative component; it can express both refusal (in the speech acts of command, request, and suggesting joint activity) and prohibition (when asking for permission). Yet, its major function is that of refusal.

Anything but that and *not again* are peripheral DFs of disagreement. Additional corpus analysis of these formulae showed that *not again* is preferred in situations when the interlocutor repeatedly addresses the speaker with a suggestion or resumes previous conversation. If the participants of the dialogue are speaking about a new topic or subject, *anything but that* is preferred.

3.3 Exploring pragmatic potential of the DFs

The observations on the Russian and English DFs presented in Sections 3.1 and 3.2 can be summarized as shown in Table 2. The most distinct group of DFs is formed by the Russian *vot eščë* and *eščë čego* and the English *no way* and *not a chance*: they all are dominant, manifesting strong association with both refusal and prohibition. (Having said that, it should be noted that the dominance of the Russian DFs is much less pronounced than that of their English counterparts.)

The next group is presented by DFs with strong-medium association with the subclass of prohibition, and medium-weak association with the subclass of refusal: the Russian *ni v koem slučae* and *i reči byt' ne možet*, and the English *out of the question*, *not on your life*, and *under no circumstances*.

The pair of DFs *ni za čto* (Russian) and *you wish* (English) comprises the group distinguished by weak association with the subclass of prohibition and strong-medium association with the subclass of refusal.

The Russian *tol'ko ne èto,* along with the English *anything but that* and *not again*, are peripheral DFs in the class of disagreement.

Language	DF	Refusal	Prohibition	Periphery
Rus	**vot eščë**	strong	strong	
Rus	**eščë čego**	strong	strong	
Eng	**no way**	strong	strong	
Eng	**not a chance**	strong	strong	
Rus	**ni v koem slučae**	weak	strong	
Rus	**i reči byt' ne možet**	weak	medium	
Eng	**out of the question**	medium	medium	
Eng	**not on your life**	weak	medium	
Eng	**under no circumstances**	weak	medium	
Rus	**ni za čto**	strong	weak	
Eng	**you wish**	medium	weak	
Rus	**tol'ko ne èto**			strong
Eng	**anything but that**			strong
Eng	**not again**			strong

Table 2: Profiles of Russian and English DFs.

3.4 Russian-English equivalence finalized

The translation equivalence between the Russian and the English DFs which has been confirmed in the course of our study can be summed up as follows (Table 3):

Vot eščë and *eščë čego*, when used in the function of refusal, carry a connotation of negative evaluation; the same applies to their common English equivalent *not a chance* (when expressing refusal) and to the English equivalent of *eščë čego*, the DF *you wish*. To express prohibition, both of these Russian DFs can be translated by the dominant English DF *no way*.

	not a chance	out of the question	no way	not on your life	under no circum-stances	anything but that	you wish	not again
eščë čego	+	+	+				+	
vot eščë	+		+					
ni za čto	+		+	+				
ni v koem slučae	+	+	+		+			
i reči byt' ne možet	+	+	+					
tol'ko ne èto						+		+

Table 3: Revised distribution of English equivalents across Russian DFs. Cells with darker grey background correspond to confirmed equivalents; slate-grey cells correspond to confirmed equivalents with limited functionality.

In the case of expressing refusal, *ni za čto* is an emotionally loaded DF. Its closest English equivalent is the dominant formula *no way* (particularly in commands and suggestions about joint activity, where it expresses categorical refusal). *Ni za čto* can also be translated by means of the other dominant English DF, *not a chance* – mainly in questions about the speaker's intentions and in suggestions about joint activity. Equally well, *ni za čto* can be rendered by *not on your life* – primarily in response to commands and suggestions about joint activity, where both DFs, as a rule, express refusal to carry out the action on the speaker's own will rather than due to impossibility.

Ni v koem slučae can be translated as *no way* both in the function of refusal and prohibition; in certain cases it can also be translated by *out of the question* (for refusal to follow advice and for commissive prohibition) and by *under no circumstances* (in response to request for advice and as commissive).

The most preferable English equivalent of *i reči byt' ne možet* – both in the function of refusal and prohibition – appears to be *no way*, yet the Russian DF can also be translated by *out of the question*.

The peripheral Russian DF of refusal *tol'ko ne èto* is typically used for negative evaluation; its English equivalents are *anything but that* (for negative assessment and refusal) and *not again* (when the speaker is reluctant to repeatedly commit an action or to be told once again something they have already heard).

3.5 Comparison with lexicographic description

In order to assess the reliability of our results we chose to compare them to the descriptions of the target DFs presented in the Russian-English Dictionary of Idioms (REDI, hereafter) [10]. This dictionary was chosen as one of the most comprehensive and academically acclaimed Russian-English lexicographic sources; it is based on parallel translations of Russian fiction literature, thus providing common ground for the comparison. Table 4 shows how the findings of the present study about the equivalence between the subsets of the six Russian and eight English DFs align with their description in REDI: column A contains the equivalents stated in this study but missing in REDI; column B presents the equivalents that are indicated in REDI but have not been confirmed in this study; and column C lists the equivalents where both sources agree.

The DF *vot eščë* is described in REDI as belonging only to the subclass of refusal (whereas our results suggest that it belongs to both subclasses); besides, the other English equivalents of *vot eščë* in the dictionary are characterized by strong expressivity and emotionality, cf.: *what (on earth) are you talking about!*, *you've got to be kidding!*, *you must be out of your mind*. As for the equivalent suggested by our research – *not a chance* – it is not present in REDI. We think that, as expressive a DF as *vot eščë* is, it can also appear in contexts without distinct emotional expression; in such cases *not a chance* would be the most apt translational equivalent.

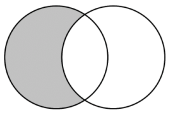
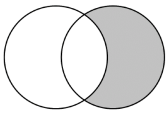
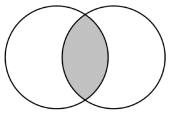
DF	A. This study REDI 	B. This study REDI 	C. This study REDI 
vot eščë	not a chance	∅	∅
eščë čego	not a chance you wish out of the question	not on your life	no way
ni za čto	not a chance	∅	not on your life no way
tol'ko ne èto	anything but that not again	[unavailable]	∅
ni v koem slučae	not a chance out of the question	∅	no way under no circumstances
i reči byt' ne možet	no way	∅	out of the question

Table 4: Russian-English equivalents, comparison between this study and REDI (∅ denotes an empty set, i.e. absence of English equivalents in either of the sources or in their intersection, respectively).

The DF *eščë čego* is referred by REDI to the subclass of refusal, with the most characteristic speech acts being suggestion about joint activity and request; this partially agrees with our data, where this DF is dominant in both subclasses, refusal (most prominently, in commands) and prohibition (particularly, in the speech acts of asking for permission and offering help). Among the English equivalents of *eščë čego* REDI lists two of our DFs – *no way* and *not on your life*. While we admit that *no way*, being a dominant DF, may also be capable of corresponding to *eščë čego* in translation, it is not the case with *not on your life*. Unlike its Russian counterpart, *eščë čego*, the English DF *not on your life* can be used in speech acts of prohibition only to a limited extent (see Section 3.3); moreover, when reacting to commands, *not on your life* should be translated with the Russian DF *ni za čto*. REDI does not list *not a chance* and *you wish* as English equivalents of the Russian *eščë čego*, whereas our study shows that there is equivalency between them, as all of these DFs contain a negative evaluative component.

The Russian DF *ni za čto* is defined in REDI as 'on no condition, under no circumstances', featuring two English equivalents from our list of DFs – *not on your life* and *no way*; however, REDI makes no mention of *not a chance*, which corresponds to *ni za čto* in the contexts of refusal to engage in joint activity.

The Russian DF *tol'ko ne èto* is not represented in REDI whatsoever; the dictionary contains similar DFs *tol'ko ètogo ne xvatalo* and *ešče čego ne xvatalo* and suggests the following DFs as their English equivalents: *that's the limit, that's the last straw*, etc. – that is, DFs with a strong component of indignation in their semantics. REDI also points out that these English DFs are used to express categorical and impolite refusal to accept the interlocutor's suggestion – which is dramatically different from the characteristics of the Russian DF *tol'ko ne èto*.

The Russian DF *ni v koem slučae* is primarily described by REDI as a construction occurring in non-dialogue utterances, cf.:

(13) *Ja ni v koem ne dopuskaju mysli, čto…*
'By no means do I admit the idea that…'

According to REDI, the English equivalents of *ni v koem slučae* are the DFs *not for one moment*, *there's no way*, and others; yet the dictionary also lists two of the English DFs suggested by our study: *no way* and *under no circumstances*. REDI does not mention the other two of our suggested equivalents – *out of the question* and *not a chance*, which, similarly to *ni v koem slučae*, serve to express prohibition.

The Russian DF *i reči byt' ne možet* is placed by REDI in the class of categorical refusal and rejection, whereas, according to our data, this DF can also be associated with the subclass of prohibition. Similarly to our study, REDI lists the DF *out of the question* as an English equivalent of DF *i reči byt' ne možet*, although treating it as a non-dialogue utterance of the type *X is out of the question*; besides, REDI does

not mention *no way*, which, according to our analysis, corresponds to this Russian formula in speech acts of categorical refusal.

4 Conclusions

Being pragmatic units, discourse formulae pose a difficult problem in translation. Defining cross-lingual equivalents of DFs requires multi-faceted analysis involving dictionaries as well as data from corpora and surveys. Using corpus analysis, we identified the English equivalents of the Russian DFs of disagreement. We also conducted a survey in order to establish to which of the two subclasses of disagreement – refusal or prohibition – the target Russian and English DFs belong. The results indicate that the English formulae have a broader coverage – most of them belong to both of the subclasses, whereas the Russian formulae tend to be more specialized in their affiliation with either subclass; yet, they can occasionally react to speech acts from the opposite subclass. In both sets of DFs, we identify peripheral formulae with marginal frequency and coverage. Besides, we demonstrate that the choice of DFs can be affected by finer pragmatic nuances of the context. The results of the study will be incorporated into Pragmaticon, the database of Russian discourse formulae. The approach suggested in this paper may contribute to advancement of the practices of cross-lingual lexicographic description of DFs.

Acknowledgements

This work is supported by the Ministry of Science and Higher Education of the Russian Federation (075-15-2020-793).

References

[1] Janda L.A. et al. A Constructicon for Russian: Filling in the Gaps. // Constructicography: Constructicon development across languages. 2018.
[2] Endersen A. et al. The Russian Constructicon: a new linguistic resource, its design and key characteristics // Computational Linguistics and Intellectual Technologies. 2020.
[3] Fillmore C.J., Kay P., O'Connor M.C. Regularity and idiomaticity in grammatical constructions: The case of let alone // Language. 1988. P. 501–538.
[4] Puzhaeva S. et al. Automated extraction of discourse formulas from Russian texts [Avtomatičeskoe izvlečenie diskursivnyx formul iz tekstov na russkom jazyke] // Bulletin of Novosibirsk State University. Series: Linguistics and Intercultural Communications. 2018. Vol. 16, № 2.
[5] Baranov A., Plungian V., Rakhilina E. A guide to Russian discourse words [Putevoditel' po diskursivnym slovam russkogo jazyka]. Moscow: Pomovsky & Partners, 1993.
[6] Schiffrin D. Discourse Markers. Cambridge: Cambridge University Press, 1987.
[7] Schiffrin D. Discourse markers: Semantic resource for the construction of conversation: Unpublished doctoral dissertation. University of Pennsylvania, 1982.
[8] Sharonov I. Discourse words and communicatives [Diskursivnye slova i kommunikativy] // Computational Linguistics and Intellectual Technologies. 2016. № 15. P. 22.
[9] Explanatory dictionary of Russian phraseology [Frazeologičeskij ob"jasnitel'nyj slovar' russkogo jazyka] / ed. Baranov A., Dobrovol'skij D. Litres, 2017.
[10] Lubensky S. Russian-English dictionary of idioms. Yale University Press, 2014.
[11] Rakhilina E., Bychkova P., Zhukova S. Rečevye akty kak lingvističeskaja kategorija. Diskursivnye formuly [Speech acts as a linguistic category. Discourse formulae]. // Voprosy jazykoznanija [Topics in the study of language]. 2021. Vol. 2.
[12] Hirschfeld H.O. A connection between correlation and contingency // Mathematical Proceedings of the Cambridge Philosophical Society. Cambridge University Press, 1935. Vol. 31, № 4. P. 520–524.
[13] Fisher R.A. The precision of discriminant functions // Annals of Eugenics. Wiley Online Library, 1940. Vol. 10, № 1. P. 422–429.
[14] Greenacre M. Correspondence analysis in practice. Chapman and Hall/CRC, 2007.

The types of infinitive constructions with predicatives (according to the Russian National Corpus)

Kustova G. I.
Vinogradov Russian Language Institute
of the Russian Academy of Sciences;
Moscow, Russia
galinak03@gmail.com

Abstract

The paper considers constructions «predicative + infinitive». For the first time, a class of interpretive infinitive constructions (opposed to emotional reactions) is introduced. For emotional reactions, the predicative and the infinitive refer to the same subject, the infinitives of the perception, mental, speech verbs are typical for them: *It hurts / scares to see how forests are dying* ('X sees, X is scared') → It hurts that forests are dying. For interpretive constructions, the subjects of the predicative and the infinitive do not coincide: *It is heartless to separate the mother from the children* – 'X separates, Y evaluates such an act as heartless'. The infinitives of perceptual and mental verbs in such a construction are either not used, or they denote a kind of action: *It is tactless to listen to private conversations*.

Keywords: predicate; infinitive construction; interpretation predicates
DOI: 10.28995/2075-7182-2021-20-456-463

Типы инфинитивных конструкций с предикативами (по данным Национального корпуса русского языка)

Кустова Г. И.
Институт русского языка
им. В. В. Виноградова РАН;
Москва, Россия
galinak03@gmail.com

Аннотация

В работе рассматриваются конструкции «предикатив + инфинитив». Впервые вводится класс интерпретационных инфинитивных конструкций, которые противопоставляются эмоциональным реакциям. У эмоциональных реакций предикатив и инфинитив относятся к одному и тому же субъекту, для них типичны инфинитивы глаголов восприятия, ментальные, речевые: ***Больно*** / ***страшно*** *видеть, как гибнут леса* ('X видит, X-у больно / страшно') → *Больно, что гибнут леса*. У интерпретационных конструкций субъекты предикатива и инфинитива не совпадают: ***Бессердечно*** *разлучать мать с детьми* – X разлучает, Y оценивает такой поступок как бессердечный. Инфинитивы перцептивных и ментальных глаголов в такой конструкции либо не употребляются, либо обозначают своего рода поступок: *Бестактно* ***слушать*** *частные разговоры*.

Ключевые слова: предикатив; инфинитивная конструкция; предикаты интерпретации

1 Введение

С предикативом, или категорией состояния, ср. [Shcherba 1974 / 2004], связано множество проблем, которые мы здесь не можем рассматривать, – например, является ли предикатив особой частью речи, как он соотносится с кратким прилагательным и наречием, какие подлежащные свойства имеет дативный субъект предикатива и т.п., ср. [Letuchii 2018], [Sigurðsson 2002].

Предметом рассмотрения в данной работе являются только инфинитивные конструкции с предикативом: Praed + Inf (рус. вариант обозначения: ПИнф).

Мы исходим из фундаментального постулата Московской семантической школы ([Apresjan 1974; 2006a]; [Paducheva 2004]), который выдвигали и другие исследователи и научные школы ([Wierzbicka 1985; 1988]): валентности (как и другие свойства языковой единицы) в общем случае определяются ее семантической структурой (при этом способы выражения (заполнения) валентностей, что важно для нашей темы (см. раздел 2), могут быть неканоническими, ср. [Boguslavskij 1996]). Если предикативы, обозначающие простейшие физиологические состояния и реакции, имеют только субъекта (это лицо, ср. ***Детям*** *холодно*, или окружающая среда, ср. ***На улице*** *холодно*), то предикативы более высокого ранга – эмоциональные реакции, восприятие, ментальные состояния, модальные предикативы – имеют еще одну валентность – пропозициональную, выражаемую инфинитивом (*Трудно решить эту задачу*) или зависимой клаузой (*Слышно, как дождь стучит по крыше*; *Приятно, что вы не забыли о нашей просьбе*; *Нужно, чтобы все пришли вовремя*). Иногда такая валентность обозначается единым термином «сентенциальный актант» [Letuchii 2018]. Мы будем различать на синтаксическом уровне сентенциальный актант (придаточное предложение) и зависимый инфинитив. На уровне семантической структуры, как будет показано ниже, инфинитивы имеют разный статус.

В русском языке есть разные модели сочетания главного предиката и зависимого инфинитива. Обычно они рассматриваются на материале глаголов. Для глагольных конструкций принято различать три разных статуса подчиненного инфинитива – субъектный, объектный и целевой (*пошел купить хлеба*). Нас интересуют две первые модели:

субъектный инфинитив – оба действия (и оба глагола) относятся к одному субъекту; формула – XX: *продолжил* (X) *разговаривать* (X), *удалось* (X) *отправить* (X) *телеграмму*: X-у удалось, X отправил. Назовем такую конструкцию **моносубъектной**;

объектный инфинитив: контролер (Y) и исполнитель (X). Формула –YX: *Командир приказал солдатам стрелять*: Y приказал, X должен стрелять (конструкция YX обычно может быть преобразована в придаточное: *приказал, чтобы стреляли*); *Родители запретили мальчику выходить из дома*: Y запретил, X не должен выходить. Назовем такую конструкцию **разносубъектной** (разумеется, есть случаи кореферентности Y и X: *приказал / запретил себе*, где Y = X).

Конструкции ПИнф с точки зрения соотношения субъектов обычно не рассматриваются. Между тем у сочетаний ПИнф могут быть те же два соотношения субъектов предикатива и зависимого инфинитива – моносубъектное (XX) и разносубъектное (YX), – что и у сочетаний глаголов.

Соотношение субъектов в конструкции ПИнф зависит от семантики предикатива. В литературе выделяются разные семантические классы предикативов – оценки (разных типов), ощущения, эмоции, ментальные, модальные, уместность, релевантность, эффективность, параметризуемый признак, свойства места и др., ср. [Letuchii 2018], [Serdobol'skaya, Toldova 2014], [Zimmerling 2017, 2018]. Однако нас эти классы интересуют не сами по себе, а относительно инфинитивной конструкции. Поэтому мы сразу исключаем перцептивные (*видно*, *слышно, как P*) и ментальные (*понятно*, *ясно*, *очевидно*, *известно*, *что P*) предикативы, которые не сочетаются с инфинитивом (**слышно стучать по крыше* ‘слышен стук’; **понятно читать* ‘понятен текст’).

С точки зрения инфинитивной конструкции предикативы можно свести к двум большим классам:

внутренние оценки, или эмоциональные реакции, – моносубъектные конструкции;

внешние оценки, или интерпретации, – разносубъектные конструкции.

Нас будут интересовать, в первую очередь, интерпретации, которые как особая группа до сих пор специально не описывались. Но рассматривать интерпретации мы будем на фоне и в сопоставлении с другим, более освоенным в лингвистических исследованиях классом – эмоциями (внутренними оценками). В работе [Serdobol'skaya, Toldova 2014] вводится противопоставление эмоциональной оценки (***Людям*** *интересно работать в консалтинге*), где обычно выражается субъект эмоционального состояния, и собственно оценки, где может выражаться субъект оценки (***Для меня*** *это* ***недорого***, с. 445) или другой участник – ориентир (собственно оценочный предикатив описывает ситуацию извне, «с объективных позиций» (с. 468), ср.: *Неделя началась* ***неудачно для доллара*** *и* ***удачно для акций***). Как видно из примеров, в работе [Serdobol'skaya, Toldova 2014] рассматриваются не только предикативы (в строгом смысле – сказуемые безличных

предложений), но и наречия (хотя все единицы при этом называются предикативами). И субъект, и ориентир в таких конструкциях могут выражаться дативом или предложной группой *для* Род. В приводимой ниже классификации используются несколько иные критерии противопоставления, хотя мы тоже выделяем группу эмоциональных оценок.

2 Внутренние оценки = эмоциональные реакции

The Ядро класса внутренних оценок, т.е. моносубъектного класса, составляют предикативы эмоциональной реакции. Эмоциональные предикативы (*страшно, стыдно, радостно, грустно, обидно, досадно, приятно, неприятно* и под.), как и хорошо описанные в литературе эмоциональные глаголы (ср. *радоваться*, *огорчаться* и под., ср. [Apresjan 1974], [Paducheva 2004], [Zaliznyak 2006]), обозначают комплексную психологическую ситуацию, которая включает три базовых компонента, характерных для эмоциональных реакций:

(а) Ситуация P – **причина**-содержание, является каузатором, вызывающим эмоциональную реакцию

(б) **Оценка** ситуации P – положительная или отрицательная

(в) **Переживание** – положительное или отрицательное (‘приятно’ vs. ‘неприятно’).

Переживание, т.е. состояние, само по себе не имеет валентности на ситуацию (состояние одноместно – оно имеет только субъекта), а оценка имеет пропозициональный объект P (пропозициональное содержание). В результате совмещения образуется гибридный предикат с пропозициональным актантом P. P может выражаться не только клаузой (ср.: *Мне очень больно, **что все так получается*** [Дмитрий Емец. Таня Гроттер и магический контрабас (2002)]), но и инфинитивом, ср.: *Вознесенскому было тяжело и больно **жить*** [Марина Зайонц. Если бы знать… (1990-2000)]. Кроме собственно эмоциональных оценок выделяется еще группа ментальных оценок – *удивительно, интересно, странно, смешно, забавно*.

В конструкции ПИнф происходит генерализация семантики предикативов и формируются два больших класса оценок – положительных (‘приятно’) и отрицательных (‘неприятно’). В результате процесса генерализации эмоциональные реакции пополняются предикативами из других семантических классов. Так, физиологические состояния-реакции могут «повышаться» до психологических реакций (с соответствующим изменением значения). Дальше всего на этом пути продвинулись *больно*, *горько* и *тяжело* (*Мне больно / горько, что я тебя обидел*; *Тяжело расставаться с друзьями*). В меньшей степени этим процессом затронуты *сладко* и *вкусно* (*Было удобно сидеть, было **вкусно пить** чай с мёдом* [Фридрих Горенштейн. Куча (1982)]; *Тома поняла, что ей **вкуснее смотреть**, как ест ребёнок, чем есть самой*… [Людмила Улицкая. Казус Кукоцкого, 2000]; ***Мне** очень **сладко, что** изнаночной канвой рассказа явилась самурайская книга «Хагакурэ»* [Интернет-альманах «Лебедь», 2003.09.28] ≈ ‘приятно’).

В инфинитивной конструкции с предикативами внутренней оценки семантически противопоставлены два класса инфинитивов: (1) инфинитивы перцептивных, ментальных и речевых глаголов, обозначающие «внутренние», информационные процессы в сознании человека (*видеть*, *воображать*, *сказать* и под.), которые мы будем называть информационными глаголами, и (2) инфинитивы «обычных», не-информационных глаголов.

2.1 Инфинитивы информационных глаголов.

С семантической точки зрения инфинитив не входит в сферу действия эмоционального предикатива: *страшно смотреть* не означает, что человеку страшно находиться с открытыми глазами, *страшно сказать* не значит, что страх вызывает само произнесение слов. Речь идет о другом – причиной и содержанием эмоциональной реакции, переживания является ситуация P, которая зависит от информационного глагола V: *Страшно смотреть* [V]*, как гибнут леса* [P] ≈ *Страшно, что гибнут леса* [P] (P = ‘гибнут леса’).

Субъект предикатива и инфинитива один и тот же (XX). Инфинитив V обозначает тот информационный канал, по которому поступает информация о P. Таким образом, синтаксически зависимый от предикатива инфинитив не заполняет семантическую валентность предикатива, но вводит ситуацию (пропозицию) P, которая, будучи валентностью инфинитива, семантически заполняет также валентность эмоционального предикатива (инфинитив может присоединять не только

клаузу, но и, например, номинализацию, ср.: *Мне больно* ***чувствовать*** *вашу* ***подавленность*** *и вашу* ***скорбь****!* [Борис Васильев. Вещий Олег (1996)]).

Данный семантико-синтаксический казус напоминает случай, который хорошо описан для предметных актантов – так называемое расщепление валентности: *Погладил голову ребенка – погладил ребенка по голове* (см. [Apresjan 1974: 153–155]. У *головы* как части тела есть валентность на целое (*ребенок*), но в предложении происходит переподчинение, и этот элемент высказывания (*ребенок*) приобретает связь с глаголом. Другая, еще более близкая аналогия экспликации канала восприятия – конструкции типа *видел своими глазами*. Глаза – это орган, который отвечает за восприятие, через который поступает информация из мира (здесь экспликация неотличима от дублирования).

В конструкции ПИнф ситуация Р расщепляется на собственно содержание и канал восприятия. Инфинитивы *смотреть, слушать*, *видеть*, *слышать* – канал, по которому информация Р поступает из внешнего мира; *думать/подумать*, *вспомнить/ вспоминать*, *представить, вообразить* – канал, через который информация Р поступает в активное поле сознания из памяти или актуально создается субъектом. *Страшно сказать* значит примерно то же: в сознании актуализировалась ситуация Р, о которой субъект сообщает. Актуализация ситуации Р сопровождается переживанием. В синтаксическом смысле здесь наблюдается процесс, обратный процессу расщепления предметной валентности: если там чужая, «дальняя» валентность приближалась к предикату, непосредственно подчинялась ему, то здесь своя валентность Р «отодвигается», «отдаляется» от эмоционального предикатива, и между ними возникает «мостик», прокладка в виде информационного глагола: *Страшно* (‘плохо’)*, что гибнут леса → Страшно смотреть, как гибнут леса*. При включении информационного инфинитива усиливается эмоциональная составляющая предикатива: из «очищенной» оценки (‘плохо’) он превращается (возвращается обратно) в эмоциональное переживание: ‘когда Х воспринимает ситуацию Р, Х испытывает отрицательные эмоции’.

Заметим (это важно для сравнения с предикативами второй группы – интерпретациями), что глаголы активного восприятия (*слушать*, *смотреть*) и пассивного восприятия (*слышать*, *видеть*) в эмоциональной группе практически нейтрализуются и значат одно и то же – поступление информации о ситуации Р, ср.: *страшно / неприятно / противно / приятно смотреть, как Р ≈ страшно / неприятно / противно / приятно видеть, как Р / что Р* (какие-то семантические различия, конечно, сохраняются, однако это отдельная тема, в которую мы не можем углубляться).

У некоторых предикативов данной группы (ср. *завидно*) все зависимые инфинитивы, по данным НКРЯ, – информационные (*завидно видеть, глядеть, представить, слушать, смотреть*).

2.2 Инфинитивы не-информационных глаголов.

Второй вариант инфинитива – не-информационные глаголы.

Среди конструкций с не-информационными инфинитивами встречаются случаи, когда инфинитив входит в семантическую сферу действия предикатива, т.е. когда эмоция распространяется на саму ситуацию V (V = Р): *Обидно* ***проиграть*** */* ***упустить*** *шпиона* (‘обидно, что проиграл / упустил шпиона’), – но такие случаи единичны.

В большинстве случаев не-информационный инфинитив V не входит в семантическую сферу действия предикатива, не заполняет его пропозициональную валентность, а служит (подобно информационному инфинитиву) «прокладкой», «связкой», переходным звеном к ситуации Р, которая и является сферой действия предикатива. Таким образом, конструкция оказывается смещенной, но, в отличие от случаев типа *Страшно видеть, что Р*, где Р эксплицитно выражается, слушающий должен сам реконструировать ситуацию Р, которая вызывает реакцию: *Мальчику* ***было страшно ходить*** *ночью через весь коридор в туалет* [«Столица», 1997.08.26] – *ходить* не входит в семантическую сферу действия *страшно*, т.к. здесь нет смысла ‘страшит, пугает ходьба’; мальчика страшит, пугает темнота (‘ночью’) и связанные с нею опасности; ***Олегу было страшно нажимать*** *кнопку звонка. Если дети решительно и однозначно настроены против него, он не представлял, как это переживет* [Виктор Мясников. Водка (2000)] – страх вызывает не нажатие кнопки звонка, а то, что за ним последует, – будущая встреча (возможно, неудачная).

3 Внешние оценки = интерпретации

Схема внешней оценки, как уже говорилось, – YX: *Ну, да это ваше дело, хоть и* ***неосмотрительно портить*** *здоровье* [К. М. Станюкович. Жрецы (1897)] – инфинитив здесь обозначает ситуацию V с субъектом X (‘X портит здоровье’), предикатив *неосмотрительно* обозначает внешнюю оценку субъекта-интерпретатора Y, который не участвует в ситуации V, а оценивает ее со стороны (‘Y считает поведение X-а неосмотрительным, предосудительным, не одобряет его’). Разумеется, поскольку оценка направлена на X-а, его поведение или поступки, она связана не только с Y-ом, но и с X-ом тоже. Однако такая ситуация всегда возникает в разносубъектных конструкциях: в случае *Командир приказал солдатам стрелять* приказ Y-а тоже направлен на X-а (связан с X-ом); важно, что сам Y в ситуации инфинитива не участвует.

Внешние оценки по-другому можно назвать интерпретациями. В работе [Apresjan 2006b] описаны семантические различия глагольных предикатов оценки и интерпретации. Однако проекция этой глагольной классификации на предикативы – тема отдельного большого исследования. Пока ограничимся следующими замечаниями. Суть внешней оценки (интерпретации) заключается в позиции интерпретатора: он смотрит на ситуацию не просто со стороны, как внешний наблюдатель, а, так сказать, с более высокой позиции – с позиции превосходства. Это очевидно в случаях осуждения X-а: *Безответственно / бессовестно / нечистоплотно / преступно использовать бюджетные средства для личного обогащения*. Но даже в случаях типа: *Наивно полагать, что P* – интерпретатор хотя и не осуждает субъекта, но рассматривает его как интеллектуально менее зрелого, чем он сам (субъект не разобрался в ситуации, в которой интерпретатор разобрался).

Интерпретация оценивает поведение, поступки или выбор человека, за которые тот несет ответственность, т.е., в конечном счете, – оценивает самого человека X, ср. [Kustova 2017]. При этом в случае интерпретации (*Неосмотрительно / бессовестно / недостойно* и т.д. *соглашаться на это*) сам X может не знать или не считать, что он поступает *неосмотрительно / недостойно* и т.д. (см. [Kustova 2004: 219]), – тогда как в случаях типа *Если бы вы знали, как* ***тяжело*** *и* ***больно отказывать*** [Родион Нахапетов. Влюбленный (1998)] X не может не знать, что ему тяжело и больно.

Разницу между внутренней и внешней оценкой можно проиллюстрировать, например, парой *страшный – трусливый*: *страшный* описывает эмоцию, которую испытывает субъект X по поводу какой-то ситуации P (‘X-у страшно’); *трусливый* оценивает сам факт реакции X-а на P со стороны внешнего наблюдателя, интерпретатора Y-а как тип поведения X-а (‘X трусит, боится’). На позицию другого человека или общества может встать и сам субъект, – но модель YX при этом все равно сохраняется, потому что в случае оценки себя со стороны субъект «раздваивается» на «судью» и «подсудимого».

Два основных класса интерпретаций – осуждение (неодобрение) и одобрение.

Диапазон неодобрения весьма широк: *авантюрно, безответственно, бессердечно, бессовестно, бесстыдно, бестактно, бесчеловечно, бесчестно, глупо, жестоко, зашкварно* (сленг.), *наивно, неблагородно, невежливо, неграмотно, недобросовестно, незаконно, неконструктивно, немилосердно, неосмотрительно, неосторожно, неправильно, непредусмотрительно, неразумно, нерационально, несерьезно, нескромно, несправедливо, нечестно, нечистоплотно, неэтично, низко, преступно, самонадеянно, самоуверенно, тщеславно, целесообразно, цинично*.

Отрицательных интерпретаций неизмеримо больше, чем положительных (из соображений экономии приводятся минимальные контексты из НКРЯ): ***Бессовестно выгонять*** *его на улицу*; ***Бестактно допытываться*** *у женщины, сколько она потратила*; ***Бесчестно применять*** *магию в схватке с воинами*; ***Бессердечно разлучать*** *мать с детьми*; ***Безответственно передавать*** *леса в долгосрочную аренду*.

Положительные интерпретации тоже встречаются: *С его стороны было разумно / предусмотрительно / дальновидно / правильно оставить машину перед офисом*; *С его стороны было благородно / правильно / справедливо / человечно / этично отказаться от наследства*, ср. также: *грамотно, остроумно, умно, хитро*.

Большинство этих оценок имеют этическое содержание (*жестоко* vs. *благородно*); в том числе такие предикативы, как *некрасиво* (***Некрасиво*** *обманывать друзей* / ***Некрасиво*** *грубить старшим* = ‘не следует, плохо’), которые из эстетической зоны смещаются в этическую. Некоторые оценки не относятся к этической сфере (*авантюрно, грамотно, остроумно, умно, хитро*).

Разумеется, приведенный список интерпретаций не исчерпывающий (тем более что он пополняется).

Если в моносубъектных конструкциях субъекты обычно выражаются дативом (***Ему*** *страшно подойти к обрыву*) или предложной группой *для Род.* (*Очень* ***для меня*** *тяжело переходить из 8-ой роты в 4-ую* [Александр Гнедин. Письма (1939-1941)], то в интерпретационных конструкциях субъекты X и Y не выражаются обычным способом. Субъект оценки Y (интерпретатор) может быть выражен вводными конструкциями: ***По-моему*** */* ***на мой взгляд****, неэтично / неправильно / неразумно обсуждать это с коллегами*. Указание на субъекта X инфинитивной ситуации V может осуществляться с помощью конструкции *с его стороны* / *со стороны X-а*: ***С его стороны*** *наконец просто* ***недобросовестно вести*** *неопределенное существование и, таким образом, быть вам в тягость* [К. М. Станюкович. Из-за пустяков (1881)]; ***Со стороны устроителей*** *выставки было в высшей степени* ***бестактно устраивать*** *ее именно в 1937 году* [Юрий Елагин. Укрощение искусств (1952)] (в языке XIX-начала XX вв. конструкции с дативом изредка встречались, ср.: *Нет, мне просто преступно с вами соглашаться* [Н. С. Лесков. Божедомы (1868)], но сейчас их в НКРЯ не обнаруживается).

Не все интерпретационные слова становятся предикативами и употребляются в инфинитивной конструкции ПИнф, но в последние десятилетия эта группа пополняется новыми единицами, хотя интерпретационная конструкция встречалась уже в XIX в. Вот данные из НКРЯ: левая колонка – XIX век + XX век до 1980 г., правая – конец XX (после 1980 г.) – начало XXI вв. (граница, разумеется, условна):

	До 1980 (180 лет)	После 1980 (40 лет)
Безответственно	—	5
Бессердечно	—	2
Бессовестно	10	6
Бесчеловечно	14	11
Благородно	7	4
Невежливо	21	12
Незаконно	2	6
Неосмотрительно	1	5
Неразумно	48	36
Нескромно	4	6
Несолидно	1	10
Нечестно	28	25
Неэтично	2	6
Преступно	48	18

Заметим, что цифры в колонках неравноценны, поскольку неравноценны временные интервалы. Например, *невежливо* имеет показатели 21 и 12, но 21 вхождение относится к интервалу в 180 лет, а 12 – к интервалу в 40 лет. Так что даже если в первой колонке число больше, пропорционально оно имеет меньший вес, чем второе число.

У некоторых оценочных предикативов примеры ПИнф пока единичны: *Вот и не решались показаться без оружия. Хотя, конечно,* ***грамотнее*** *было бы поскорее* ***избавиться*** … [В. Пронин. Слишком большое сходство (2017); *Однако же* ***поспешно*** *и* ***необдуманно*** *было бы* ***делать вывод****: не найден* **—** *не существовал* [«Вокруг света», 1994].

Напротив, некоторые единицы закономерно не употребляются в интерпретационной конструкции – например, слова со значением контролируемых усилий: **внимательно было читать / *сосредоточенно было слушать / *настойчиво было просить*. Такие единицы не переходят в класс интерпретаций и не функционируют в качестве предикативов (только как наречия или прилагательные: *настойчиво просил*; *настойчивая просьба*).

Интерпретационная конструкция существенно отличается от эмоциональной не только по способам выражения субъектов, но и по другим свойствам. Если большинство инфинитивов эмоциональной конструкции – информационные, причем активное и пассивное восприятие (*смотреть*

– *видеть*, *слушать – слышать*) нейтрализуются, то в интерпретационной конструкции информационные инфинитивы в исходном значении не встречаются, ср. **Несправедливо смотреть / видеть, как гибнут леса*. Если же такие глаголы все-таки попадают в интерпретационную конструкцию, с ними происходит семантический сдвиг: *Было бы, однако, совершенно* ***несправедливо видеть в*** *Кальвине уже в это время человека с совершенно сложившимися религиозными убеждениями и вполне выяснившимся враждебным отношением к католицизму* [Б. Д. Порозовская. Жан Кальвин (1898)] – *видеть* ≈ ‘усматривать, рассматривать, считать’; ***Безответственно / бесчеловечно смотреть*** *на все эти безобразия и не вмешаться* – ‘нельзя мириться с безобразиями’.

Таким образом, ситуации, обозначаемые информационными глаголами, в конструкциях интерпретации приравниваются к поступкам. Причем к поступкам приравниваются не только воззрения, которыми руководствуется человек и за которые он несет ответственность. Как поступок оценивается даже контролируемое восприятие (*смотреть*, *слушать*) в собственном смысле, например: ***Бестактно слушать*** *частные разговоры*.

4 Заключение

Сочетание с инфинитивом наиболее характерно и наиболее естественно для модальных (*нужно*, *необходимо*) и близких к ним по значению предикативов (*выгодно*, *полезно* ≈ ‘следует делать’; *бесполезно, бессмысленно, вредно, губительно, рискованно* ≈ ‘не следует делать’). Для других предикативов конструкция с инфинитивом является достаточно искусственной. Однако природа этой искусственности разная.

В случае внутренних оценок инфинитивная конструкция чаще всего является результатом «выделения», экстрапозиции информационного предиката (*страшно смотреть*, *горько сознавать*). В случае же интерпретационных предикативов за счет инфинитивной конструкции восполняется важная грамматическая лакуна. Для обычного предиката (глагола) основной является предикативная (финитная) форма (сказуемое), но, кроме того, он может иметь адъективную репрезентацию (причастие) и адвербиальную репрезентацию (деепричастие). У интерпретационных смыслов, если они не выражаются глаголом, как бы нет предикативной репрезентации, а есть только адъективная (прилагательное: *нерациональный*, *неосторожный*, *несправедливый*, *самонадеянный*) и адвербиальная (наречие: *нерационально*, *неосторожно*, *несправедливо*, *самонадеянно* – причем это особые наречия – с так называемой плавающей сферой действия, ср. [Filipenko 1998; 1999; 2003]). Предикатив восполняет грамматическую лакуну – недостаток глагольной репрезентации для интерпретационных смыслов.

Если снабдить интерпретационную лексику семантической разметкой, то различие в конструкциях ПИнф можно использовать в программах автоматической обработки текста для поиска семантической сферы действия предикатива. Разумеется, будут какие-то случаи неразличения, когда предикатив эмоциональной реакции развивает значение интерпретации, ср.:

Странно было читать эту рецензию = внутренняя оценка, реакция: ‘странная рецензия, в рецензии было написано что-то странное’;

Странно читать такие рецензии = внешняя оценка, интерпретация: ‘странно, что X читает такие рецензии’, какие именно рецензии – неизвестно; со стороны X-а это странное, неодобряемое поведение.

Однако это касается небольшого количества предикативов типа *странно*, *смешно* (и для них можно предусмотреть специальные правила). В остальных случаях автоматический анализ сочетания ПИнф может осуществляться по общему правилу с учетом семантической пометы.

Acknowledgements

Исследование выполнено при финансовой поддержке РФФИ и Национального научного фонда Болгарии, проект № 20-512-18005.
The reported study was funded by RFBR and National Science Foundation of Bulgaria (NSFB), project number 20-512-18005.

References

[1] Apresjan Yu.D. Lexical semantics [Leksicheskaya semantika]. — Moscow: Nauka, 1974.

[2] Apresyan Yu.D. Foundations of systemic lexicography [Osnovaniya sistemnoj leksikografii] // Linguistic picture of the world and systemic lexicography [Yazykovaya kartina mira i sistemnaya leksikografiya]. Ed. by Yu.D. Apresyan. — Moscow: YaSK, 2006a. — P. 33–160.

[3] Apresjan Yu.D. Lexicographic type: verbs of interpretation [Leksikograficheskii tip: glagoly interpretatsii]. // Linguistic picture of the world and systemic lexicography [Yazykovaya kartina mira i sistemnaya leksikografiya]. Ed. by Yu. D. Apresjan. — Moscow: YaSK, 2006b. — P. 145–160.

[4] Boguslavskij I.M. The Scope of lexical units [Sfera dejstvija leksicheskih edinic]. — Moscow: YaSK, 1996.

[5] Filipenko M.V. On adverbials with a floating and fixed scope (to the question of actants and non-actants of a predicate) [Ob adverbialakh s plavayushchei i fiksirovannoi sferoi deistviya (k voprosu ob aktantakh i neaktantakh predikata)] — Semiotics and informatics [Semiotika i informatika]. Issue 36. — Moscow, 1998. — P. 120–139.

[6] Filipenko M.V. Predicative adverbs in -O Predikativnye narechiya na -O // Typology and theory of language. From description to explanation [Tipologiya i teoriya yazyka. Ot opisaniya k ob"yasneniyu]. — Moscow: YaRK, 1999. — P. 503–510.

[7] Filipenko M.V. Semantics of adverbs and adverbial phrases [Semantika narechij i adverbial'nyh vyrazhenij]. — Moscow: Azbukovnik, 2003.

[8] Kustova G.I. The Types of Derived Meanings and Language Extension Mechanisms [Tipy proizvodnykh znachenii i mekhanizmy yazykovogo rasshireniya]. — Moscow: YaSK, 2004.

[9] Kustova G.I. Adjective in the text as a reduced predication [Prilagatel'noe v tekste kak redutsirovannaya predikatsiya] // Problems of functional grammar. Predicative categories in utterance and whole text [Problemy funktsional'noi grammatiki. Predikativnye kategorii v vyskazyvanii i tselostnom tekste]. — Moscow: YaSK, 2017. — P. 224–246.

[10] Letuchii A.B. Predicatives [Predikativy]. Materials for the corpus grammar of the Russian language [Materialy k korpusnoi grammatike russkogo yazyka]. Issue III. Parts of speech and lexical and grammatical classes [Chasti rechi i leksiko-grammaticheskie klassy]. — Sankt-Peterburg: Nestor-Istoriya, 2018. — P. 136–192.

[11] Paducheva E.V. Dynamic models in lexical semantics [Dinamicheskie modeli v semantike leksiki]. — Moscow: YaSK, 2004.

[12] Serdobol'skaya N.V., Toldova S.Yu. Constructions with evaluative predicatives in Russian: participants in the evaluation situation and the semantics of the evaluative predicate [Konstruktsii s otsenochnymi predikativami v russkom yazyke: uchastniki situatsii otsenki i semantika otsenochnogo predikata]. — Acta Linguistica Petrotolitana. Transactions of the Institute for Linguistic Studies [Trudy Instituta lingvisticheskikh issledovanii]. Vol. X. Part 2. — Sankt-Peterburg: Nauka, 2014. — P. 443–477.

[13] Shcherba L.V. Language system and speech activity [Yazykovaya sistema i rechevaya deyatel'nost']. — Moscow: URSS, 2004 (1-st ed. – 1974).

[14] Sigurðsson H.A. To be an Oblique Subject: Russian vs. Iselandic // Natural Language and Linguistic Theory. —2002. — Vol. 20.

[15] Wierzbicka A. Lexicography and conceptual analysis. — Ann Arbor: Karoma, 1985.

[16] Wierzbicka A. The semantics of grammar. — Amsterdam: Benjamins, 1988.

[17] Zaliznyak Anna A. Polysemy in language and ways of its representation [Mnogoznachnost' v yazyke i sposoby ee predstavleniya]. — Moscow: YaSK, 2006.

[18] Zimmerling A.V. (2017) Russian Predicatives in the Perspective of Sociolinguistic Experiment and Corpus Grammar [Russkie predikativy v zerkale eksperimenta i korpusnoi grammatiki]. // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2017) [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Po materialam Mezhdunarodnoy Konferentsii “Dialog” (2017)],]. Issue 16. Vol. 2. Moscow, pp. 466–481.

[19] Zimmerling A.V. Impersonal constructions and dative-predicative structures in Russian // Voprosy Jazykoznanija. — 2018, No. 5, pp. 7–33.

Animacy in the use of anaphoric and demonstrative pronouns in Russian and French

Alexander Letuchiy
HSE University
Moscow, Staraya Basmannaya st., 21/4
alexander.letuchiy@gmail.com

Elena Nikishina
HSE University
Moscow, Staraya Basmannaya st., 21/4
helene_nikichina@mail.ru

Abstract

The article focuses on the role of animacy in Russian and French pronominal systems. Although animacy is a grammatical category only in Russian, while in French it is not reflected in the behavior of nouns, it turns out that some animacy-based restrictions on the use of anaphoric and demonstrative pronouns are common for the two languages. We address syntactic restrictions that affect the following types of uses: (i) use of anaphoric pronouns in copular constructions; (ii) repetition of anaphoric pronouns for the sake of clearness and / or emphasis; (iii) deictic use of anaphoric pronouns; (iv) anaphoric use of demonstrative pronouns. In all the four cases, except, perhaps, the fourth one, pronouns tend to have an animate referent, while inanimate ones are more problematic. We conclude that these restrictions mainly result from the fact that animate objects have a greater discourse importance and more often become the main subject of the discourse than inanimate ones. At the same time, degree of strictness of restrictions sometimes differ between the two languages: for instance, demonstrative pronouns in the anaphoric use tend to have an animate antecedent in Russian, while for French, this tendency is weaker.

Key words: Russian; French; animacy; pronouns; pronoun repetition; copular constructions; anaphoric uses; deictic uses

DOI: 10.28995/2075-7182-2021-20-464-472

Роль одушевлённости в употреблении анафорических и указательных местоимений в русском и французском языках

Александр Летучий
Национальный исследовательский университет
Высшая школа экономики
Москва, ул. Старая Басманная, 21/4
alexander.letuchiy@gmail.com

Елена Никишина
Национальный исследовательский университет
Высшая школа экономики
Москва, ул. Старая Басманная, 21/4
helene_nikichina@mail.ru

Аннотация

В статье обсуждается роль одушевлённости в системах местоимений русского и французского языков. Хотя только в русском одушевлённость является грамматической категории, а во французском она не отражается в морфологии существительных, выясняется, что местоимения в двух языках подпадают под параллельные друг другу ограничения на употребления. Мы рассмотрим данные следующих типов: (i) употребление анафорических местоимений в связочных конструкциях; (ii) конструкции с повтором анафорических местоимений; (iii) дейктические употребления анафорических местоимений; (iv) анафорические употребления дейктических местоимений. Во всех этих случаях, кроме отчасти последнего, существенно легче употребляются местоимения с одушевлёнными референтами. Мы делаем вывод, что эти ограничения в первую очередь связаны с большей дискурсивной значимостью одушевлённых участников по сравнению с неодушевлёнными. В то же время строгость ограничений в двух языках не всегда совпадает: тенденция к употреблению русского указательного местоимения *тот* с одушевлённым антецедентом в русском явно сильнее, чем параллельная ей во французском.

Ключевые слова: русский язык; французский язык; одушевлённость; местоимения; конструкции с повтором; связочные конструкции; анафорические употребления; дейктические употребления

1 Введение

Наша работа посвящена роли одушевлённости в русской и французской системах местоимений. На первый взгляд, основание для сравнения здесь довольно зыбкое, потому что одушевлённость занимает разное положение в системах двух языков. В русском она является полноценной грамматической категорией (см. [Klenin 1983], [Grannes 1984], [Крысько 1994] о складывании этой категории): с одной стороны, она выступает как словоклассифицирующая для существительных, с другой, проявляется в их склонении:

(1) *Я увидел стол / быка*.

Как и род, падеж и число, одушевлённость используется при согласовании прилагательных, причастий и местоимений-прилагательных:

(2) *Я увидел большой стол, стоявший в углу*.

(3) *Я увидел высокого человека, стоявшего в углу*.

(4) *Поговорим про твой дом*.

(5) *Поговорим про твоего дядю*.

Как многие грамматические категории, грамматическая одушевлённость не всегда стопроцентно совпадает с семантической. Рассмотрим примеры (6) и (7), в которых показано, что одну и ту же команду можно назвать *«Зенит»* и *чемпион России.* Несмотря на то, что референт у этих выражений один, первое (как все названия команд) является грамматически неодушевлённым, а второе (как весь ряд обозначений статуса команды, таких как *лидер, чемпион, бронзовый призёр, претендент на золото, соперник* и т.д.) — одушевлённым:

(6) *Армейцы в предыдущем туре премьер-лиги обыграли «Зенит»*. [Антон Сычев. ЦСКА разгромил «Анжи» в гостях // Известия, 2014.03.24]

(7) *Динамовцы неожиданно обыграли чемпиона России со счетом 1:0.*
[Шунин призвал фанатов «Динамо» сильно не радоваться победе над «Зенитом». https://www.gazeta.ru/sport/news/2020/08/27/n_14852929.shtml?utm_source=yxnews&utm_medium=desktop]

В отличие от русского языка, во французском для существительных не существует грамматического различия по одушевлённости. Что же касается системы местоимений, то они во французском языке как раз чувствительны к одушевлённости антецедента, хотя оппозиция по одушевленности реализуется лишь в определённых контекстах и только в позиции косвенного объекта (см. обсуждение данного противопоставления в Grevisse & Goosse 2008):

(8) *Tu dois répondre à tes parents — Tu dois leur répondre* ‘Ты должен ответить своим родителям — Ты должен им ответить’

(9) *Tu dois répondre à leurs lettres — Tu dois y répondre* ‘Ты должен ответить на их письма — Ты должен на них ответить’

(10) *Je pense à mon ami — Je pense à lui* ‘Я думаю о своем друге — Я думаю о нём’ (при одушевлённом косвенном объекте).

(11) *Je pense à mon projet — J’y pense* ‘Я думаю о своем проекте — Я думаю о нём’ (при неодушевлённом косвенном объекте).

(12) *Il s'occupe de ses enfants — Il s'occupe d'eux* 'Он занимается своими детьми — Он занимается ими'

(13) *Il s'occupe de ses papiers — Il s'en occupe* 'Он занимается своими документами — Он ими занимается'

Стандартным способом оформления объекта во французском языке являются местоимения-клитики ряда *le, la, les*, *lui, leur* (см. Black 1982, Wust 2010 об их распределении и линейном расположении). Однако данные примеры показывают, что одушевлённость влияет на способ оформления: при данных глаголах при одушевлённых непрямых объектах используется местоимение-клитика *leur* (8) или сочетание предлога с ударным местоимением (10), (12), а при неодушевлённых — специализированные местоимения *y* (заменяющее группу с предлогом *à*) и *en* (заменяющее группу с предлогом *de*). В то же время сами авторы книги [Grevisse, Goosse 2008] отмечают, что распределение способов маркирования по одушевлённости может нарушаться под влиянием определённых стилистических и прагматических факторов: приводятся примеры типа (14), где местоимение *y* отсылает к одушевлённому участнику:

(14) *C'est un homme équivoque, ne vous y fiez pas.*
'Это неоднозначный человек, не доверяйте ему'.

Целью настоящей статьи является показать, что, несмотря на все различия, частные ограничения на одушевлённость референта, существующие в русском и французском языках, в ряде случаев параллельны друг другу. Заметим, что ранее некоторыми авторами (например, Летучий 2015, Циммерлинг 2020) уже указывалось, что характеристика по одушевлённости не сводится к противопоставлению типов склонений, а распадается на два или несколько грамматических противопоставлений.

Оставшаяся часть статьи организована следующим образом. В части 2 обсуждается роль одушевлённости в ограничениях, которые налагаются в русском и во французском на употребления анафорических местоимений. Часть 3 продолжает анализ тех же местоимений, но уже в их дейктических употреблениях. В части 4 кратко анализируется вклад одушевлённости в употребление указательных местоимений. В заключении (часть 5) подводятся итоги работы.

Используемые в работе примеры взяты из Национального корпуса русского языка (www.ruscorpora.ru) и с сайтов, найденных поиском в системе Google. Некоторые французские примеры были переведены или оценены нашей информанткой, носительницей французского языка Неж Рошан, кроме того, используются примеры и их оценки, данные в исследованиях и учебных пособиях по французскому языку. Кроме этого, отдельные русские примеры, не иллюстрирующие ключевых утверждений, построены авторами. Оценки русских примеров получены интроспекцией и, в ряде случаев, опросом информантов, при этом данные оценки могут показаться дискуссионными.

2 Ограничения на употребление анафорических местоимений

2.1 Ограничение 1. Употребление местоимений в связочных конструкциях

Первое из рассматриваемых ограничений характеризует связочные конструкции. Для русского языка речь идёт о конструкциях с нулевой связкой (ниже НС). В целом НС сочетается и с одушевлёнными, и с неодушевлёнными субъектами:

(15) *Иван Куринной — опытный фотограф…* [коллективный. Фотополигон // «Русский репортер», № 48 (78), 18–25 декабря 2008]

(16) *Баку — красивый, вальяжный город на берегу моря*. [Токарева Виктория. Своя правда // «Новый Мир», 2002]

Однако если субъект выражается анафорическим местоимением, то в некоторых конструкциях с НС (как правило, в тех, где вторая часть является именной группой)[1] это местоимение в норме относится к одушевлённому референту **(ограничение 1)**. Для неодушевлённых возможно почти исключительно выражение субъекта с помощью указательного местоимения *это* (см. о его употреблениях Падучева 1985):

(17) *Он человек, близкий к гениальности, он красивый человек.* [Василий Аксенов. Звездный билет // «Юность», 1961]

(18) ??*Я много раз был в Пятигорске.* ***Он*** *красивый город.* (ср. нормальное *Это красивый город*).

В НКРЯ практически не обнаруживается примеров, где местоимения ряда *он*, употреблённые в связочной конструкции, сочетались бы с неодушевлённой именной группой. Немногочисленные примеры выглядят стилистически маркированными: в примере (19) из «Театрального романа» более типичным и нейтральным было бы употребление конструкции *Это чужой мир*. А пример (20), возможно, несколько архаичен.

(19) *Я в него не пойду.* ***Он —*** *чужой мир. Отвратительный мир!* [М. А. Булгаков. Записки покойника (Театральный роман) (1936-1937)]

(20) *У Достоевского народ хорош не потому только, что* ***он*** *простой народ и бедный народ, а потому, что он народ верующий, православный*. [К. Н. Леонтьев. Достоевский о русском дворянстве (1891)]

Сходное ограничение наблюдается и во французских конструкциях с глаголом-связкой *être.* Как правило, если субъект выражен анафорическим местоимением, оно имеет одушевлённый антецедент. Если антецедент неодушевлённый, обычно используется указательное местоимение *ce*:

(21) *Je connais Jean depuis longtemps.* ***Il*** *est une personne extraordinaire. /C'est une personne extraodrinaire.*
'Я давно знаю Жана. Он замечательный человек. / Это замечательный человек'.

(22) ??*J'aime Nice.* ***Elle*** *est une belle ville* (ср. нормальное *C'est une belle ville*).
'Мне нравится Ницца. Она красивый город' (ср. нормальное 'Это красивый город').

При этом указательное местоимение *ce* допустимо и при одушевлённых, и при неодушевлённых референтах, ср. *C'est une personne extraordinaire*.

2.2 Ограничение 2. Употребление в конструкции с повтором референта

Ещё одно общее для двух языков ограничение связано с некоторыми конструкциями с повтором референта. Так, французская эмфатическая конструкция с повтором местоимения более естественна, когда референт местоимения одушевлён:

(23) ***Lui****, il avait envie d'aller au cinéma* 'Ну а он хочет пойти в кино'.[2]

(24) ??***Lui****, il a été acheté chez le bouquiniste* 'Ну а она была куплена у букиниста' (например, о книге).

[1] Легко заметить, что конструкции, где после связки следует не именная группа, а прилагательное, не подпадают под это ограничение: примеры вида *J'ai apporté des pommes. Elles sont bonnes* 'Я принес яблоки. Они вкусные' приемлемы, в противоположность примерам типа ??*Elles sont de bons fruits* 'Они хорошие фрукты'.
[2] В данном примере в разговорной речи возможно даже опущение местоимения-подлежащего: *Lui avait envie d'aller au cinéma*, в этом случае нестандартное оформление подлежащего имеет тот же выделительный эффект.

В русском языке полностью параллельной французской конструкции нет. Однако если референт сначала выражается местоимением, а затем поясняется с помощью полной именной группы, в этом контексте одушевлённые референты тоже несколько уместнее неодушевлённых:

(25) ***Он**, старый, поселковский бухгалтер, прожил много лет и много на своем веку поездил.* [Владимир Маканин. Утрата (1984)]

(26) ??***Он**, старый чемодан, давно уже без дела стоял в углу.*

Ограничения 1 и 2, видимо, объясняются сходным образом. Для **ограничения 2** естественнее всего дискурсивное объяснение: в дискурсе одушевлённые референты более значимы, чем неодушевлённые, именно они в первую очередь подвергаются выделению и являются наиболее естественными заполнителями для конструкций с двойным выражением референта, которые делают участника более значимым (см. [Kibrik 2011]. Для **ограничения 1** объяснение менее очевидно, но можно предположить, что при отсутствии связочного глагола (русский) или полнозначного глагола (французский) местоимение приобретает особую выделенность — в клаузе нет выраженного глагола, который с большой вероятностью был бы наиболее выделен интонационно и в смысловом отношении. Именно для одушевлённых референтов естественно такое выделенное положение местоимений (см. также обсуждение противопоставления слабых и сильных местоимений в Cardinaletti, Starke 1999, Testelets 2003).

3 Ограничения на дейктические употребления местоимений

Ещё одно ограничение касается уже не анафорических, а дейктических употреблений местоимений. В отличие от рассмотренных выше контекстов, данное ограничение по-разному проявляется в русском и во французском. Русское местоимение *он / она / оно /они* имеет, как известно, не только анафорические, но и дейктические употребления. В этом случае ((28), (29)) говорящий указывает жестом на объект, о котором говорит, а в предшествующей коммуникации этот объект не встречается. При этом местоимение в дейктическом употреблении выделено интонацией и/или несёт специфические маркеры указания типа *вот* (28) или *вон*. В примере (27) дополнительно указывает на дейктическое употребление положение местоимения *его* перед глаголом *спросить*.

(27) *Не веришь, **его** спроси, — указывая на Потапенко, отвечал Каландадзе.* (А.Ф, Ростков. Первые гвардейцы-танкисты. http://nkosterev.narod.ru/vov/mem_2/rostk_34.html)

(28) [Ипполит (Юрий Яковлев), муж, 47, 1928] *ботиночки на тонкой подошве/ **вот он** знает. **Он** всё знает.* [Эльдар Рязанов, Эмиль Брагинский. Ирония судьбы, или С легким паром, к/ф (1975)]

Употребление *он / она / оно / они* в дейктической функции почти всегда связано с одушевлёнными участниками. Реплика с указанием типа (29) выглядит странно, если говорящий указывает на неодушевлённый объект:

(29) *??Ты ведь мебель хотел покупать? Купи (вот) его* (указывая на стол).

Во французском языке основные анафорические местоимения: *il, elle, ils, elles, la, le, les, lui, leur* — не имеют дейктической функции. В частности, это связано с тем, что они не могут, в отличие от ряда *он* в русском языке, нести на себе интонационное выделение[3]. Дейктическую функцию принимают на себя специальные конструкции и маркеры: например, сочетание анафорического местоимения с частицами с адвербиальным значением *là* и *ci* (при этом *ci* употребляется значительно реже).

[3] Это относится, конечно, к приглагольным местоимениям (pronoms conjoints), а не к ударным (pronoms toniques — *lui, elle, eux, elles*), которые ведут себя во фразе более автономно.

Носитель французского языка, владеющий русским, переводит примеры типа *Ты ведь машину хотел продать? Продай ему* [реплика сопровождается указательным жестом или акцентным выделением]*, он как раз хочет купить* или *Ты хотел знать, что посмотреть в Киеве? Спроси его* [указательный жест или акцентное выделение]*, он как раз оттуда приехал*, используя сочетания частицей *là*. Примеры без этой частицы он считает неграмматичными в дейктическом значении:

(30) *Demande(-le) à **lui-là** / Demande(-le) lui à **lui-là**.*[4]

(30') *#Demande(-le)-**lui**.*
'Спроси (это) вот у него [указательный жест]'.

(31) *Vends-la (lui) à **lui-là**.*

(32) *#Vends-la-**lui**.*
'Продай (её) вот ему [указательный жест]'.

При этом в сочетаниях с *là* действует ограничение на одушевлённость, похожее на русское. Дейктические употребления такого рода невозможны для неодушевлённых объектов — ср. примеры (33) и (34), где местоимения занимают позицию прямого дополнения: первый из них (с одушевлённым референтом) возможен, а второй (с неодушевлённым референтом) носитель считает сомнительным:

(33) [Ты думаешь, кого позвать на День рождения?] *Invite **lui-là** / **#Invite-le**.*
'Пригласи вот его [указательный жест]'.

(34) [Ты хотел купить фрукты?] ??*Achète **lui-là / #Achète-le**.*
'Купи вот его (например, арбуз) [указательный жест]'.

Для неодушевленного референта в этом контексте используется другая местоименная стратегия — указательное местоимение *ça* или *celui-là*:

(34') *Achète celui-là / Achète ça.*

Данные ограничения ещё более объяснимы, чем рассмотренные в части 2 ограничения на анафорические употребления. Стандартно дейктические местоимения выделены в дискурсе — они не имеют текстового антецедента, поэтому к их употреблению необходимо привлечь внимание, чтобы интерпретировать их правильно. Одушевлённые объекты в целом сильнее выделены, чем неодушевлённые.

Кроме того, одушевлённые объекты в глазах говорящих хорошо дифференцированы друг от друга. Неодушевлённые объекты хорошо представляются как единая масса — не случайно для них возможны указательные конструкции типа *Возьмите вот это Prenez ça*, где *это* может указывать и на один, и на несколько предметов.

4 Немного об указательных местоимениях

Выше мы говорили только об основных для каждого из языков анафорических местоимениях: о ряде *он / она / оно / они* для русского и о местоимениях *il / elle, le / la* и т.д. для французского. Скажем кратко о другом типе местоимений — указательных: русском *тот* и французской паре *celui-ci* и *celui-là*. При этом, подобно тому, как анафорические местоимения могут употребляться указательно, верно и обратное: местоимения, у которых главное употребление — указательное, могут употребляться с текстовым антецедентом, то есть анафорически.

[4] Во втором варианте примера (30) непрямой объект, видимо, получает дополнительное выделение, поскольку выражен и местоимением-клитикой *lui*, и предложной группой *à lui.* Однако в данной статье мы не рассматриваем отдельно этот тип конструкций.

Местоимение *тот* может употребляться и как субстантив (аналог именной группы), и как адъектив (модификатор именной группы). В адъективном употреблении *тот* допускает и одушевлённые, и неодушевлённые антецеденты, ср. *Я был в* ***том*** *городе, Я знаю* ***того*** *человека.* Однако в своём анафорическом субстантивном употреблении (см. о нём в работах [Крейдлин, Чехов 1988], [Подлесская 2020] и др.) *тот* сочетается почти исключительно с одушевлёнными антецедентами, как в (35). Исключения типа (36) встречаются и грамматически приемлемы, однако довольно редки.

(35) *После смерти тверского князя Михаила Ярославича, основавшего город, Старица досталась в удел одному из его четырех сыновей, а* ***тот*** *в свою очередь завещал ее своему сыну Семену…* [М. Б. Бару. Таракан на канате // «Волга», 2016]

(36) *Иномарка от неожиданности вздрагивает, начинает гудеть и останавливается. За ней — фура, а* ***ту*** *подпирает уже вся колонна, проделавшая по пыли, грязи и колдобинам долгий путь.* (https://zavtra.ru/blogs/2011-11-1553)

Во французском языке наиболее близка по функциям к местоимению *тот* пара местоимений *celui-ci* и *celui-là*. Если они употреблены вместе, в контексте противопоставления, то они указывают на два объекта, в разной мере удалённые от говорящего и слушающего. Например, в пространственном контексте (37) *celui-ci* обозначает близкий к участникам коммуникации объект, а *celui-là* — более удалённый.

(37) *Mets les fleurs dans les vases, les roses dans* ***celui-ci*** *et le jasmin dans* ***celui-là***.
‘Поставь цветы в вазы, розы в эту, а жасмин в ту’.

Эти употребления никак не ограничены одушевлённостью референтов и могут относиться и к одушевлённым, и к неодушевлённым именам. Однако нас будет интересовать другое, анафорическое употребление, где употребляется одиночное местоимение.

(38) *En ce qui concerne le protocole sanitaire,* ***celui-ci*** *reste inchangé!*
(https://www.instagram.com/competencesco/)
‘А что касается эпидемиологического протокола, **он** остается без изменений’.

По-видимому, у анафорического употребления указательных местоимений есть некоторая тенденция к одушевлённости референта, как в (39) (она отражается и в том, что в грамматических исследованиях и учебных пособиях на данные местоимения обычно приводятся примеры с одушевлённым антецедентом). Однако она явно менее заметна, чем для русского *тот*: примеры с неодушевлёнными объектами, к которым относится анафорическое *celui-ci* в значении ‘тот, последний’ (ср. (40)), без труда находятся в Интернете, хотя примеров с одушевлёнными референтами явно больше.

(39) *…Emmanuel Macron a enjoint au président brésilien Jair Bolsonaro de respecter l'Accord de Paris sur le climat. En retour,* ***celui-ci*** [Болсонару] *a mis en scène son mépris pour le « premier monde »* (https://www.lexpress.fr/...)
‘… Эмманюэль Макрон призвал бразильского президента Жаира Болсонару соблюдать Парижское климатическое соглашение. В ответ **тот** [Болсонару] выказал презрение к странам «первого мира»’.

(40) *…une jeune femme s'indigne d'avoir été sanctionnée d'une amende de 135 euros à cause d'un tract, qu'elle dit avoir « ramassé par terre ».* ***Celui-ci*** *se trouvait dans la poche arrière de son pantalon.* (https://www.liberation.fr/checknews/2020/09/15/...)
‘… молодая женщина возмущается наложенным на неё штрафом в 135 евро за листовку, которую она, по её словам, «подняла с земли». **Она** лежала у неё в заднем кармане штанов’.

Объяснение асимметрии, существующей в русском, может крыться в выделенности участников: тяготение *тот* к одушевлённым контекстам (и сходную, хотя и слабую, тенденцию в случае *celui-ci*) можно объяснить большей дискурсивной выделенностью одушевлённых участников.

Когда эти местоимения начинают функционировать как анафорические, они прежде всего закрепляются за наиболее важными для коммуникации участниками, являющимися её основной темой. Как правило, таковыми являются именно одушевлённые участники, поэтому именно они стандартно являются антецедентами *mot* и *celui-ci* в значении ‘он, последний’.

5 Заключение

В данной статье мы рассмотрели ограничения на употребление русских и французских местоимений. Как выяснилось, хотя в грамматической системе двух языков местоимения занимают разное место, в обоих языках их функционирование подпадает под ряд ограничений, связанных с одушевлённостью.

Ограничения накладываются на различия между системами местоимений двух языков. В частности, во французском языке анафорические местоимения употребляются дейктически только при наличии специальных маркеров, а в русском это не обязательно. Однако в обоих системах дейктические употребления ограничены, в основном, одушевлёнными референтами.

В некоторых случаях ограничения различаются по степени жёсткости. Как было показано выше, местоимение *mot* в анафорическом употреблении довольно сильно тяготеет к одушевлённым антецедентам, а для французского *celui-ci* эта тенденция слабее.

Наблюдаемые ограничения объясняются, в основном, различными дискурсивными свойствами одушевлённых и неодушевлённых референтов. Во-первых, одушевлённые объекты гораздо чаще, чем неодушевлённые, являются основной темой высказывания или текста и в гораздо большей мере способны нести на себе выделение. Во-вторых, одушевлённые объекты лучше индивидуализируются, неодушевлённые же скорее могут представляться как нечленимая масса и часто обслуживаются местоимениями типа *это*, не дифференцированными по числу референтов.

Кратко скажем о теоретической значимости наших результатов. Описанные нами асимметрии не объясняются и в русле подхода Я. Г. Тестельца (2003), ориентированного на различии между адъюнктными и аргументными группами. Тестелец утверждает, что в русском существуют не классы слабых и сильных местоимений, а «слабые позиции» — адъюнктные позиции в предложных и именных группах. Тем не менее, рассматриваемые нами контексты явно не относятся к слабым — местоимения в них допустимы, ограничения затрагивают лишь те из них, которые имеют неодушевлённые референты (антецеденты). Следовательно, предпочтительна трактовка в терминах А. Кардиналетти и М. Штарке (1999), которые считают, что противопоставлены сами местоимения с одушевлённым антецедентом (сильные) и с неодушевлённым (слабые). В то же время и этот подход не безупречен: он сталкивается с проблемой исчисления контекстов. Наши ограничения не относятся к ядру класса контекстов, описанных в [Cardinaletti, Starke 1999]: он включает сочинительные конструкции, модификацию местоимений адвербиалами (типа *даже* или *только*) и др. Учёт наших контекстов потребовал бы сильного дробления языковых данных: пришлось бы считать, например, что контексты с нулевой связкой и прилагательным в предикатной позиции не различают сильные и слабые местоимения (ср. *Он ещё молодой / ещё свежий*), а контексты с нулевой связкой и именной группой в предикатной позиции различают (*Он хороший человек / *хороший город*).

Поэтому мы предпочитаем считать, что двух по-настоящему противопоставленных синтаксических классов местоимений в русском всё-таки нет. На поведении местоимений сказываются дискурсивные особенности одушевлённых и неодушевлённых объектов и связанные с ними особенности именных групп и местоимений с одушевлёнными и неодушевлёнными референтами. В силу того, что речь идёт не о двух классах местоимений, а о влиянии на местоимения дискурсивных свойств, это влияние градуально и связано со свойствами конкретных конструкций: например, в конструкции типа *Он ещё свежий* местоимение *он* допустимо, в частности, потому, что прилагательное *свежий* требует слова, с которым бы согласовывалось по роду и числу (*он*), тогда как в **Он хороший город* этого требования нет, и местоимение *он* неприемлемо, в отличие от *это*, снижающего дискурсивную выделенность субъекта. При этом различия в дискурсивной выделенности одушевлённых и неодушевлённых референтов приводят к грамматическим запретам или хотя бы ограничениям тогда, когда конструкция предполагает сильную выделенность участника, выраженного местоимением.

Благодарности

Мы выражаем благодарность участникам конференции Groupe d'Études en Linguistique Textuelle Constrastive (GELiTeC) (Форли, Италия, проводилась онлайн, 13-15 мая 2021 г.), в частности, Ольге Иньковой, Златке Генчевой, Ольге Артюшкиной, Малгожате Новаковской за интересные вопросы и плодотворную дискуссию. Исследование выполнено при финансовой поддержке РФФИ и ЧНФ в рамках научного проекта № 20-512-26004 (The reported study was funded by RFBR and GACR, project number 20-512-26004).

Литература

[1] Black James R. (1982), The structure and placement of French clitic pronouns, University of London Ph.D. dissertation.
[2] Cardinaletti Anna, Starke Michal (1999), The typology of structural deficiency: on the three grammatical classes, University of Venice Working Papers in Linguistics (UVWPL), Venice.
[3] Grannes Alf (1984), Impersonal animacy in 18th century Russian, Russian Linguistics, Vol. 8 (3), pp. 295–311.
[4] Grevisse Maurice, Goosse André (2008), Le bon usage. 14e édition, Editions De Boeck Université, Bruxelles.
[5] Kibrik A.A. (2011), Reference in discourse, Oxford University Press, Oxford.
[6] Klenin Emily (1983), Animacy in Russian: a new interpretation, Slavica Publishers, Columbus, OH.
[7] Kreidlin G.E., Chekhov A.S. (1988), Sootnošenie semantiki, aktual'nogo členenija i pragmatiki v leksikografičeskom opisanii anaforičeskix mestoimenij (na material mestoimenij gruppy TOT) [Interaction of semantics, topic / comment structure and pragmatics in the lexicographic description of anaphoric pronouns (on the example of TOT group of pronouns)], Problemnaja gruppa po eksperimental'noj i prikladnoj lingvistike. Predvaritel'nye publikacii [Research group in experimental and applied linguistics preprints], Russian Language Institute, Moscow, Vol. 178.
[8] Krys'ko V.B. (1994), Razvitije kategorii oduševlennosti v istorii russkogo jazyka [The development of the animacy category in the history of Russian], Lyceum, Moscow.
[9] Letuchij A.B. (2015), Mestoimenija i oduševlennost' v russkix konstrukcijax s nulevoj svjazkoj [Pronouns and animacy in Russian zero copula constructions], Zimmerling A.V., Lyutikova E.A., Konoshenko M.B. (eds.), Tipologija morfosintaksičeskix parametrov. Materialy meždunarodnoj konferencii [Typology of morphosyntactic parameters. Proceedings of the international conference], Volume 2, pp. 200–225, MPGU editions, Moscow.
[10] Paducheva E.V. (1985), Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'ju [Utterance and its relations with the real world], Nauka, Moscow.
[11] Podlesskaya V.I. (2020), "A tot Perovskoj ne dal vslast' pospat'": prosodija i grammatika anaforičeskogo tot v zerkale korpusnyx dannyx ["A tot Perovskoj ne dal vslast' pospat'": prosody and grammar of the anaphoric tot (based on the corpus data)], Kompjuternaja lingvistika i intellektual'nye texnologii. Po materialam ježegodnoj meždunarodnoj konferencii "Dialog" (Moscow, June 17-20, 2020) [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2020"], Volume 19 (26), pp. 628–643.
[12] Testelets Ya.G. (2003), Are there strong and weak pronouns in Russian?, Formal Approaches to Slavic Linguistics #11: The Amherst Meeting 2002, Michigan Slavic Publications, pp. 515–538.
[13] Tsimmerling A.V. (2020), Oduševlennost'. Russkij jazyk [Animacy. The Russian language], Trudy instituta russkogo jazyka im. V.V. Vinogradova RAN [Proceedings of Vinogradov Russian Language Institute of Russian Academy of Sciences], Vol. 24, pp. 43–56.
[14] Wust Valerie (2010), L2 French Learners' Processing of Object Clitics: Data from the Classroom, L2 Journal, Vol. 2, pp. 45–72.

The semantic component ‘scale’ in the meaning of a discourse particle *uzh*

Irina Levontina
Russian Language Institute RAS
irina.levontina@mail.ru

Abstract

The modal particle *uzh* is perhaps the most difficult Russian discourse word to describe since its semantics is highly elusive. The existing descriptions are rather abstract and poorly correlate with various cases of usage of *uzh*. Besides, they do not take into consideration several crucial components of this particle’s meaning. For instance, in phrases like *Uzh ya-to znayu* (‘I do know’) one can notice a hugely important component of meaning - the idea of a scale. One can say *Ya-to etot sekret znayu, a vot drugim nevdomek* (‘I do know the secret, whereas others have no idea about it’), and in this example, uzh would be irrelevant. *Uzh ya-to eto znayu* presupposes that others probably know it too, but it’s me who knows it for sure. This very idea of a scale and poles together with the idea of the exceedance of expectations (which is also important for the meaning of *uzh*) constitutes the semantic contribution that this particle makes. Moreover, *uzh* partly smooths the opposition between the central and other elements of a multitude, because it does not exclude them from consideration, it just gives emphasis to that one.

The aim of this research is to examine those types of *uzh* usage, where the idea of a scale is most clearly actualized. Probably, if we understand how the significant components of this particle’s meaning function, we will get closer to the development of a complete picture of its usage. For example, the idea of a scale within the meaning of *uzh* is expressed in the context of a special question (*Zachem uzh tak zlo?* ‘Why so mean?’). In an argument *uzh* often implies that the speaker was almost ready to back down, but not to this extent - like in a famous poem by Daniil Kharms called «Liar» (1930). The idea of a scale is vividly realized in the context of an implicit (*Gde uzh mne!*, ‘How can I…’) or explicit negation. It is especially interesting to pay attention to the peculiar effects of the combination of uzh with comparative forms (*luchshe uzh,* ‘it would be better...’). The usage of *uzh* in standard word combinations *raz uzh, esli uzh, togda uzh* has its restrictions, also connected with the idea of a scale. The development of a modal meaning in a temporal word, which brings the transformation of a timeline into a scale of expectations or possibilities, is quite typical.

Keywords: Particles; discourse; Russian Language; scale
DOI: 10.28995/2075-7182-2021-20-473-482

Семантический компонент ‘шкала’ в значении дискурсивной частицы *уж*

И. Б. Левонтина
Институт русского языка им. В. В. Виноградова РАН

irina.levontina@mail.ru

Аннотация

Модальная частица *уж*, возможно, самое трудное для описания дискурсивное слово русского языка с плохо уловимой семантикой. Существующие очень абстрактные толкования с трудом соотносятся с разнообразными употреблениями уж. Кроме того, в них не вполне учтены некоторые существенные компоненты значения этой частицы. Так, во фразах типа *Уж я-то знаю* хорошо заметна чрезвычайно важная часть значения *уж* – идея шкалы. Можно сказать: *Я-то этот секрет знаю, а вот другим невдомек*, но *уж* было бы в этом случае неуместно. *Уж я-то это знаю* предполагает, что другие, возможно, знают тоже, но обо мне это можно сказать совершенно точно. Именно эта идея шкалы и полюса и составляет, в сочетании с идеей превышения ожиданий, также чрезвычайно важной для значения уж, собственный вклад частицы. Более того, *уж* даже

отчасти гасит противопоставление, поскольку оно показывает, что остальные элементы множества не исключаются из рассмотрения, а просто данный элемент оказывается особо выделен. Задача настоящей работы – рассмотреть несколько типов контекстов употребления уж, в которых идея шкалы реализуется наиболее ярко. Возможно, если мы поймем, как работают важные компоненты смысла частицы, это приблизит нас к построению целостной картины ее функционирования. Так, идея шкалы в значении *уж* проявляется в контексте специального вопроса (*Зачем уж так зло?*). В споре уж часто выражает идею, что говорящий готов был бы пойти на уступки, но не до такой степени – как в знаменитом стихотворении Д. Хармса «Врун» (1930). Идея шкалы ярко реализуется в контексте имплицитного (*Где уж мне!)* или эксплицитного отрицания. Особенно интересно обратить внимание на своеобразные эффекты сочетания уж с компаративом (*лучше уж*). В типичных сочетаниях раз уж, если уж, тогда уж использование частицы имеет определенные семантические ограничения, также связанными с идеей шкалы. Развитие у временно́го слова модального значения, при котором происходит трансформация временно́й шкалы в шкалу ожиданий или возможностей, вполне типично.

Ключевые слова: частицы, дискурс, русский язык, шкала

0.

Модальная частица *уж*[1], возможно, самое трудное для описания дискурсивное слово русского языка. Не случайно Е. В. Урысон характеризует модальное *уж* как «особое разговорное словечко с почти неуловимой семантикой, некий "наполнитель" высказывания, придающий ему идиоматичную разговорную окраску» [Урысон 2007: 539]. Эту частицу неоднократно исследовали, ей даже специально посвящено некоторое количество работ [Paillard 1986-87, Mendoza 1999, 2000, Урысон 2007, Левонтина 2008]. Довольно подробное перечисление контекстов употреблений частицы *уж* содержится в таких словарях, как «Словарь структурных слов русского языка» [Морковкин 1997], «Словарь русских частиц» [Шимчук, Щур 1999].

В работе Д. Пайара «*Уж,* или необсуждаемое» [Paillard 1986-87] предлагается толкование *уж*, смысл которого таков: *уж* маркирует некоторую ценность p как не подлежащую обсуждению и одновременно противопоставляет ее ценности p', которая эксплицитно или имплицитно задана в предшествующем контексте, однако в качестве преодоленной, отвергаемой, недействительной. Это, как и другие толкования инварианта *уж*, ориентировано в первую очередь на тот круг контекстов *уж*, который в словарях обычно рассматривается как реализация первого из его модальных значений (*Уж я-то знаю*). В словаре [Шимчук, Щур 1999] *уж* здесь толкуется в том смысле, что оно «выделяет из некоторого множества предмет, признак или событие, которые в ситуации выбора по тем или иным причинам должны быть рассмотрены в первую очередь».

Сходно с пайаровским толкование Цыбатова [Zybatow 1990]: *уж* указывает на выраженную в левом контексте или выводимую неуверенность в том, что входящее в сферу действия частицы утверждение соответствует действительности и подчеркивает, что оно не может не соответствовать действительности. На эту же тему – толкование И. Мендосы [Mendoza 2000]:

уж

(a) маркирует какую-то ценность W как действительную;

(b) причем имплицируется также ценность W', которая недействительна;

(c) причем ценность W' имплицитно или эксплицитно фигурирует с предшествующем тексте или может быть выведена.

Как было показано в [Левонтина 2008], эти очень абстрактные толкования с трудом соотносятся с чрезвычайно разнообразными употреблениями *уж*. Кроме того, в них не вполне учтены некоторые существенные компоненты значения этой частицы. Так, в работе отмечалось, что в фразах типа *Уж я-то знаю* хорошо заметна чрезвычайно важная, на наш взгляд, часть значения *уж* – идея **шкалы**. Можно сказать: *Я-то этот секрет знаю, а вот другим невдомек*, но *уж* было бы в этом случае неуместно. *Уж я-то это знаю* предполагает, что другие, возможно, знают тоже, но обо мне это можно сказать совершенно точно. Именно эта идея шкалы и полюса и составляет, в сочетании с идеей превышения ожиданий, также чрезвычайно важной для значения *уж*, собственный вклад частицы в значение высказывания, в то время как идея **противопоставления** (ср. толкование Пайара), скорее всего, наводится контекстом. Более того, *уж* даже отчасти гасит противопоставление, поскольку оно показывает, что остальные элементы множества не исключаются из рассмотрения, а просто данный элемент оказывается особо выделен[2].

[1] Временно́е *уж* как вариант или синоним *уже* здесь не рассматривается.

[2] Отчасти это соотносится с компонентом ‘в первую очередь’ в дескрипции Шимчук и Щур.

Задача настоящей работы – рассмотреть несколько типов контекстов употребления *уж*, в которых идея шкалы реализуется наиболее ярко. Возможно, если мы поймем, как работают важные компоненты смысла частицы, это приблизит нас к построению целостной картины ее функционирования.

1.1.

Чрезвычайно наглядно идея шкалы в значении *уж* проявляется в контексте специального вопроса:

(1) *[nemoki, nick] И учитель абсолютно бессилен и бесправен по сути. [irga101, nick] Почему уж так бессилен? Я вот взяла да уволилась, когда мне не понравилась одна школа. [Сегодня в топе блогов история учительницы (блог) (2008)]*

Как можно заметить, второй говорящий оспаривает именно **степень** бессилия, а не само его наличие. Первый говорящий заявляет об **абсолютном** бессилии, а второй возражает, что оно не абсолютно. Гораздо хуже звучал бы диалог: *Учитель бессилен! – Почему уж бессилен?* Он, однако, становится более естественным при наличии продолжения типа: *Возможностей у учителя, конечно, мало, но какие-то рычаги все же есть*, где проясняется, что второй говорящий тоже оспаривает именно степень.

В подобных случаях сочетание вопросительного слова с *уж* обычно сопровождается показателем высокой степени, чаще всего *так*:

(2) *Да и перед Западом как будто непонятно становилось: отчего **уж** я **так** не оправдываюсь, **ни единым словом**? может, в чём-то клевета и права? [А. И. Солженицын. Бодался теленок с дубом (1967-1974)]*

Как мы видим, здесь Солженицын сомневается именно в правильности такой **категоричности** – он говорит не о том, что зря не оправдывался, а о том, что зря не оправдывался совсем, *ни единым словом*.

В следующих примерах речь также идет о чрезмерности:

(3) *— С мальчиком дела плохи! — угрюмо произнес Володя. — Вы же, наверное, все слышали, вам Харламов рассказывал… — Рассказывал, но я не понимаю, почему уж так плохи дела, вторичное кровотечение наступит не обязательно [Юрий Герман. Дорогой мой человек (1961)]*

(4) *Профессор Серебряков тоже человек. Зачем **уж** так презирать его? Он не гангстер, не половой психопат, он хотел жить, любил женщину, по-своему, в меру своих сил, и годами без устали занимался одним — писал, писал, писал, писал. [Юрий Трифонов. Предварительные итоги (1970)]*

В первом случае говорящий считает прогноз слишком пессимистичным, во втором – отношение к чеховскому герою чрезмерно, несправедливо критическим. Фразы без *так* – *Почему **уж** плохи дела, Зачем **уж** презирать его* – гораздо менее естественны. Конечно, такое возможно, если само по себе слово ссылается на очевидную шкалу, указывая на ее крайнюю точку: *Зачем **уж** насмерть, Зачем **уж** ненавидеть, Почему **уж** гениально*. Но это нетипично. С другой стороны, подобные употребления *уж* в контексте указания на высокую степень без обозначения самого признака (часто в сочетании со словом *так*) как раз очень естественны:

(5) *Я купил за три тысячи долларов «белый» военный билет <…>. Майор, военком, сказал мне, что где-то там в бумагах напишет мне косоглазие и отсутствие обеих конечностей. Когда я ужаснулся: мол, зачем **уж** так?! — он резонно заметил: — Повесткой безногого не вызовешь [Виктор Слипенчук. Зинзивер (2001)]*

Говорящий не против того, чтобы ему приписали болезнь, но отсутствие ног – это чересчур.

(6) *— От женщины, которая таким делом занимается, может вытошнить. — Ну, зачем **уж** так. Хорошую женщину никакое дело не испортит. [И. Грекова. На испытаниях (1967)]*

(7) *— Куличи ладно — только не вздумай их святить, не буду есть. — Да почему **уж** так? — <…> — А бабушка всегда святила, и красила. Что ж это, не наша вера? [А. И. Солженицын. На изломах (1996)]*

Рецензенты «Диалога» справедливо отмечают, что во всех этих случаях фигурируют исключительно вопросы о причине и цели (*Почему уж, зачем уж, отчего уж, к чему уж*). На самом деле другие вопросы тут тоже возможны: ***На что уж** тут так обижаться?; **Что уж** такого особенного он знает?; **Сколько уж** он на нас потратил?*; ср. также примеры:

(8) *Надя была в полубреду. Она произносила имена Г.Шенгели и В.Нарбута с какими-то подозрениями (оказывается, и им Осип читал свое стихотворение). А **кого уж** тут подозревать, если я знаю теперь 14 слушателей, а где гарантия, что их не было больше? [Эмма Герштейн. Вблизи поэта (1985-1999)]*

(9) *— А мне вчера исполнилось восемьдесят лет. — Поздравляю, — говорю я. Он вздыхает. — Да **с чем уж** там поздравлять? — Ну хотя бы с тем, что вы до этого возраста дожили! [Владимир Войнович. Замысел (1999)]*

Действительно, однако, вопросы о цели и причине в данном случае наиболее типичны. Как отметил один из рецензентов, здесь перед нами не прямые вопросы, а более сложные речевые акты – вопросы, ставящие под сомнение сказанное собеседником. Конечно, отдельно трудно себе представить фразу **Что **уж** ты купил?* (звездочка поставлена рецензентом). Однако возможен диалог, в котором первый говорящий попрекает, что, мол, все в доме куплено им, а второй парирует: *Что **уж** ты такого купил?* – то есть, может, что-то и купил, не так много и не такое дорогое. Вопрос *Что ты купил?* подразумевает, что говорящий хочет получить перечень покупок. Вопрос *Что уж ты купил?* подразумевает, что говорящий оспаривает степень существенности покупок, которую собеседник, по его мнению, преувеличил. При этом фраза *Что уж ты купил?* – это все-таки вопрос, и на него вполне естественно услышать ответ типа: *Ну не так уж и мало: вот, например, этот холодильник, да и микроволновку тоже.* Этим она отличается от фраз типа *Где уж мне!*, которые никакого ответа не предполагают; о них речь пойдет в следующем разделе.

Здесь стоит еще обратить внимание на то, насколько разный эффект вызывает добавление разных частиц в контекст специального вопроса. Сравним *уж* с частицей *-то* (*Где-то он теперь?)* [Левонтина 2016]. Как будто обе частицы как-то усугубляют вопрос и притом усложняют иллокутивную цель, но совершенно по-разному. *-То* сообщает высказыванию функцию погружения в фантазии или воспоминания, при этом ответ от собеседника возможен, но не особенно ожидаем. *Уж* отчасти превращает вопрос в возражение. Естественно, что эти частицы предпочитают контексты с разными вопросительными словами, и если для *уж* вопросы с *зачем* и *почему* наиболее характерны, то для *-то* они как раз практически невозможны.

1.2.

Идея шкалы ярко реализуется в контексте имплицитного отрицания – например, в следующем типе контекстов *уж*: *Какой уж тут отдых!; Куда уж мне судить об этом;*

(10) *Не без трепета приступаю я к написанию этого текста. Если уж Витгенштейна, по его мнению, не понял <…> сам Рассел, сумевший привести в замешательство великого Фреге, то где **уж** нам. [В. А. Успенский. Витгенштейн и основания математики (2002)]*

(11) *Подсчитали: пахать и то нельзя на этом комбайне, где **уж** тут копать картошку. [Анатолий Азольский. Лопушок // «Новый Мир», 1998]*

(12) *И затем, немного возвысив голос: — Исполнилось пророчество: "Не зарастёт священная тропа!.." Не зарастёт, думаю. Где **уж** ей, бедной, зарасти. Её давно вытоптали эскадроны туристов… [Сергей Довлатов. Заповедник (1983)]*

Уж здесь факультативно: возможно и *Где тут копать картошку* и т. п. *Уж* усиливает риторический эффект, показывая, что ситуация не просто отличается от стандартной, а максимально от нее далека.

В контекстах этого типа предпочитаются совсем другие вопросительные слова, чем в предыдущем случае, и при этом они десемантизированы: в сочетаниях *где уж* и *куда уж* отсутствует темпоральная семантика.

Уж хорошо сочетается и с эксплицитным отрицанием, причем оспаривается именно степень выраженности какой-то характеристики. В этом отношении характерно частое использование *уж* в составе оборота *не так (уж) и*:

*(13) Музыкой все это предприятие назвать было трудно, но <…> в жестянку, подвешенную к гармонике, монеты падали не так **уж** и редко… [Дина Рубина. Медная шкатулка (сборник) (2015)]*

*(14) Пострадавшему на всякий случай вкатили в живот семнадцать уколов от гипотетического бешенства, и все его жалели, хотя было не так **уж** и больно. [Алексей Варламов. Купавна // «Новый Мир», 2000]*

Уж в составе этого оборота факультативно:

(15) И выстрел наконец грохнул — не так и далеко, в сосняке. [Василь Быков. Болото (2001)]

(16) "В общем, я не так и стара", — Таня чуть не подпрыгнула от этой мысли. [Василий Аксенов. Пора, мой друг, пора (1963)]

Однако использование этой частицы в подобных контекстах чрезвычайно типично. Идея шкалы в значении частицы очень органично сочетается с ними, и нередко частица оказывается прагматически почти обязательной.

Рассмотрим сочетания: *не такой уж бедный, уж не такой бедный, уж не бедный, уж и бедный*. Во всех этих случаях, в том числе без слова *такой*, которое само по себе указывает на степень, *уж* указывает, что говорящей считает чрезмерной приписываемую кому-то степень бедности. Если шкала не очевидна и не введена, то высказывание с *уж* будет странным: *?уж не итальянский (а французский)*.

1.3.

В связи с ключевой для *уж* идеей шкалы особенно интересно обратить внимание на своеобразные эффекты сочетания *уж* с компаративом (чаще всего *уж лучше / лучше уж* и *скорее уж / уж скорее*, но возможно также *уж правильнее / красивее / разумнее / честнее было бы* и т. п.). Сравним два диалога:

a.- Вам налить чаю? – Лучше кофе

и

b. - Вам налить чаю? – **Уж** лучше кофе.

Различие между ними очевидно. Репликой без частицы *уж* говорящий просто сообщает о своих предпочтениях, в то время как реплика *Уж лучше кофе* означает, что ни чай, ни кофе не соответствуют его желаниям: например, он голоден и вообще-то хотел бы поесть, или он рассчитывал, что предложат что-нибудь покрепче чая и кофе, но из двух не очень хороших вариантов второй все-таки более приемлем. Приведем еще несколько примеров:

*(17) Чем так торговать, как говорится, лучше **уж** воровать. [Борис Екимов. Пиночет (1999)]*

*(18) Тут было какое-то смутное чувство, подсказывавшее, что лучше **уж** я — униженный, чем я — отрёкшийся от себя. [Фазиль Искандер. Мученики сцены (1989)]*

*(19) — Говорят, ты стал писателем? Я растерялся. Я не был готов к такой постановке вопроса. **Уж** лучше бы она спросила: "Ты гений?" Я бы ответил спокойно и положительно.. [Сергей Довлатов. Чемодан (1986)]*

*(20) — Тише, тише, вы всё не так делаете, вы нам только мешаете, вы **уж** лучше помолчите. [Владимир Войнович. Иванькиада, или рассказ о вселении писателя Войновича в новую квартиру (1976)]*

Естественно, что *уж* хорошо сочетается и с единицами типа *получше*:

(21) *— Если ты сейчас же не перестанешь, я отберу у тебя ведро и отдам его другой девочке! — Хорошей? — спросила Гуля. — **Да уж получше** тебя, — ответила мама. [Елена Ильина. Четвертая высота (1945)]*

(22) *Теперь стало поживописнее. Идет маленький бритый татарин какой-нибудь в чибитейке, или глупый чуваш, или разряженная мордовка. **Все уж получше**. [В. А. Соллогуб. Тарантас (1845)]*

Говорящий не настаивает, что предлагаемый вариант хорош, но альтернативный вариант точно хуже. Та же идея реализуется и в сочетании *уж не хуже*:

(23) *И Экссон Петролеум не смог бы без Ходорковского, который до недавнего совсем времени был прекрасным покровителем и мог разруливать пусть хуже Абрамовича, но **уж не хуже** Фридмана. [Сергей Доренко. Левые силы - перезагрузка (2003) // «Завтра», 2003.08.13]*

(24) *— Был бы у тебя брат, сам бы сейчас завтраки готовил! — Да **уж не хуже** бы получилось. [Юлия Лавряшина. Улитка в тарелке (2011)]*

(25) *— Тебе бы, Петрович, с лекциями выступать. — А что, — не смутился Елахов, — **уж не хуже** некоторых бы выступил. [Влада Валеева. Скорая помощь (2002)]*

Напротив, в сочетании *уж не лучше* обсуждается хороший вариант, однако он не может быть настолько же хорош, как альтернативный:

(26) *Жена вкусно готовит… **Уж не лучше** меня, наверное… [Ольга Новикова. Мне страшно, или Третий роман // «Звезда», 2003]*

1.4.

Сходная идея решения в условиях сокращенного выбора представлена в еще одном типичном для *уж* круге употреблений – в сочетаниях *раз уж, если уж, тогда уж*:

(27) *Я совершенно не уверен, выйду ли я отсюда, но если **уж** выйду, то плюну на всё, что я здесь пережил и видел, и забуду их, чертей, на веки вечные. [Ю. О. Домбровский. Факультет ненужных вещей, часть 2 (1978)]*

(28) *Если **уж** говорить о сделке, то следует исходить из принципа "деньги вперёд". [Борис Дмитриев. Ни пяди назад // «Коммерсантъ-Власть», 1998]*

(29) *Раз **уж** всего напечатать нельзя, то ясно, что будет произведён отбор публикуемых произведений, — а именно только того, что хочет начальство. [И. М. Дьяконов. Книга воспоминаний (1995)]*

(30) *А теперь, раз **уж** проснулись, я вас посмотрю. Послушала сердце, измерила давление. Вполне прилично! Ничего похожего на то, что было. [И. Грекова. Перелом (1987)]*

(31) *Раз **уж** поехали… к югу, как ты выражаешься, надо соответственно и вести себя… Или уж сиди дома, не езди. А куда к югу-то? [Василий Шукшин. Печки-лавочки (1970-1972)]*

(32) *Он сказал: — Тогда я сделаю вот что. Я тебя поцелую. — И это лишнее, — возразил Краснопёров. Забулдыга постоял в раздумье. Затем взглянул на Красноперова и твёрдо произнёс: — Тогда **уж** я как минимум — спою. [Сергей Довлатов. Иная жизнь (1984)]*

(33) *Раз **уж** заговорил про Ленинград, стоит вспомнить забавный эпизод, связанный с одним из моих самых любимых драматических актёров. [И. Э. Кио. Иллюзии без иллюзий (1995-1999)]*

(34) *Понимаете? Наберите ещё раз. А если опять никто не отвечает… — Ну тогда **уж** прямиком в милицию, — пообещал я [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]*

В контекстах **сокращенного выбора** реализуется та идея, которую Д. Пайар считает центральной для значения *уж*, – идея необсуждаемого. Выбор производится из ограниченного круга возможностей, остальные исключены из рассмотрения. Здесь важно обратить внимание, что *уж* уместно в контекстах сокращенного выбора вовсе не всегда. Когда сокращение выбора не

приводит к увеличению возможностей, *уж* не используется. Например, один человек просит другого заехать куда-либо по дороге. Сравним два возможных ответа. В первом случае собеседник отвечает:

*(35)a. Я не знаю, буду ли завтра выезжать, но если **уж** буду, то, конечно, заеду.*

Здесь задается некая шкала затрат времени и сил, и *уж* очень естественно. Ср., однако, другой ответ:

*b.??Я не знаю, буду ли завтра выезжать, но если **уж** буду, то вообще в другую сторону.*

Это очень странная фраза, здесь уместна была бы не частица *уж*, а частица *даже*, поскольку в данном случае возможная поездка никак не продвигает ситуацию в направлении возможности выполнения просьбы. Шкала не выстраивается, и *уж* оказывается неуместным.

Вообще использование дискурсивных слов в условных, причинных, уступительных контекстах всегда сопряжено с невероятно разнообразными семантическими и прагматическими эффектами; об этом есть большая литература, см., например, [König 1991]. Стоит также отметить, что что в поведении *уж* в подобных контекстах обнаруживается сходство с другим дискурсивным словом – *хоть*. В знаменитой работе П. Паршина [Паршин 1988] семантика *хоть* описывается через понятие деонтического диалога. *Хоть* является средством переговоров, торга в широком смысле и дает совершенно разные эффекты в зависимости от роли говорящего (проситель или контролер) и от характера сценария (позитивного или негативного): *– Дай хоть рубль! – Хоть десять, мне не жалко! / – Хоть копейку, ничего не дам. Хоть* и *уж* – слова со скалярной семантикой, приспособленные для такого рода торга, и механизмы образования прагматических эффектов у них довольно похожи. Ср. *– Позволь **уж** мне попробовать! – Ладно **уж**! / - Нет **уж**, ни за что.*

1.5.

В споре уж часто выражает идею, что говорящий готов был бы пойти на уступки, но не до такой степени – как в знаменитом стихотворении Д. Хармса «Врун» (1930):

(36)— А вы знаете, что У?
А вы знаете, что ПА?
А вы знаете, что ПЫ?
Что у папы моего
Было сорок сыновей?
Было сорок здоровенных —
И не двадцать,
И не тридцать,-
Ровно сорок сыновей!
— Ну! Ну! Ну! Ну!
Врешь! Врешь! Врешь! Врешь!
Еще двадцать,
Еще тридцать,
Ну еще туда-сюда,
*А **уж** сорок,*
Ровно сорок, –
Это просто ерунда!

Интересно, что здесь в паре с *уж* выступает *еще*: тридцать *еще* туда-сюда, а *уж* сорок – никак. Идея шкалы видна очень наглядно. Обе частицы могут присоединяться как к теме, так и к реме – и даже, как еще в этом примере – и к теме, и к реме (*Еще тридцать – еще туда сюда*). Здесь сказано *Уж сорок – это ерунда*, но возможно и *Сорок – это уж ерунда*.

Именно обыгрывание идеи шкалы является в этом стихотворении одним из основных источников комизма. Если в первой итерации числа 20-30-40 действительно задают шкалу, то в других непонятно, почему одно менее правдоподобно, чем другое:

Что на небе
Вместо солнца

Скоро будет колесо?
Скоро будет золотое —
Не тарелка,
Не лепешка, –
А большое колесо!
<…>
Что под морем-океаном
Часовой стоит с ружьем?
<…>
Ну, с дубинкой,
Ну, с метелкой,
Ну еще туда-сюда,
А с заряженным ружьем –
Это просто ерунда!

Конечно, тарелка или лепешка в качестве заменителя солнца ничем не лучше и не хуже колеса, а часовой с ружьем может стоять под морем с тем же успехом, что часовой с метелкой. Однако использованная частица навязывает представление о шкале правдоподобия. Эта шкала совершенно фантастична, и при этом она задана имплицитно и подается как нечто само собой разумеющееся. За счет высокой степени суггестивности и создается это ощущение невероятного абсурда[3].

В подобных контекстах *уж* часто фигурирует в сочетании с другими частицами: *ну, прям, да, и, так*:

(37) *Как — есть чувство, что вы немного отец этим ребятам?» — «Немного есть», — уныло соглашался Владимир. —* ***Да уж****, — не верила Ася. —* ***Уж прям уж****. — Что вы все такие ядовитые? — вдруг яростно схватывалась Таня [Ирина Полянская. Сельва (1996)]*

(38) *Везет же тебе… Мой и не ездит и прав не имеет и к машине только чтобы сесть рядом подходит… Будущая жена. Везет же тебе…* ***Да уж прям****. [Учимся водить (2007-2008)] [*

(39) *Тебе-то, молодому, еще все нипочем, а товарищ генерал у вас — пожилые, им бы поберечься. —* ***Ну, уж и*** *пожилые, — обиделся генерал слегка игриво. — Я еще таких молодых двоих заменю [Г. Н. Владимов. Генерал и его армия (1994)]*

(40) *— Как это благородно с вашей стороны, несмотря на ночь, не отказаться от визита, — сказал он с ледяной учтивостью. — Да ради рюмки он и до Москвы доползет, — сказала Адель и указала лекарю на дверь. —* ***Ну уж и*** *ради рюмки, — обиделся Иванов. [Булат Окуджава. Путешествие дилетантов (Из записок отставного поручика Амирана Амилахвари) (1971-1977)]*

[3] Другой источник комического в этом стихотворении тоже имеет лингвистическую природу. В последней части говорится:

Что до носа
Ни руками,
Ни ногами
Не доехать,
Не допрыгать,
Что до носа
Не достать!
– Ну! Ну! Ну! Ну!
Врешь! Врешь! Врешь! Врешь!
Ну, доехать,
Ну, допрыгать,
Ну еще туда-сюда,
А достать его руками -
Это
Просто
Ерунда!

Здесь обыгрывается омонимичность фразы *Это просто ерунда*. Во всех фрагментах она означает ‘Это бессмысленно’, а в последней – ‘Это очень легко’.

*(41)— **Ну уж и** обрадуются, — сказал Соколов. — Так же, как спортсмены радуются, когда не они, а кто-нибудь другой устанавливает рекорд. [Василий Гроссман. Жизнь и судьба, часть 2 (1960)]*

*(42)— Ну, может быть, там какие-нибудь деликатные женские подробности, — шутливо нахмурился прокурор.— Вот всё вам **так уж и** выложить! [Ю. О. Домбровский. Факультет ненужных вещей, часть 4 (1978)]*

Возможны даже самостоятельные реплики *Ну уж! Уж прям!* и. п., выражающие реакцию на реплику, которая кажется слишком неправдоподобной.

2.

Наличие в значении модального *уж* компонентов ‘шкала’ и ‘превышение ожиданий’ совершенно естественно: они связаны с тем временным значением *уж*, в котором оно близко к временно́му *уже* и указывает на опережение. Развитие у временно́го слова модального значения, при котором происходит трансформация временно́й шкалы в шкалу ожиданий или возможностей, вполне типично. Об этом применительно к значению частиц *уже* и *уж* см. [Урысон 2007; 536-360]. Ср. следующий пример, в котором невозможно различить временную шкалу и шкалу ожиданий:

*(43)Мы, не помню **уж** зачем, приехали в Сталинабад, и по пути обратно в горы заехали на ГАЗ-51 на рынок. [Александр Городницкий. «И жить еще надежде» (2001)]*

*Не помню **уж*** – значит ‘раньше помнил, но за давностью лет забыл’ или ‘не помню, да это и несущественно’?

Итак, мы рассмотрели, как семантический компонент ‘шкала’ реализуется в некоторых типах контекстов уж. Как следует поступать дальше – включать ли этот компонент в инвариантное описание уж или в толкование только отдельных значений (лексем) этой частицы – зависит от идеологии описания.

Разумеется, упомянутыми выше семантическими компонентами значение *уж* не исчерпывается. Так, важнейший элемент этой смысла частицы составляет идея объяснения, ссылки на что-то, влияющее на итоговую оценку ситуации говорящим (фраза *Уж очень умный* не может быть сказана просто при описании человека – она уместна, если говорящий объясняет, почему этот человек ему нравится/подходит или, напротив, не годится)[4]. При этом говорящий считает объяснение достаточным, не требующим дальнейшего обсуждения (*Так уж здесь принято*), часто даже блокирующим его; ср. контексты типа *Не знаю уж, как они познакомились…*, где *уж* ясно указывает, что говорящий не имеет в виду обсуждать знакомство, а хочет обсудить что-то, что было потом. С этим компонентом смысла также связаны разнообразные семантические и прагматические эффекты, которые мы не можем проанализировать в небольшой статье.

Пожалуй, общую идею частицы *уж* можно описать примерно так: *уж* указывает на то, что ситуация отклоняется от нормы или ожиданий говорящего/слушающего в такой степени (большой или маленькой), что ссылка на это полностью объясняет итоговую оценку говорящим ситуации. Однако, чтобы подробно показать, как этот смысл взаимодействует с разными типами контекстов, нужна была бы целая книга.

Acknowledgements

Работа выполнена при поддержке РФФИ, грант 19-012-00291.

[4] Этот компонент смысла подробно описан применительно к противительным и уступительным словам.

References

[1] Levontina 2008 — Levontina I. B. Zagadki chasticy uzh [Riddles of the particle uzh] // Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog» (Bekasovo, 4–8 iyunya 2008 g.). Vyp. 7(14). M., 2008.

[2] Levontina 2016 – Levontina I. B. O nekotoryh maloizuchennyh upotrebleniyah chasticy -to [On some poorly studied uses of the particle -to] // Yazyk: poiski, fakty, gipotezy: Sbornik statej k 100-letiyu so dnya rozhdeniya akademika N. Yu. Shvedovoj / Otv. red. M. V. Lyapon. – M.: LEKSRUS, 2016. – 816 s. S. 359-375.

[3] Morkovkin 1997 — Morkovkin V. V. (red.). Slovar' strukturnyh slov russkogo yazyka [Dictionary of structural words of the Russian language]. M., 1997.

[4] Parshin 1988. Parshin P.B. Ustupka i antiustupka v deonticheskom dialoge (funkcionirovanie leksemy hot') [Concession and anti-concession in the deontic dialogue (functioning of the lexeme xot')] // Referenciya i problemy tekstoobrazovaniya. Moskva: Problemnaya gruppa \"Logicheskij analiz yazyka\", 1988. S. 146-168.

[5] Uryson 2007 — Uryson E. V. Uzhe i uzh: variativnost', polisemiya, omonimiya? [Uzhe and uzh: variability, polysemy, homonymy?] // Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog» (Bekasovo, 30 maya — 3 iyunya 2007 g.) M., 2007. S. 531–541.

[6] Shimchuk, Shchur 1999 — Shimchuk E. G., Shchur M. G. Slovar' russkih chastic. [Dictionary of Russian Particles]. Frankfurt am Main, 1999.

[7] König 1991. *König, E.* The Meaning of Focus Particles. A Comparative Perspective. London: Routledge 1991.

[8] Mendoza 1999 — *Mendoza I. Uzhe* und *uzh* in der modernen russischen Standardsprache // Die Welt der Slaven. 1999. No. 44(2). S. 213–224.

[9] Mendoza 2000 — *Mendoza I.* Zur Geschichte von Partikeln: russisch *uzhe* und *uzh* // Linguistik online. 2000. No. 2.

[10] Paillard 1986–1987 — *Paillard D. Už* ou l'indiscutable // Bulletin de linguistique appliquée et générale. 1986/87. No. 13. P. 190–213.

[11] Zybatow 1990 — *Zybatow L.* Was die Partikeln bedeuten. Eine kontrastive Analyse Russisch–Deutsch. München, 1990.

Gender and Case in Russian Nouns Denoting Professions and Social Roles

Magomedova V. D.
independent researcher
varya.magomedova@gmail.com

Slioussar N. A.
NRU HSE, Moscow
SPbU, Saint Petersburg
slioussar@gmail.com

Abstract

In the present paper, we analyzed a group of Russian nouns denoting professions and social roles. Historically, these nouns were masculine; in modern Russian, they can also be used with feminine agreement, but only nominative forms are regarded as normative (e.g. *etot / eta vrač* 'this$_{M/F}$ doctor'). We showed that oblique case feminine forms occur naturally using the Web-as-corpus approach and conducted three experimental studies. We discovered that offline rating and online processing of such forms depends on their case. Firstly, this is a unique example of the properties of the form influencing the properties of the lexeme. Secondly, the fact that all oblique forms are regarded as marginal and that locative was found to be significantly worse than other oblique cases points to a deep connection between grammatical gender and inflectional classes and to the crucial role of affix syncretism in morphological processing. This presents a challenge for different approaches in theoretical morphology.

Keywords: grammatical gender; inflectional class; case; agreement; Russian
DOI: 10.28995/2075-7182-2021-20-483-491

Род и падеж у русских существительных, обозначающих профессии и социальные роли

Магомедова В. Д.
независимый исследователь
varya.magomedova@gmail.com

Слюсарь Н. А.
НИУ ВШЭ, Москва
СПбГУ, Санкт-Петербург
slioussar@gmail.com

Аннотация

В этой статье мы анализируем группу существительных русского языка, обозначающих профессии и социальные роли. Исторически эти существительные относились к мужскому роду. В современном русском языке с ними допустимо согласование и по женскому роду, но только формы именительного падежа считаются нормативными (ср. *этот / эта врач*). Используя интернет в качестве источника примеров, мы показали, что носители русского языка используют также формы косвенных падежей и изучили их в трех экспериментальных исследованиях. Мы установили, что обработка предложений, содержащих такие формы с согласованием по женскому роду, — как оценка их приемлемости (оффлайн-обработка), так и скорость, с которой читаются такие предложения (онлайн-обработка) — зависит от падежа этих форм. Во-первых, это можно рассматривать как уникальный случай, когда грамматические категории словоформы влияют на грамматические категории лексемы. Во-вторых, то, что все формы косвенных падежей оцениваются носителями как маргинальные, но при этом формы предложного падежа оказались наиболее проблемными, указывает на глубинную связь между категорией рода и словоизменительными классами (склонениями) и на ключевую роль синкретизма при морфологической обработке. Эти результаты представляют определенные сложности для различных подходов в рамках теоретической морфологии.

Ключевые слова: грамматический род; склонение; падеж; согласование; русский язык

1 Introduction

This paper analyzes a group of Russian nouns denoting professions and social roles. Historically, these nouns were masculine, but in modern Russian, they can also be used with feminine agreement: e.g. *etot / eta vrač* 'this$_{M/F}$ doctor'. These nouns have several interesting features, and we will focus on one of them: a complex relationship between gender and case features.

According to different sources, feminine agreement is grammatical only in the nominative case (e.g. Graudina et al. 1976; Zaliznjak 2002). Zaliznjak (2002) even suggests representing these nouns as two separate lexemes: a masculine noun and a feminine noun with a defective paradigm, rather than one common gender lexeme. However, oblique case forms with feminine agreement are attested. For example, Sitchinava (2011) notes that "according to Internet data, the phrase *etu vrača* 'this$_{F.ACC.SG}$ doctor$_{ACC.SG}$' is relatively frequent in the modern electronic communication", but does not provide any further details.

In this study, we analyzed naturally occurring oblique case forms with masculine and feminine agreement using the Web-as-corpus approach and conducted three experimental processing studies. Our primary goal was to find out whether the status of feminine forms (their prevalence, their perceived grammaticality, their online processing) depends on their case. Foreshadowing the results, the answer was positive. This is interesting as a unique example of the tail wagging the dog (the properties of the form influencing the properties of the lexeme), but may also have wider implications.

Case hierarchies are introduced in many formal and functional linguistic frameworks, and formal theories also draw a principal distinction between structural and inherent cases. Furthermore, cases differ dramatically in terms of their frequency. For individual case affixes, frequency is also an important property; another crucial property is syncretism. Finding out which of these factors affect production and processing of the relevant feminine noun forms is important for understanding the status of case paradigms and case affixes in the mental grammar and for modelling these phenomena in theoretical morphology.

2 Previous studies

Previous studies dedicated to the nouns denoting professions and social roles focused on agreement in the nominative case. Several experimental studies (Panov 1968; Novikov & Priestly 1999) analyzed the choice of masculine and feminine gender in agreeing verbs and adjectives. They found that semantic agreement is more frequent with verbs than with adjectives. Corbett (2006) incorporated these conclusions in his theory of agreement. A group of studies comes from the field of language acquisition because children acquire semantic agreement relatively late (Dizer 2007; Dobrova 2013; Rodina & Westergaard 2012; Rodina 2014; Tseitlin 2009).

Garnham and Yakovlev (2015) compiled a list of 160 nouns; for every noun, they marked whether it has a corresponding feminitive and how stereotypically female or male the denoted profession or social role is[1] (this factor was found to play a role in the studies on other languages). According to the first parameter, the nouns were divided into those having a normative pair (e.g. *učitel' – učitel'nica* 'teacher'), those having a colloquial pair (e.g. *parikmaxer – ?parikmaxerša* 'hairdresser') and unpaired (e.g. *psixolog* 'psychologist'). Garnham and Yakovlev conducted the first online processing experiment measuring sentence-by-sentence reading times.

Slioussar and Generalova (2018) measured word-by-word reading times in their study. They demonstrated that feminine agreement always triggers reading time delays compared to masculine agreement, but the size of this delay depends on gender stereotypes associated with a given profession or social role. No previous studies looked at the processing of oblique case forms.

3 Corpus study

To assess the frequency of masculine and feminine agreement patterns for different case forms, we conducted a corpus study. We could not use the Russian National Corpus (www.ruscorpora.ru) or simi-

[1] This required a separate experiment. Participants were asked, for example: "You see 100 paediatricians. How many of them do you think are female?". The answer could be given using a scale (0%, 10%, 20% etc.).

lar sources because they mainly contain edited texts, and chose Web-as-corpus approach. We selected 43 unpaired nouns from the list by Garnham and Yakovlev (2015): 42 nouns ending in a consonant, like *psixolog* 'psychologist', and the word *sud'ja* 'judge'. We searched for the combinations of a masculine or feminine agreeing pronoun (*moj* 'my', *naš* 'our' and *etot* 'this') and a target noun in all six cases in singular.

We used the Google search engine and analyzed the raw numbers that it provides. We understand the risks involved, for example, duplicate (and multiplicate) hits that recite one actual phrase. However, our primary goal was to find out whether all case forms are attested and to establish very approximate frequency patterns.

For some stereotypically masculine professions like *švejcar* 'doorman' or *mexanik* 'mechanic', no feminine agreement was attested. For 30 of the 43 selected nouns the sum of all search results for feminine forms did not reach 5000 hits, and the sum of all oblique case forms was below 30 (most often, less than 20 hits). Results for the remaining 13 nouns (the number of hits and percentages for every case) are given in Table 1. Table 2 presents the same 13 nouns with masculine agreement for the sake of comparison.

Noun	**Nom**	**Gen**	**Dat**	**Acc**	**Ins**	**Loc**	**Total**
dizajner 'designer'	11350 (96.4%)	274 (2.3%)	30 (0.3%)	32 (0.3%)	86 (0.7%)	2 (<0.1%)	58484
kosmetolog 'cosmetologist'	7210 (89.9%)	122 (1.5%)	212 (2.6%)	219 (2.7%)	260 (3.3%)	0	8023
fotograf 'photographer'	44600 (99.1%)	265 (0.6%)	62 (0.1%)	43 (0.1%)	43 (0.1%)	0	45013
menedžer 'manager'	11070 (87.2%)	350 (2.8%)	383 (3.0%)	261 (2.0%)	631 (5.0%)	2 (<0.1%)	12697
nevrolog 'neurologist'	11880 (94.4%)	469 (3.7%)	132 (1.1%)	63 (0.5%)	38 (0.3%)	0	12582
pedagog 'pedagogue'	15080 (97.1%)	105 (0.7%)	99 (0.6%)	45 (0.3%)	200 (1.3%)	0	15529
pediatr 'pediatrician'	12600 (89.6%)	452 (3.2%)	330 (2.4%)	363 (2.6%)	316 (2.2%)	1 (<0.1%)	14062
professor 'professor'	8110 (92.5%)	347 (4.0%)	69 (0.8%)	79 (0.9%)	158 (1.8%)	1 (<0.1%)	8764
psixolog 'psychologist'	9430 (80,2%)	278 (2,4%)	220 (1,9%)	1567 (13,3%)	257 (2,2%)	0	14852
stomatolog 'dentist'	14202 (96.2%)	416 (2.8%)	119 (0.8%)	20 (0.1%)	6 (0.1%)	0	14763
vrač 'doctor'	597500 (98.5%)	2876 (0.5%)	2895 (0.5%)	2289 (0.4%)	829 (0.1%)	18 (<0.1%)	606407
xirurg 'surgeon'	4952 (96.2%)	154 (3.0%)	21 (0.4%)	10 (0.2%)	13 (0.2%)	0	5150
sudja 'judge'	14430 (40.0%)	7850 (21.8%)	5609 (15.6%)	5396 (15.0%)	2169 (6.0%)	574 (1.6%)	36028

Table 1: Google search results for target nouns with feminine agreement

On the one hand, it is obvious that the share of nominative forms in Table 1 is dramatically larger than in Table 2. Only the noun *sud'ja* 'judge' that belongs to the 2nd declension (according to the *Rus-*

sian Grammar (Shvedova, ed., 1980)) does not show this tendency. This leads to the conclusion that the problem with oblique forms of other nouns is associated with the system of Russian inflectional classes. We will come back to this observation in the discussion section.

Noun	Nom	Gen+Acc	Dat	Ins	Loc	Total
dizajner ‘designer’	1073900 (43.0%)	1113600 (44.6%)	152470 (6.1%)	151470 (6.1%)	3641 (0.2%)	10395081
kosmetolog ‘cosmetologist’	151840 (45.5%)	84240 (25.2%)	67990 (20.4%)	28440 (8.5%)	1231 (0.4%)	333741
fotograf ‘photographer’	335400 (48.2%)	163720 (23.5%)	101600 (14.6%)	91540 (13.2%)	3317 (0.5%)	695577
menedžer ‘manager’	16647000 (57.3%)	1976600 (6.8%)	4721600 (16.2%)	5707700 (19.6%)	13714 (0.1%)	29066614
nevrolog ‘neurologist’	44380 (52.9%)	19800 (23.6%)	15198 (18.1%)	4158 (5.0%)	322 (0.4%)	83858
pedagog ‘pedagogue’	222400 (36.4%)	191400 (31.3%)	85300 (14.0%)	105100 (17.2%)	6828 (1.1%)	611028
pediatr ‘pediatrician’	221460 (65.1%)	56640 (16.6%)	38071 (11.2%)	23490 (6.9%)	658 (0.2%)	340319
professor ‘professor’	280300 (61.2%)	124500 (27.2%)	24720 (5.4%)	24640 (5.4%)	3633 (0.8%)	457793
psixolog ‘psychologist’	134500 (57.5%)	39860 (17.0%)	20450 (8.8%)	37990 (16.2%)	1216 (0.5%)	234016
stomatolog ‘dentist’	131050 (59.4%)	52460 (23.8%)	25390 (11.5%)	10635 (4.8%)	981 (0.5%)	220516
vrač ‘doctor’	1308000 (48.5%)	750600 (27.8%)	416100 (15.4%)	158000 (5.8%)	66414 (2.5%)	2699114
xirurg ‘surgeon’	116400 (23.5%)	63600 (12.9%)	44800 (9.1%)	18280 (3.7%)	4078 (0.8%)	247158
sudja ‘judge’	102300 (28.3%)	188669 (52.2%)	17250 (4.8%)	38700 (10.7%)	14548 (4.0%)	361467

Table 2: Google search results for target nouns with masculine agreement

On the one hand, it is obvious that the share of nominative forms in Table 1 is dramatically larger than in Table 2. Only the noun *sud’ja* ‘judge’ that belongs to the 2nd declension (according to the *Russian Grammar* (Shvedova, ed., 1980)) does not show this tendency. This leads to the conclusion that the problem with oblique forms of other nouns is associated with the system of Russian inflectional classes. We will come back to this observation in the discussion section.

We can also compare our results to the distribution of cases that Slioussar and Samoilova (2015) calculated for all animate nouns in singular in the syntactically disambiguated subcorpus of the National Russian Corpus: 60.7% nominative forms, 16.6% genitive, 6.2% dative, 8.8% accusative, 6.8% instrumental and 1.0% locative. Nominative is the most frequent, but by far not as frequent as it is in Table 1.

On the other hand, Table 1 shows that feminine agreement is attested in all oblique cases and is not limited to singular examples. Locative forms are underrepresented, but locative case is in general very infrequent with animate nouns. No oblique case appears to be substantially more frequent than the others, so we will turn to experimental studies to explore if there are any differences between them.

4 Experimental study

We conducted three experiments studying how oblique feminine forms are judged offline and processed online.

4.1 Grammaticality judgement experiment

53 native Russian speakers (18 to 55 years old) volunteered to take part in this experiment. They were asked to evaluate sentence grammaticality using the scale from 1 (absolutely ungrammatical) to 5 (perfectly grammatical). The experiment was run on the IbexFarm platform (www.spellout.net).

We selected 15 unpaired nouns that denote stereotypically feminine professions from the list compiled by Garnham and Yakovlev (2015). With each noun, we created five stimulus sentences with five different oblique case forms, as in (1a) or (1b). The nouns were modified by pronouns (*naš* 'our', *etot* 'this' etc.) showing gender agreement. We distributed 75 stimulus sentences across five experimental lists using the Latin square principle, so that every participant sees each noun only once (in one out of five oblique cases). As a result, every list contained 15 stimulus sentences, as well as 30 filler sentences used for distraction.

(1) a. *Ja uznal o svoem diagnoze ot našej vrača.*
I learned about self's diagnosis from our$_{F.GEN.SG}$ doctor$_{GEN.SG}$
'I learned about my diagnosis from our doctor'.

b. *Ja obratilsja s etoj problemoj k našej vraču.*
I appealed with this problem to our$_{F.DAT.SG}$ doctor$_{DAT.SG}$
'I asked our doctor about this problem'.

We found that all oblique forms were judged as equally marginal: genitive received the average rating of 2.0, dative — 2.0, accusative — 1.9, instrumental — 2.0, and locative — 1.8. We used ordinal logistic regression with mixed effects (intercepts) by participant and by item for the statistical analysis, and it did not reveal any significant differences, as expected. These results agree with the corpus data above. However, since oblique feminine forms are infrequent, but most definitely possible, we devised another experiment to zoom in on the potential differences between them.

4.2 Ranging experiment

35 native Russian speakers (19 to 45 years old) volunteered to participate. We selected 30 out of 75 stimulus sentences used in the previous experiment: six sets with six nouns in five oblique cases. Rather than showing participants one sentence from each set, we presented all sentences from one set at once (in a random order) and asked participants to range them from the worst to the best using the 1 to 5 scale. The experiment was run using the PsychoPy software (https://www.psychopy.org).

The data from four participants were discarded because they used only 1s and 5s (all other participants did not always used the whole scale, but at least did not limit themselves to its extremes). After that, we calculated the average ratings: 4.0 for instrumental, 3.4 for accusative, 3.0 for genitive, 2.9 for dative and 1.4 for locative. Using the same statistical methods as in the previous experiment, we found that locative is significantly worse than all other oblique cases (loc vs. acc: β=-4.38, *SE*=0.41, z=-10.69, p<0.01; loc vs. dat: β =-3.34, *SE*=0.30, z=-11.16, p<0.01; loc vs. gen: β=-4.05, *SE*=0.37, z=-11.06, p<0.01; loc vs. ins: β=-3.71, *SE*=0.33, z=-11.22, p<0.01). No other differences were significant. We will come back to these results in the discussion section.

4.3 Self-paced reading experiment

The third experiment was designed to study online processing. 68 native Russian speakers (18 to 55 years old) volunteered to take part in it. We selected 24 unpaired nouns from the list in (Garnham & Yakovlev 2015) and created 48 stimulus sentences like (2a-c) in two experimental conditions: with masculine and with feminine agreement (every noun was used in two sentences). In this experiment, target nouns appeared not only in the oblique cases, but also in nominative. In all sentences, the gender

and case of the target noun were unambiguously signaled by an agreeing adjective and, in some cases, a preposition.

(2) a. *Za stolom sidit molodoj / molodaja bibliotekar' v sinem pidžake.*
at table sits young$_{M.NOM.SG/F.NOM.SG}$ librarian$_{NOM.SG}$ in blue jacket.
'A pretty librarian wearing a blue jacket is sitting at the table'.

b. *Petr uznal ot opytnogo / opytnoj vrača o svoem diagnoze.*
Peter learned from experienced$_{M.GEN.SG/F.GEN.SG}$ doctor$_{GEN.SG}$ about self's diagnosis.
'Peter learned about his diagnosis from an experienced doctor'.

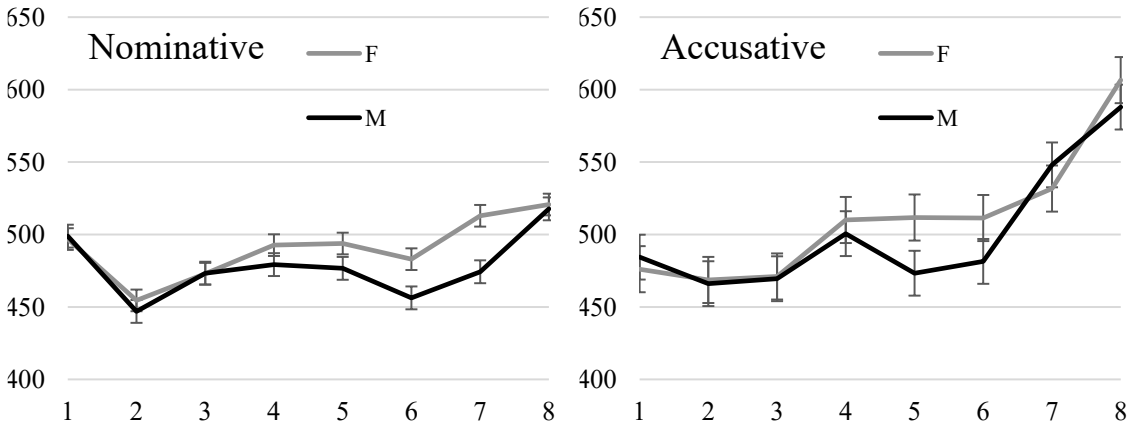
c. *Vanja priglasil populjarnogo / populjarnuju dietologa na večernee šou.*
Vanya invited popular$_{M.ACC.SG/F.ACC.SG}$ dietologist to evening show
'Vanya invited a popular dietologist to the evening show.'

All examples with a particular case had the same syntactic structure. So the target noun was always the fifth word, except for the sentences with accusative case, in which it was the fourth. In all sentences, three words followed the target noun. We created two experimental lists that contained 48 stimulus sentences in one of the two conditions and 108 filler sentences.

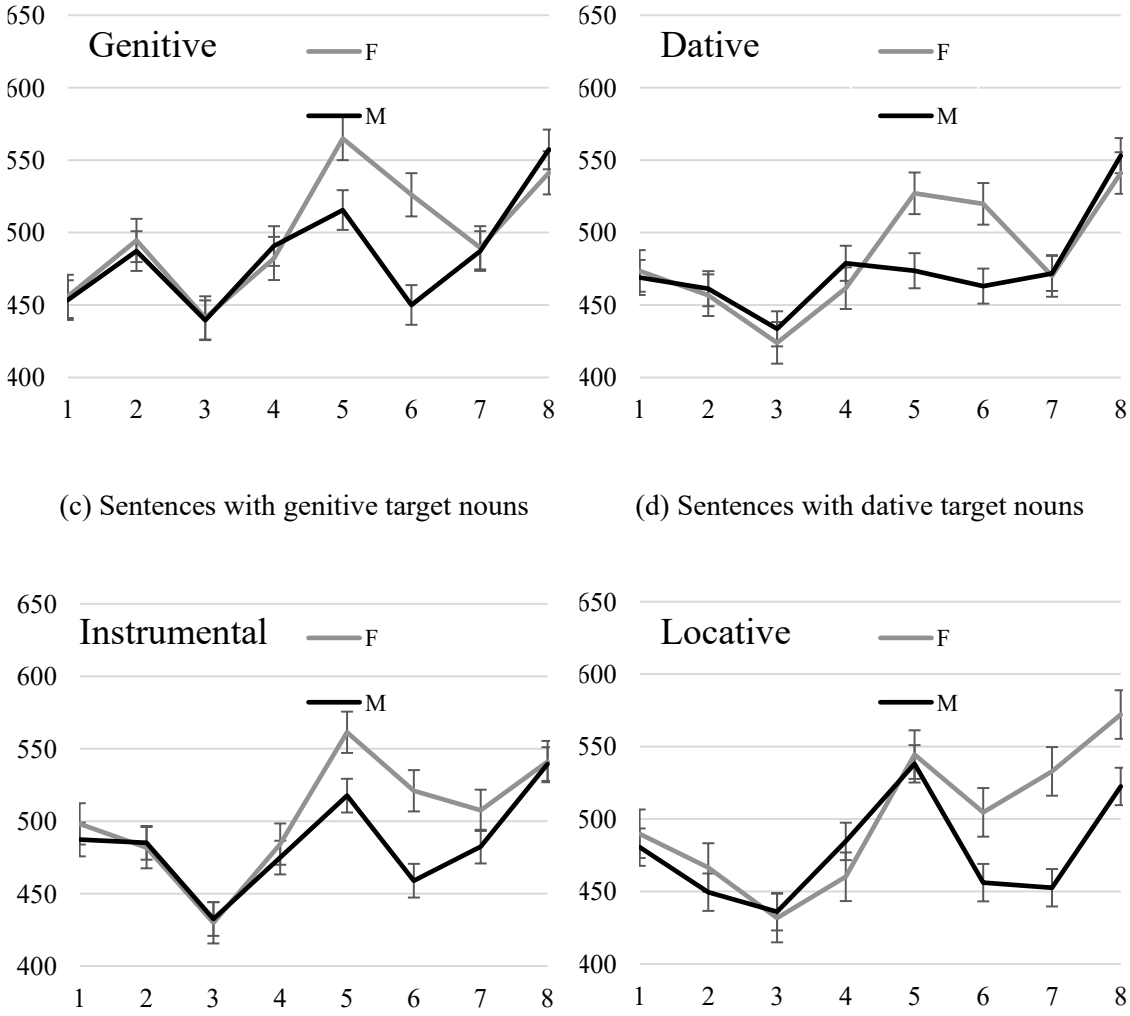
The experiment was run on the IbexFarm platform (www.spellout.net). We used the classic word-by-word self-paced reading methodology. In each trial, a sentence first appeared masked: all letters were replaced by dashes while spaces and punctuation marks remained intact. Participants were asked to press the space bar to reveal a word and re-mask the previous one. As a result, word-by-word reading times could be measured. One third of the sentences were followed by forced choice comprehension questions to ensure that the participants were reading properly.

We analyzed participants' question-answering accuracy and reading times. No participant made more than 3 errors, so no data were discarded based on this parameter. Reading times that exceeded a threshold of 2.5 standard deviations, by region and condition, were excluded (Ratcliff 1993). In total, 3.7% of the data were excluded as outliers.

Average reading times per region (word) in different case groups and experimental conditions are presented in Figure 1. Even in the nominative group, feminine agreement takes more time to process than masculine agreement. This was previously observed by Slioussar and Generalova (2018) who also demonstrated that the size of the delay depends on the stereotypes associated with different professions. Processing of sentences with oblique cases has not been studied before.



(a) Sentences with nominative target nouns (b) Sentences with accusative target nouns



(c) Sentences with genitive target nouns (d) Sentences with dative target nouns

(e) Sentences with instrumental target nouns (f) Sentences with locative target nouns

Figure 1: Average word-by-word reading times in different groups (in ms)

For every case group, we compared word-by-word reading times in the two conditions using linear regressions with mixed effects (intercepts) by participant and by item. On the target noun (N region), the differences in the genitive (β=40.08, *SE*=19.04, z=2.02, p=0.04), dative (β=39.15, *SE*=17.91, z=2.19, p=0.03) and instrumental (β=43.33, *SE*=20.78, z=2.09, p=0.04) groups reached significance. On the following word (N+1 region), there were significant differences in every group (nominative: β=26.43, *SE*=12.55, z=2.11, p=0.04; genitive: β=48.02, *SE*=12.33, z=3.90, p<0.01; dative: β=66.22, *SE*=14.84, z=4.46, p<0.01; accusative: β=37.01, *SE*=15.24, z=2.43, p=0.02; instrumental: β=61.01, *SE*=13.80, z=4.42, p<0.01; locative: β=37.80, *SE*=14.57, z=2.59, p=0.01).

In the N+2 region, the difference between the two conditions was significant only in the locative group (β=67.79, *SE*=11.87, z=5.71, p<0.01). The same was true for the N+3 region (β=49.04, *SE*=17.82, z=2.75, p<0.01), which is the last word of the sentence. In other words, the differences in the sentences with two structural cases, nominative and accusative, reach significance later than in the sentences with non-structural cases. In the locative group, the delay associated with feminine agreement develops later than in the other groups and is more sustained.

5 Discussion and conclusions

Let us summarize the results. The corpus study demonstrated that oblique feminine forms are dramatically less frequent than nominative forms, which definitely cannot be explained by general differences in case frequency. At the same time, all case forms are attested. Only locative is underrepresented, but it is in general the least frequent case in animate nouns. The grammaticality judgment study confirmed that oblique feminine forms are perceived as marginal. However, the ranging experiment that zoomed on the differences between oblique cases and the self-paced reading experiment showed that locative case differs from the others. To explain this result, let us first consider why the words like *vrač* 'doctor' have problems with developing into a common gender noun with a full paradigm.

Russian has many common gender nouns (mostly denoting personal qualities, but also professions and social roles, like *kollega* 'colleague' or *sudja* 'judge') that belong to the 2nd declension ending in *-a/ja* in the nominative singular. Apparently, this is possible because this class historically contains both masculine and feminine nouns, although the former are a minority. The 1st declension with a zero affix in the nominative singular has no feminine nouns. Some feminine nouns like *mat'* 'mother' do have a zero affix in the nominative singular, but they belong to the 3rd declension, in which all oblique case affixes in the singular sub-paradigm are different. We argue that this is the reason why the words like *psyxolog* 'psychologist' are easily used with feminine agreement only in the nominative.

This points to a deep connection between the grammatical gender and declension, which is hard to explain in various morphological theories. For example, in the Distributed Morphology framework inflectional class is a feature stored on a syntactic node (e.g. Kramer 2015). As syntactic trees are parsed successively, either gender may be expected to influence declension or vice versa. In non-structural theories, for example, the Optimality Theory, it is easier to explain how various factors including inflectional classes may influence gender assignment (e.g. Rice 2005). Some non-structural analyses can even predict gender assignment variation (e.g. Doleschal 2000). However, these approaches do not offer an explanation why certain factors play a more important role than the others in a particular case in a particular language.

Now let us come back to locative — why does it differ from other oblique cases? This cannot be explained by case frequency: although locative is the least frequent in animate nouns, differences between other cases would also be expected. Locative is low in different case hierarchies, but instrumental is even lower. Apparently, the only possible explanation is affix syncretism: in other oblique cases in singular, affixes of the 1st declension do not coincide with the 2nd and 3rd declension, but the locative affix *-e* is the same in the 1st and 2nd declension. Prima facie, this could seem advantageous because the 2nd declension contains the majority of feminine nouns. But the effect is the opposite because these nouns have a different paradigm. After finishing the ranging experiment, one of our participants noted that locative seemed the worst to her and added a very telling comment in (3).

(3) *Kak budto eto ne vrač, a kakaja-to vrača*.
as if this not doctor$_{\text{NOM.SG(1st declension)}}$ but some$_{\text{F.NOM.SG}}$ doctor$_{\text{NOM.SG (non-existent 2nd declension noun)}}$

The role of affix syncretism in production and processing was discussed in several experimental studies on different languages, including Russian (e.g. Badecker & Kuminiak 2007; Chernova et al. 2020; Hartsuiker et al. 2003; Slioussar 2018). This question is interesting both for the models of production and processing and for theoretical morphology, in which different approaches to syncretism and to the role of concrete morphemes can be found. For example, in Distributive Morphology relying on the principle of Late Insertion, this role is assumed to be very limited. Our results shed new light on these problems. In particular, in all previous studies, syncretism increased the incidence of errors in production and made them less noticeable in comprehension, and we are the first to get the opposite result.

Acknowledgements

The reported study was partially supported by RFBR and GACR, project number 20-512-26004 (Исследование выполнено при частичной финансовой поддержке РФФИ и ЧНФ в рамках научного проекта № 20-512-26004).

References

[1] Badecker W., Kuminiak F. Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak // Journal of memory and language. – 2007. – Vol. 56. – Pp. 65–85.

[2] Chernova D., Slioussar N., Alekseeva S. Orthographic processing of case forms of Russian nouns in a sentence [Osobennosti orfograficheskoj obrabotki padezhnyx form russkix sushchestvitel'nyx v kontekste predlozhenija] // Tomsk State University Bulletin – 2020. – Vol. 454. – Pp. 45–54.

[3] Corbett G. Agreement. – Cambridge University Press, Cambridge, 2006.

[4] Dizer E. Children mastering the category of gender in the situation bilingualism and trilingualism [Osvoenie kategorii roda v ramkax detskogo dvu- i trex'jazychija] // Semantic categories in child speech [Semanticheskie kategorii v detskoj rechi]. – Nestor-Istorija, Saint Petersburg, 2007. – Pp. 244–265.

[5] Dobrova G. R. Overcoming the contradiction between the semantics and the form of animated personal nouns in the process of assimilating the category of gender by children [Preodolenie protivorechija mezhdu semantikoj i formoj odushevljonnyx lichnyx sushchestvitel'nyx v protsesse usvoenija det'mi kategorii roda] // Verb and noun categories in functional grammar [Glagol'nye i imennye kategorii v sisteme funktsional'noj grammatiki]. – Nestor-Istorija, Saint Petersburg, 2013. – P. 64.

[6] Doleschal U. Gender assignment revisited // Gender in Grammar and Cognition. Part I: Approaches to Gender. – Mouton de Gruyter, Berlin, 2000. – Pp. 117–166.

[7] Garnham A., Yakovlev Y. The interaction of morphological and stereotypical gender information in Russian // Frontiers in Psychology. – 2015. – Vol. 6. – Art. 1720. – Pp. 1–12.

[8] Graudina L.K., Itskovich V. A., Katlinskaja L. P. Grammatical correctness of Russian speech [Grammaticheskaja pravil'nost' russkoj rechi]. – Nauka, Moscow, 1976.

[9] Hartsuiker R. J., Schriefers H. J., Bock K., Kikstra G. M. Morphophonological influences on the construction of subject-verb agreement. – Memory and cognition. – 2003. – Vol. 31. – Pp. 1316–1326.

[10] Kramer R.T. The morphosyntax of gender. – Oxford University Press, Oxford, 2015.

[11] Novikov Y., Priestly T. Gender differentiation in personal and professional titles in Contemporary Russian // Journal of Slavic Linguistics. – 1999. – Vol. 7. – Pp. 247–263.

[12] Panov M.V. Russian language and Soviet society: Morphology and syntax of the modern Russian literary language [Russkij jazyk i sovetskoe obshchestvo: Morfologija i sintaksis sovremennogo russkogo literaturnogo jazyka] – Nauka, Moscow, 1968.

[13] Rice C. Optimizing Russian gender: A preliminary analysis // Formal Approaches to Slavic Linguistics 13. – Michigan Slavic Publications, Ann Arbor, 2005. – Pp. 265–275.

[14] Rodina Y., Westergaard M. A cue-based approach to the acquisition of grammatical gender in Russian // Journal of Child Language. – 2012. – Vol. 39. – Pp. 1077–1106.

[15] Rodina Y. Variation in the input: child and caregiver in the acquisition of grammatical gender in Russian // Language Science. – 2014. – Vol. 43. – Pp. 116–132.

[16] Shvedova N. (ed.) Russian grammar [Russkaja grammatika]. – Vol. 2. – Nauka, Moscow, 1980.

[17] Sichinava D. V. Gender. Materials for the project of the corpus description of Russian grammar [Rod. Materialy dlja proekta korpusnogo opisanija russkoj grammatiki]. – Access mode: http://rusgram.ru – Manuscript, 2011.

[18] Slioussar N., Generalova A. Grammatical characteristics and gender stereotypes in the processing of gender agreement in Russian [Grammaticheskie xarakteristiki i gendernye stereotipy pri obrabotke soglasovanija po rodu v russkom jazyke] // The 8th International Conference on Cognitive Science: Abstracts. – Institut psixologii RAN, Moskva, 2018. – Pp. 933–935.

[19] Slioussar, N., Samojlova M. Frequencies of different grammatical features and inflectional affixes in Russian nouns [Chastotnosti razlichnyx grammaticheskix xarakteristik i okonchanij u suschestvitel'nyx russkogo jazyka] // Proceedings of the conference 'Dialogue'. – Access mode: http://www.dialog-21.ru/digests/dialog2015/materials/pdf/SlioussarNASamoilovaMV.pdf.

[20] Slioussar N. Forms and features: The role of syncretism in number agreement attraction // Journal of Memory and Language. – 2018. – Vol. 101. – Pp. 51–63.

[21] Tseitlin S. On word formation in child speech [Ocherki po slovoobrazovaniju v detskoj rechi]. – Znak, Moscow, 2000.

[22] Zaliznjak A. A. Russian nominal inflection, and selected works on modern Russian language and general linguistics [Russkoe imennoe slovoizmenenie s prilozheniem izbrannyx rabot po sovremennomu russkomu jazyku i obshchemu jazykoznaniju]. – Jazyki slavjanskoj kul'tury, Moscow, 2002.

Morphological annotation of social media corpora with reference to its reliability for linguistic research

Mariia Michurina
ABBYY Lab, MIPT,
Dolgoprudny, Russia
RSUH, Moscow, Russia
marimitchurina@gmail.com

Alexandra Ivoylova
ABBYY Lab, MIPT,
Dolgoprudny, Russia
RSUH, Moscow, Russia
aleksandra.ivoilova@abbyy.com

Nikolay Kopylov
ABBYY Lab, MIPT,
Dolgoprudny, Russia
nikolay.kopylov@abbyy.com

Daniil Selegey
ABBYY Lab, MIPT,
Dolgoprudny, Russia
daniil.selegey@abbyy.com

Abstract

This paper presents the results of the study devoted to the applicability of SOTA methods for morphological corpus annotation (based on GramEval2020) for analytical sociolinguistic research. The study shows that statistically successful technologies of morphosyntactic annotation for such purposes create a number of problems for researchers if they are used purely i.e. without any linguistic knowledge. In this paper, methods for improving the morphological annotation, successfully implemented in GICR, from the point of view of its reliability are presented.

Keywords: automatic morphotagging, morphosyntactic annotation, lemmatization, NLP evaluation, morpho-parsers for Russian, language of social media

DOI: 10.28995/2075-7182-2021-20-492-504

Морфоразметка корпуса текстов из социальных сетей с точки зрения надежности лингвистических исследований

Мария Мичурина
ABBYY Lab, МФТИ,
Долгопрудный, Россия
РГГУ, Москва, Россия
marimitchurina@gmail.com

Александра Ивойлова
ABBYY Lab, МФТИ,
Долгопрудный, Россия
РГГУ, Москва, Россия
aleksandra.ivoilova@abbyy.com

Николай Копылов
ABBYY Lab, МФТИ,
Долгопрудный, Россия
nikolay.kopylov@abbyy.com

Даниил Селегей
ABBYY Lab, МФТИ,
Долгопрудный, Россия
daniil.selegey@abbyy.com

Аннотация

В работе приводятся результаты проведенного исследования по применимости SOTA-методов морфоразметки русскоязычных корпусов (по данным GramEval2020) для аналитических социолингвистических исследований. Показано, что механическое применение статистически успешных технологий разметки для таких целей порождает ряд проблем для исследователя - теоретического лингвиста. Приводятся методы улучшения разметки с точки зрения надежности получаемых результатов, успешно примененные при создании новой версии ГИКРЯ.

Ключевые слова: автоматическая морфоразметка, морфосинтаксический анализ, лемматизация, оценка систем автоматической обработки текста, морфопарсеры для русского языка, язык социальных медиа

1 Introduction

In modern linguistic research, the so-called mega corpora [3], or extra-large corpora [5], created according to the Web as Corpora (WAC) technology and containing billions of words, are widely used. It is quite obvious that manual annotation in such corpora is an unbearable task for a linguist. Thus, the only option is automatic annotation.

This paper examines the quality of automatic morphosyntactic annotation of mega corpora for sociolinguistic studies of Russian, processed by SOTA methods, which participated in the GramEval2020 competition [10]. The integral morphosyntactic parser for Russian[1] [1] was selected for our evaluation as it had achieved best results in the evaluation (hereinafter referred to as IMParser). The research was carried out within the framework of the new version of the General Internet Corpus of Russian (GICR) [2]. The GICR is one of the four existing mega corpora of Russian (the other three are ruTenTen [8], Aranea [4] and Taiga [14]. Unlike ruTenTen and Aranea, GICR is a differentiated corpus, i.e., divided into segments depending on the source of the texts. From our point of view, it is the advantage of GICR as it allows us to test the IMParser on texts from different segments of Russian social networks that may be hard for parsers. Taiga, on the other hand, is a corpus designed for computational linguists and NLP-specialists, not linguistic researchers [14].

Thus, the work evaluates the progress in the field of automatic corpus annotation over the past few years: the TnT parser [6], [13], used in the first version of the GICR, is a typical representative of statistical automatic parsers (for example, Aranea was annotated with the Tree Tagger [16], and ruTenTen with Tree Tagger and RFTagger [17]; both mentioned parsers, just like TnT, use hidden Markov models and, therefore, the quality of their annotation does not differ much. IMParser is, in a sense, a typical representative of the new generation of parsers. Consequently, on the one hand, they are the standard representatives of the parsers of their generation and allow us to assess the progress of text processing methods in general. On the other hand, both have been used in the GICR, and it is important for us to evaluate the improvement in the annotation quality.

There are a number of morphological parsers that use different formats and quite a few of them have their own tagsets (e.g., SynTagRus, OpenCorpora, RNC, MSD-GICR, MULTEXT-East, etc.). However, there is a Universal Dependencies (UD) project [11], which annotation guidelines unite more and more languages and corpora. Its use seems quite promising to us. It is the UD annotation scheme that IMParser uses.

According to the purpose of this study, we were faced with the following tasks:

- Evaluate the work of the parser in relation to various phenomena that should be of interest to users of such corpora as GICR; determine the benefits of integrating morphosyntactic annotation;
- Propose a new pipeline for corpus annotation, which gives a satisfactory final result from the point of view of a linguistic researcher, including adjusting the work of the parser;
- Assess the applicability of UD as a corpus annotation scheme for linguistic and sociolinguistic studies of the Russian language.

The GICR is intended primarily for theoretical linguistic research. Therefore, the quality of lemmatization, PoS-labelling, and disambiguation is important here. In this regard, in our work we carried out not only a numerical assessment, comparing the percentage of parsing accuracy, but also a manual quality assessment. Moreover, not only the quality of data annotation processing is important, but also its speed (the standard sizes of mega corpora force their developers to pay attention to it). Tests on GramEval data have shown that solutions based on pretrained BERT models are slightly better than fine-tuned ELMo ones, but they are much slower.

Below in this article, the results of solving the listed problems will be considered in detail: in the second paragraph, we will talk about the work of IMParser on the GICR data, in the third, the adaptation of the UD scheme for tagging the GICR will be discussed.

[1] https://github.com/DanAnastasyev/GramEval2020

2 The IMParser and its evaluation

The overall accuracy for five genres of the modern Russian language ("news, social media and electronic communication, wiki-texts, fiction, poetry; Middle Russian texts are used as the sixth test set" [10]) of the IMParser is the following: 0.916 versus the baseline accuracy of 0.804 (rnnmorph for lemmatization and morphology and UDPipe for syntax).

The IMParser is a combination of three interacting elements. Firstly, this is a fairly simple classifier that predicts the lemmatization rule for a word form, secondly, it is a morphological parser based on embeddings, and thirdly, a dependency parser. For morphological analysis, the model can use different versions of embeddings (character-level embeddings as well as two variants of contextual embeddings that have already proven themselves in NLP: BERT and ELMo); moreover, the parser uses grammeme embeddings that contain information about the grammatical meanings of a word and give the parser information about the interaction of this word with others. Models with contextual embeddings, especially BERT model, have shown the best quality. Dependency parser uses Edmonds' algorithm for finding minimum spanning trees on directed graphs for decoding; it produces syntactic parsing within the UD guidelines. All the three elements interact with each other: "The latter model uses shared representations between the morphological parser, the lemmatizer and the dependency parser" [1]. Thus, the parser simultaneously processes lemmatization, morphological tagging and syntactic parsing. A similar approach to automatic data annotation has already been used before, for example, in ETAP-4 [7], and the GramEval2020 rules were based on the decision that morphology and syntax should be analyzed simultaneously and be related.

2.1 Quality evaluation methods of automatic annotation

One of the main tasks of this study is manually evaluating the quality of automatic annotation of IMParser, which is of particular interest because of its multitask approach.

During the GramEval2020 competition, quality testing was carried out. It was aimed at cross-system comparison: the organizers automatically compared the manual annotation (gold set, inaccessible to the participants of the competition) and parser annotations, and published the average score, paying special attention to errors common to all systems, which directly follows from the objectives of the competition. We are interested in meaningful analysis, including analysis of particular ambiguous units. We want to understand to what extent the integral parsing methods are applicable for labelling corpora, which are intended not for NLP tasks, but for studies of the language. General quality metrics are important, but some types of errors may be unacceptable for language research.

As part of our research, we manually compared the quality of the IMParser to the end-to-end morphological analysis of the TnT parser for the Russian language. The main concern is both lemmatization and PoS-tagging. A special quality evaluation of IMParser's verb and noun lemmatization was carried out due to cases with a complex paradigm in these parts of speech. For this experiment, 10,000 tokens of random sentences were taken from VKontakte segment of the GICR. We also focused on the quality of lemmatization of out-of-vocabulary (OOV) words: lexemes that are absent from both standard dictionaries and training data. Social network texts usually possess newly created lexemes [15] and therefore a morphosyntactic parser has to cope well enough with such things. Finally, to test the claim that integrated morphosyntactic parsing improves the quality of disambiguation, some experiments with full and PoS homonyms were carried out.

2.2 Quality evaluation results

An experiment comparing the annotation of the TnT-parser and the IMParser gave the following results: the TnT-parser is not good enough in disambiguation, like other parsers in its category. However, IMParser has serious problems with lemmatization of non-homonymous word forms, namely verbs (see Fig. 1). This, apparently, is due to the fact that the parser does not use a dictionary for lemmatization. Non-dictionary approach should give an advantage that is important for processing texts of social networks: for the parser there is no fundamental difference between dictionary and OOV words. However, the presence of hallucinations, a serious negative consequence, was discovered (by hallucinations we mean cases when the parser generates lemmas that do not exist in the language). For this analysis, the ELMo model (trainable ELMo LSTM) was used as a compromise in terms of the quality and the parsing speed.

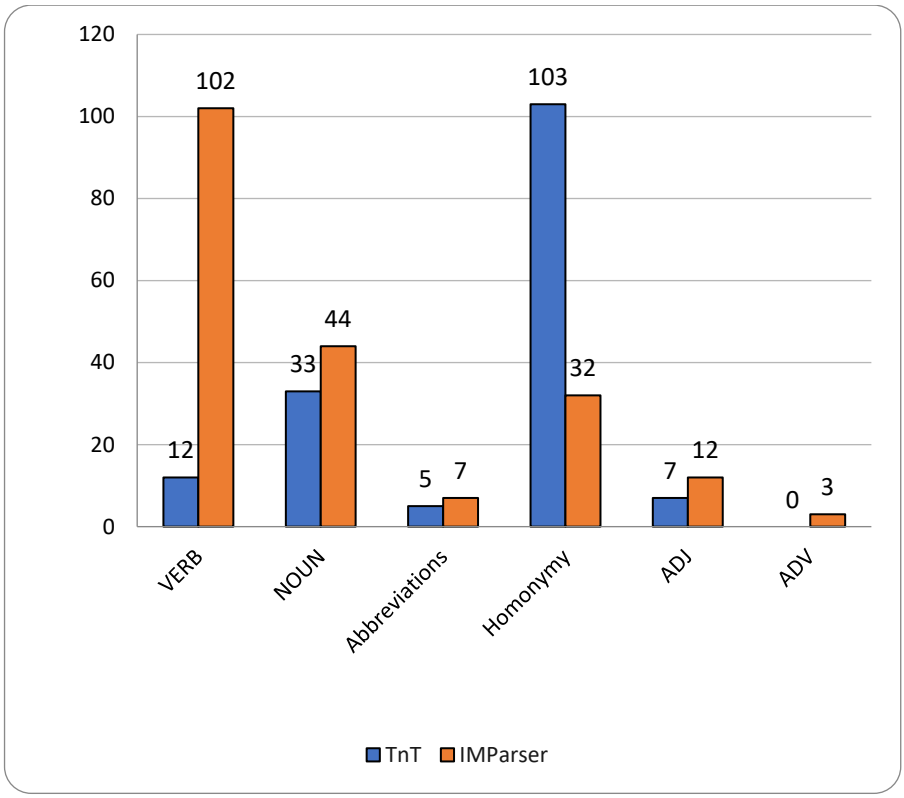


Fig. 1. Comparison of lemmatization errors per 10,000 tokens

In contrast to lemmatization, the IMParser copes with PoS-tagging better than the TnT parser (see Fig. 2). This is especially noticeable in nouns and adverbs tagging.

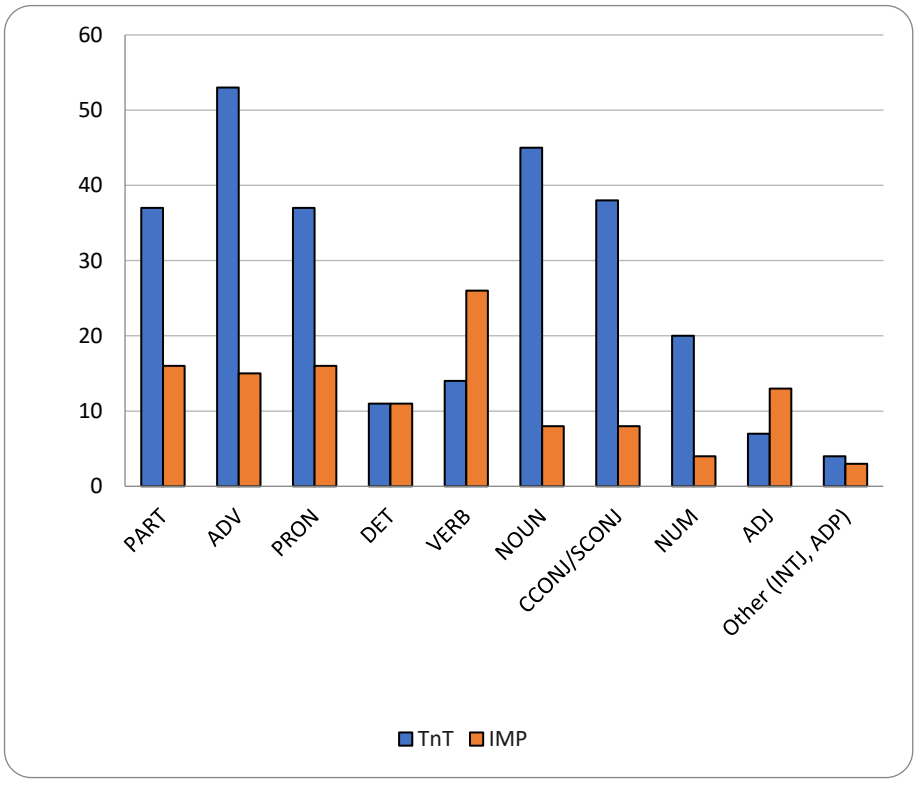


Fig. 2. PoS-tagging errors (tags show PoS that should be given instead of the wrong ones)

The quality of the IMParser for most of the other parameters (including parsing speed, PoS and feature tagging) turned out to be good, and this made it possible to use the parser for automatic parsing of the GICR.

2.3 Lemmatization problem

Since the quality of lemmatization shown by the IMParser was noticeably worse than the one performed by the TnT-parser, it was decided to concentrate on lemmatization tests, including the "problematic" parts of speech, i.e. verbs and nouns. To assess the quality of lemmatization, a dataset of 10 000 tokens was parsed, using BERT and ELMo models. As far as the data source, VKontakte as the "dirtiest" and the most difficult segment to parse was taken. All word forms with errors (typos, spelling errors) were excluded since GramEval did not presuppose spelling corrections. Lemmatization of nouns and verbs was assessed manually. For the results, see table 1.

	Verbs	Nouns
BERT	5.1	3.4
ELMo	7	5

Table 1. Lemmatization error rate, verbs and nouns, per 10,000 tokens

The results of quality evaluation of the IMParser revealed a serious problem with lemmatization: even when using the BERT model, which showed the best results in the competition, hallucination errors were found. Hallucination errors (see table 2) play a particularly important role because from a human point of view they are difficult to explain and predict comparing to disambiguation errors. Therefore, such errors may lead to users' mistrust in corpus annotation or incorrect data and statistics in studies if a linguist gives the annotation too much credence. We have also found that both models systematically miscalculate lemmas for word forms in uppercase ("ПЯТЫЙ" is defined as "пяты", and "РЕБЯТА" as "ребят"), although the parser was trained, among other things, on social network data where the uppercase is quite common. However, there are only a few errors in disambiguation.

Wordform	**Right lemma**	**BERT**	**ELMo**
потерь	потеря	потеь	потерьа
подсел	подсесть	подйти	подсеть
льдах	лед	льер	льд
прилечу	прилететь	прилестить	прилечуть
пою	петь	повать	поть
берите	брать	беыть	берить
бегите	бежать	бяться	бегять
шипящими	шипеть	шипить	шипть
стань	стать	стань	стть
зажгли	зажечь	зжечь	зажгть

Table 2. Examples of hallucination errors

If we compare the quality figures from the competition, parser results seem quite good (dev / test sets in lemmatization: 98.3% / 95.8% ELMo and 98.5% / 96.4% BERT, respectively). Nonetheless, the evaluation of the two most important parts of speech showed that it is impossible to use the results of the lemmatization in the corpus without a dictionary check. Perhaps a difference between the competition numbers and our noun-verb experiment arises from the fact that GramEval evaluation script took into account unchangeable word forms when lemmas always correspond to the original word (e.g. conjunctions, particles and prepositions), therefore, improving the lemmatization quality rate.

The IMParser doesn't use dictionary-based lemmatization as a matter of principle, moreover, it is pointed out that its approach to lemmatization, i.e. the compilation of rules for modifying word forms according to the training corpus (less than 1,000 classes of rules in total) and the application of these rules for test data lemmatization "is less likely to hallucinate an invalid lemma than in the sequence-to-sequence approach" [1]. Table 3 presents statistics on lemmas with hallucination errors for the same sample of nouns and verbs in 10,000 random tokens.

	Verbs	Nouns
BERT	49.3	61.67
ELMo	44.16	87.06

Table 3. Hallucination error rate to overall lemmatization error rate, verbs and nouns, per 10,000 tokens

2.4 Difficult disambiguation based on syntax

In order to assess how joint processing of morphology and syntax affects the quality of parsing (in particular, lemmatization), an experiment was carried out with full homonyms "пла́чу" and "плачу́", "сто́ит" and "стои́т".

In addition to these pairs, there was an attempt to experiment with the word form "лечу" (lemma "лететь"), but it turned out that there was no such word form in the training set at all, and the parser gave either "лечить" or hallucination errors in all the cases. This is a serious problem because the error found affects the language core (the verb "лететь" is part of the basic vocabulary and is more frequent than "лечить") and can provoke users' distrust of the corpus, while we strive to raise the level of confidence of linguistic users in the web corpus. Such errors should be excluded, for example, through the use of a dictionary. We think that an experiment with a large number of frequently used Russian verbs is needed in order to objectively assess the scale of the problem.

The form and grammatical features of "плачу́" and "пла́чу" verbs match fully, but "пла́чу" is an intransitive verb and cannot have a direct object, whereas the verb "плачу́" is a transitive one. Moreover, supposedly only the verb "плачу́" can have an argument with the preposition "за" (e.g., "плачу за обучение"). We assume that if the analysis of morphological and syntactic characteristics as well as lemmatization is processed simultaneously, the verb "плачу́" with a direct object or with a noun phrase with the preposition "за" is more likely to be lemmatized as "платить" than cases of "плачу́" without a direct object.

For the experiment, 221 sentences were selected from VKontakte and LiveJournal segments with the lemma "платить", 120 of them contain the word form "плачу́" with a direct object (DOBJ), and the remaining 101 without a direct object. In addition, in 98 sentences out of total 221 the verb "плачу́" has an argument with the preposition "за". The analysis was carried out both with BERT and ELMo models. The results are presented in table 4.

	BERT	**ELMO**
Average score (платить)	25.3	26.7
Платить + DOBJ	34.2	35
Платить – DOBJ	15	17
Платить + “за”	24.5	25.5
Платить – “за”	26	27.6

Table 4. Percentage of right lemmas of the verb “плачу́”

Having analyzed the statistics obtained, we can conclude that syntax in cases with direct object truly contributes to correct disambiguation. However, the preposition “за” does not affect the parsing. Moreover, as the anonymous reviewer rightly pointed out, the preposition “за” can actually occur in combination with the verb “плакать”:

*Каждый вечер **плачу за** тобой. Вернись быстрее ты домой.*
*Каждый вечер слушаю эту музыку и **плачу за** ним!*
*мне больно, и я **плачу за** тех кто живёт на донбассе.*

It was found that the parser is much more likely to lemmatize the verb as “плакать”, possibly due to a skew in the training dataset, but the statistics on the training set is as follows:

Lemma	**Total amount in train data**	**Word form**	**Amount of word forms in train data**
плакать	65	плачу (лемма: плакать)	4
платить	151	плачу (лемма: платить)	2

That is, although the number of the “плачу” option with the “плакать” lemma is formally twice as large, the absolute numbers are too small.

A similar case is represented by the words “сто́ит” and “стои́т”: in this form they differ only in stress, but for the first variant the lemma will be “стоить”, and for the second “стоять”. Also, the verb “стоить” is transitive, but the verb “стоять” is not. The following experiment was based on this difference.

For the experiment, 213 sentences were selected from the VKontakte segment with the “стоить” lemma, in 113 of them the word form “stand” with a direct object (DOBJ) occurs, in the remaining 100 the verb goes without a direct object. The results are presented in table 5.

	BERT	**ELMO**
Average score (стоить)	83.5	66.6
Стоить + DOBJ	89.3	78.7
Стоить – DOBJ	77	53

Table 5. Percentage of right lemmas of the verb “сто́ит”

In this case, it is obvious that the number of correct lemmas is higher for the direct object verb.

To sum up, we can say that syntax improves the quality of disambiguation, and it is worth noting that the cases selected for experiments are quite rare and complex.

2.5 PoS and grammatical disambiguation based on syntax

Homonymy, including PoS and grammatical one, often causes errors in automatic morphological parsing. It makes researchers pay special attention to this issue in studies related to automatic morphological labeling. According to the developers of GICR 1.0, the quality of the disambiguation of 'complicated' cases was 90% for adjectives and 68% for nominalized adjectives (nouns), as well as one of the worst indicators – 66% for accusative of animate nouns [12].

The purpose of this experiment was to test how well two models we are considering will distinguish between adjectives and substantives, as well as the coinciding forms of nouns in the nominative, genitive and accusative.

For the first experiment, the most frequent nouns derived from adjectives with no morphological transformation were extracted from "A New Frequency Dictionary of Russian" [9] (some words denoting abstract concepts were excluded, e.g., "основное", "главное", "целое"). In total, four words were selected:

- прошлое
- ученый
- русский
- больной

We used VKontakte and LiveJournal segments as a data source. 200 sentences were selected for each pair of words (100 sentences for a noun, 100 sentences for an adjective). These sentences were labelled manually according to the experiment task in such a way that if a word does not have a nominal head, then the word form gets the "noun" tag. Cases with a paired structure, where the ellipsis of the nominal head is obvious, were annotated as adjectives. Complicated cases with homonymy were tagged according to the semantics of a construction. The results of this experiment are presented in table 6.

	BERT	**ELMo**
Прошлое (NOUN)	95	99
Прошлый (ADJ)	97	95
Ученый (NOUN)	99	99
Ученый (ADJ)	88	80
Русский (NOUN)	85	78
Русский (ADJ)	100	96
Больной (NOUN)	97	94
Больной (ADJ)	77	63
Mean NOUN	94	92.5
Mean ADJ	90.5	83.5

Table 6. Adjectives and nominalized adjectives (nouns)

The quality of disambiguation in these cases are obviously higher than that of the TnT-parser (the accuracy of the latter is 68% for substantives on average, while IMParser shows 83% accuracy). Thus, we may expect the improvement in quality of such cases.

For the second experiment the following words were selected from the frequency dictionary with coinciding word forms in nominative and accusative:

- время
- дело
- жизнь
- слово
- место

The volume of the selected data, as in the previous experiment, was 100 sentences for each noun in the nominative and 100 sentences for each noun in the accusative; all data was reviewed and tagged manually. See the results in table 7.

Lexeme	**BERT**	**ELMo**
дело, Nom	100	100
дело, Acc	96	95
время, Nom	99	97
время, Acc	99	99
место, Nom	96	94
место, Acc	99	97
слово, Nom	95	97
слово, Acc	96	89
жизнь, Nom	98	96
жизнь, Acc	97	99
Mean, Nom	97.6	96.8
Mean, Acc	97.4	95.8

Table 7. Percentage of correct features, Nom and Acc

Finally, lexemes with matching accusative and genitive forms were selected:

- бог
- ребенок
- человек
- друг
- отец

The results of the experiment with the same volume of data are shown in table 8.

Lexeme	BERT	ELMo
бог, Acc	89	86
бог, Gen	97	98
ребенок, Acc	97	95
ребенок, Gen	99	100
человек, Acc	97	91
человек, Gen	100	99
друг, Acc	95	97
друг, Gen	98	98
отец, Acc	94	93
отец, Gen	100	100
Mean, Acc	94.4	92.4
Mean, Gen	98.8	99

Table 8. Percentage of correct features, Acc and Gen

Thus, we see that the correct scores for IMParser do not fall below 86 for a particular lexeme.

2.6 Lemmatization and PoS-tagging quality of out-of-vocabulary words

The aim of the next experiment was to check how well the models selected for the study deal with lemmatization and PoS-tagging of OOV words (neologisms, nonce words, slang, borrowings, etc.). For the experiment, 468 random lexemes with 639 occurrences were selected that were found neither in the Compreno dictionary nor in the training data, word forms with typos were removed. The results are presented in table 9.

	BERT	ELMo
Lemmatization	85.26	80.34
PoS-tagging	91.88	89.1
Different lemmas of the same lexemes (based on lemmatization of lexemes with several occurrences)	17.58	18.68

Table 9. Percentage of right lemmas of out-of-vocabulary occurrences

BERT and ELMo commit up to 15% and 20% of lemmatization errors respectively; but we should bear in mind that these are “complex”, non-dictionary words that are not found in the training corpus. It is difficult to establish correct lemmas for some of them (e.g., is the lexeme “друзьяшки” plurale tantum?). Unfortunately, there is no way to correct the lemmatization of neologisms and slang, so this percentage of accuracy is final.

2.7 Lemma corrections and grammatical system adaptation. Correction results

Since we came to the conclusion that the lemmatization quality of the selected models is not high enough, it was decided to improve the results. The possible way of correction is to use a dictionary. This hybrid method is not innovative, as well as not the only one available, but it has its advantages: firstly, it is a high speed of work, and secondly, predictability and verifiability of results.

Thus, we decided to use Compreno dictionary that contains more than 200,000 lemmas and more than 6.6 million word forms with PoS-tags and grammatical features. This dictionary was used to correct the morphology of GICR 1.0 (with the TnT parser) and helped to significantly improve the accuracy of disambiguation (by 30% for some categories of nouns).

We converted the dictionary to UD format; then word forms, PoS and grammatical features were checked: if all the three parameters matched but the lemma was different, then the lemma was replaced with the correct one taken from the dictionary.

	Verbs	**Nouns**
BERT	5.1	3.4
BERT (with Compreno)	0.7	2
ELMo	7	5
ELMo (with Compreno)	1.1	2.5

Table 10. Lemmatization error rate, verbs and nouns,
before and after using Compreno dictionary, per 10,000 tokens

This decision significantly improved the quality of lemmatization. It shows that a high-quality corpus should still be based on a dictionary with the inflection model; however, certain problems are still there:

- Homonyms with completely identical grammatical features cannot be resolved (e.g., “честный” and “честной” in oblique cases, “небо” and “нёбо” because of the letter “ё” as “е” is often replaced by “ё”);
- If the parser gave a false PoS or grammatical tag, then the lemma either will not be corrected, or it may be changed to a wrong one.

Although certain results have already been achieved in correcting lemmatization, some work still needs to be done. Moreover, there are no simple solutions to the above-mentioned problems. The importance of correct lemmas is obvious: corpus user should not think how to avoid errors of automatic lemmatization and compose corpus queries with disjunction of word forms but can safely use the lemma search.

3 GICR annotation within UD guidelines

The UD framework is an actively developing project with one of the best annotation formats. There were a lot of discussions around it and changes to it continue to be made today. We do not claim to change the UD Russian tagset as a whole, but we would like to slightly adjust the tagset that will be used in the GICR to simplify it for theoretical linguists who, unlike computational linguists who are actively using UD treebanks, have no experience with this tagset. At present, the situation with the UD format for the GICR is as follows.

The following changes have already been made:

- The PROPN tag, which denotes a proper name, has been replaced with NOUN (to avoid ambiguities in words like “Президент”, “Дед” and other non-proper names written with a capital letter);
- the particle “бы/б” would not be labelled as AUX, because it will probably seem strange to linguists dealing with the Russian language: the tag for this particle would be PART

To do:

- Although according to the terms of GramEval 2020 and the UD guidelines “Pass” tag (passive voice) should only be related to participles, 13% of “-ся/сь” verbs in standard training UD corpora of Russian possess the “Pass” tag (and not the “Middle” tag). It is due to the large amount of training data that is very difficult to verify manually. Because of this, the parser gets additional errors.
- A single tag is required for foreign words. At the moment, the annotation of foreign words is carried out in an ambiguous way: foreign words representing the names of large companies, cities, etc., are labelled as PROPN or NOUN, and all other foreign words are marked as X. The difference between NOUN and X is too subjective and therefore it is better to unify it in some way.

Furthermore, we have identified some features that may cause difficulties for linguists working with the Russian language. In particular, the following ones:

- transitivity / intransitivity of verbs in not labelled;
- there is no separate tag for “предикатив”;
- no tags for Plurale / Singulare Tantum in the category “Number”;
- Plurale Tantum nouns may have a genus, but the source of such information is unknown. (It is not discussed within Russian UD guidelines, but the training corpora contain this annotation. Thus, the parser also assigns gender to Plurale Tantum words).

The UD format has many advantages, including universality and readability. It covers more and more languages and treebanks. We tried to simplify its use for theoretical linguists a little in the GICR corpus without changing the general concept.

4 Conclusion

As a result of this study, the following conclusions were obtained:

- The importance of analyzing automatic markup errors from the standpoint of theoretical linguistics has been shown, including analysis of the annotation scheme;
- A number of problems have been identified that prevent the use of automatic markup for the needs of linguistic researchers without adjustments;
- The need for vocabulary support at some stage of the pipeline has been proven.

We believe that not only the national corpus, but also the web corpus should be treated as a serious source, so it should include:

A) a clear, human-readable annotation format, which the UD format successfully handles;
B) vocabulary support to ensure correct analysis of at least the language core;
C) a thorough analysis of the entire pipeline, taking into account the impact of segmentation and tokenization, which possess special features of social media texts, on the final annotation quality.

In the near future, based on the results of the research, the Compreno dictionary with UD format annotation and a sample of the GICR corpus annotation (silver standard) will be made available to the public.

Acknowledgements

This research was financially supported by the Russian Government Program of Competitive Growth of Kazan Federal University, and by RFBR, grant № 17-29-09163.

References

[1] *Anastasyev D. G. (2020)*, Exploring Pretrained Models for Joint Morpho-Syntactic Parser of Russian, in Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2020”, Moscow, June 17–20, 2020.

[2] *Belikov V., Kopylov N., Piperski A., Selegey V., and Sharoff S. (2013)*, Corpus as language: from scalability to register variation, in Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.

[3] *Belikov V., Selegey V. and Sharoff S. (2014)*, Preliminary considerations towards developing the General Internet Corpus of Russian, in Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2012", Moscow, pp. 37-49.

[4] *Benko V. (2014)*, Aranea: Yet Another Family of (Comparable) Web Corpora, in Proceedings of Text, Speech and Dialogue, 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014, Springer International Publishing Switzerland, 2014, pp. 257–264.

[5] *Benko V. and Zakharov V. P. (2016)*, Very Large Russian Corpora: New Opportunities and New Challenges, in Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”, Moscow, June 1–4, 2016.

[6] *Brants T. (2000)*, TnT – A Statistical Part-of-Speech Tagger, in Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, April 29 – May 3, 2000, Seattle, WA.

[7] *Inshakova E. S, Iomdin L. L., Mitushin L. G., Sizov V. G., Frolova T. I., and Zinman L. L. (2019)*, SinTagRus segodnya (SinTagRus today), in Trudy Instituta Russkogo Yazyka im. V.V. Vinogradova, t. 21, Moscow, 2019, pp. 14-41.

[8] *Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., and Suchomel V. (2013)*, The TenTen Corpus Family, in Proceedings of the 7th International Corpus Linguistics Conference, Lancaster, 2013, pp. 125–127.

[9] *Lyashevskaya O. N. and Sharoff S. A. (2009)*, A Frequency Dictionary of Russian. Access mode: http://dict.ruslang.ru/freq.php.

[10] *Lyashevskaya O. N., Shavrina T. O., Trofimov I. V., and Vlasova N. A. (2020)*, GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing, in Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialogue 2020”, Moscow.

[11] *de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006)*, Generating typed dependency parses from phrase structure parses, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC).

[12] *Selegey D., Shavrina T., Selegey V., and Sharoff S. (2016)*, Automatic morphological tagging of Russian social media corpora: training and testing, in Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialogue 2016”, Moscow.

[13] *Sharoff S. and Nivre J. (2011)*, The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge, in Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2011”, Bekasovo, pp. 591–605.

[14] *Shavrina T., Shapovalova O. (2017)*, To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser, in proc. of “CORPORA 2017”, international conference, Saint-Petersbourg, 2017. P. 78-84

[15] *Shavrina T. O. (2017)*, Metody obnaruzhenia i ispravlenia opechatok: istoricheskiy obzor (Methods of mistypes detection and correction: a historical review), in Voprosy Yazykoznania, 2017, №4, pp. 115–134.

[16] *Helmut Schmid (1994)*, Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

[17] *Helmut Schmid and Florian Laws (2008)*, Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging, in COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK.

Experiments on human incremental parsing of English

Leonid Mityushin, Leonid Iomdin
A.A. Kharkevich Institute for Information Transmission Problems
Russian Academy of Sciences
mit@iitp.ru, iomdin@iitp.ru

Abstract

Experiments have been carried out in which human subjects incrementally constructed dependency trees of English sentences. The subjects were successively presented with growing initial segments of a sentence, and had to draw syntactic links between the last word of the segment and the previous words. They were also shown a fixed number of lookahead words following the last word of the segment. The results of the experiments show that lookahead of 1 or 2 words is sufficient for confident incremental parsing of English declarative sentences.

Keywords: incremental parsing; human parsing; dependency tree; English language
DOI: 10.28995/2075-7182-2021-20-505-513

Эксперименты по инкрементальному построению синтаксической структуры английских предложений человеком

Леонид Митюшин, Леонид Иомдин
Институт проблем передачи информации им. А.А. Харкевича
Российская академия наук
mit@iitp.ru, iomdin@iitp.ru

Аннотация

Были проведены эксперименты, в которых испытуемые в инкрементальном режиме строили структуры синтаксических зависимостей для английских предложений. Испытуемым последовательно предъявлялись растущие начальные отрезки предложений, и они должен был проводить синтаксические связи между последним словом отрезка и предыдущими словами. Они также могли видеть ограниченный правый контекст – заданное число слов, следующих за последним словом отрезка. Эксперименты показали, что правый контекст размером в 1 или 2 слова достаточен для уверенного построения синтаксических структур английских повествовательных предложений.

Ключевые слова: инкрементальный синтаксический анализ; синтаксический анализ, производимый человеком; дерево зависимостей; английский язык

1 Introduction

In this work experiments are described on incremental construction of dependency trees by humans. The subjects[1] in the experiments were linguists well experienced in syntactic annotation. In each experiment, the subject was successively shown growing initial segments of an English sentence and had to draw syntactic links between the last word of the segment (**the active word**) and the previous words (**the left context**). In addition to the initial segment, at each step the subject could see a fixed number of words following the active word (**lookahead**).

This work is a repetition for English of the experiments on human incremental parsing of Russian carried out by the authors previously [Mityushin, Iomdin, 2019], and our motivation was to find out

[1] Here and below the word *subject* is used in the meaning 'a person taking part in an experiment' (the equivalent of the Russian term *испытуемый*).

whether the results for English and Russian would be similar. We could not expect it a priori as these languages are not closely related; actually, their morphology and syntax are typologically very different.

In situations of uncertainty the subjects were supposed to use so-called tentative syntactic links instead of ordinary ones. An additional option in the English experiments was a temporary increase of the lookahead size. This proved to be very useful, especially in the case of zero default lookahead. Three series of experiments were carried out for default lookahead sizes 0, 1 and 2, with 100 sentences processed in each series. The results show that lookahead of 1 and especially 2 words is enough for quite confident incremental parsing (as is the case with Russian).

This leads to the conclusion that natural language texts have certain implicit properties that make incremental parsing effective. It is reasonable to assume that such properties may be rooted in specific features of human text generation, most probably associated with gradual, incremental deployment of information and meaning (the course of which the linguist reconstructing the parse tree tries to guess). Although produced by the authors independently**,** this assumption seems to be much in line with the theory of dynamic syntax, which appeared in 1990s [see Blackburn, Meyer-Viol, 1994] and has been actively developed since, involving differently structured languages [see e.g. Kempson et al., 2001; Tugwell, 2006; Kempson et al., 2011; Kempson, Gregoromichelaki, 2017]. The ideas of dynamic syntax lie at the intersection of linguistics, cognition and brain science, and heavily rely on incrementality.

2 Syntactic model

As in the experiments with Russian, we use the representation of syntactic structures of sentences in the form of dependency trees introduced by I. Mel'čuk [1974, 1988] within the framework of the Meaning ⇔ Text theory. This representation is used in the ETAP multifunctional multilingual linguistic processor [Iomdin et al., 2012]. In this format, the nodes of a dependency tree are the words of the sentence, and the links are labelled with names of syntactic relations. In the current version of the ETAP English syntax, 64 relations are used [Apresjan et al., 1989]; a similar set of relations is described by Mel'čuk and Pertsov [1987].

To facilitate incremental construction of dependency trees, certain formal changes were made to subtrees for phrases containing prepositions or conjunctions. In the ETAP syntax, prepositions/ conjunctions dominate the noun/verb groups that follow them. For example, the sentences *She lives in Paris* and *She lives in luxury* have the following dependency trees:

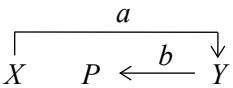
```
She    <-.       predicative          She    <-.       predicative
lives  --' --.   --                   lives  --' --.   --
in     --. <-'   1st completive       in     --. <-'   adverbial
Paris  <-'       prepositional        luxury <-'       prepositional
```

Here for each syntactic link entering a word, the corresponding name of syntactic relation is shown. Being presented with the initial segment *She lives in ...* , the subject cannot confidently decide which type of link connects the words *lives* and *in*. The sentences *He drank tea and coffee* and *He drank tea and left* have the following dependency trees:

```
He     <-.          predicative        He     <-.          predicative
drank  --' --.      --                 drank  --' --. --.  --
tea        <-' --.  1st completive     tea        <-'   |  1st completive
and        --. <-'  coordinative       and        --. <-'  coordinative
coffee     <-'      coordinate-        left       <-'      coordinate-
                     conjunctional                          conjunctional
```

Being presented with the initial segment *He drank tea and ...* , the subject cannot decide which word is the head of the coordinative link: *tea* or *drank*.

In order to avoid these difficulties, we used a different representation of prepositional/conjunctional constructions. The construction $X \xrightarrow{a} P \xrightarrow{b} Y$, where P is a preposition or a conjunction, a is an arbitrary syntactic relation, and b is one of the four relations: prepositional, coordinate-conjunctional, subordinate-conjunctional, or comparative-conjunctional, is replaced with the

construction  . This transformation is purely technical and can be performed automatically in both directions. The new representation allows us to postpone making decisions about preposition/conjunction attachment until the relevant content words appear. It is worth noting that the new form of these constructions is accepted in multiple syntactic models today. In particular, it conforms to the principles of the Universal Dependencies framework [see Osborne, Gerdes, 2019].

3 Experimental setup

Processing of a sentence is organized as a dialogue supported by a specially created computer program. The input to the program is a sentence in the form of a string of characters; the program splits it into words using blanks as separators. Isolated punctuation marks, such as dashes, are coupled with adjacent words, usually those on the left. The dialogue consists of N–1 steps numbered 2, 3, ... , N, where N is the number of words in the sentence. At step K, the subject is presented with a dialogue text file which shows the first K words of the sentence with the adjacent punctuation plus a fixed number of words following the word K (lookahead). When the last word of the sentence is shown, it is accompanied by the message [end of sentence]; until this message appears, the subject has no information about the length of the sentence.

At step K, the dialogue file also shows the partial syntactic structure (PSS) created at the previous steps on the words of the left context [1, ... , K–1] of word K. PSS is the main data structure supported by the program. For a given sentence, PSS may be any set of syntactic links between the words of the sentence with the additional condition that these links form either a single well-formed directed tree or a union of disjoint well-formed trees. At the beginning of the experiment, PSS is empty. The task of the subject is to add to PSS new syntactic links, so as to obtain a complete dependency tree of the sentence after step N.

At each step, the subject can create new syntactic links between the active word K and the words of its left context. There are two types of link: ordinary and tentative. Ordinary links are "permanent", they are added to PSS at the moment of creation and are supposed to remain there. New tentative links are added to the so called "tentative pool" (another data set supported by the program, also empty at the beginning). The links of the tentative pool can be added to or removed from PSS at any moment; in particular, they can be added to PSS at the moment of creation.

If necessary, it is also allowed to add to PSS or remove from it ordinary links whose both ends belong to the left context [1, ... , K–1]. However, these actions are considered as "error correction", and the subject is instructed to prevent them as much as possible.

There is also a different sort of action: the subject can ask the program to show one additional word of lookahead. As a result, the subject gets the same dialogue file with one word added; the active word remains the same. If the command to show an additional word is given M times in a row, M additional words will be shown. At subsequent steps, the lookahead size returns to its default value.

We always presume that processing a given sentence results in producing its correct complete syntactic structure (the correspondence between sentences and their syntactic structures is discussed in Section 7). The subject's performance on a sentence is measured by three indicators: the number of corrected errors, the number of created tentative links and the number of commands to increase lookahead. In an ideal situation, all these numbers are equal to zero; in real practice the subjects are instructed by the experimenters to avoid making corrections as much as possible and to keep the number of tentative links and lookahead expansions to a minimum. Accordingly, unless there is a significant risk of error, it is preferable to use ordinary links and refrain from increasing lookahead.

4 Example

In this section we illustrate the use of ordinary links, the main building material for dependency trees. The input sentence is "*That was how the words occurred in the old Latin poem.*", and the lookahead size is 2 words. If everything goes correctly, at step 6 the subject is presented with the following dialogue file:

```
That was how the words occurred in the ......
 1   That      <-.  predicative
 2   was       --'  --
 3   how            --
 4   the       <-.  determinative
 5   words     --'  --
 6   occurred
     in
     the

---------------------------------------
|  * -->  6         |
|  6 -->  2  was    |
|  6 -->  3  how    |
|  6 -->  5  words  |
---------------------------------------

   TENTATIVE LINKS
---------------------------------------
|  create and add to PSS  |    -->
|  create                 |    -->
|  add to PSS             |    -->
|  remove from PSS        |    -->
|  increase lookahead     |
---------------------------------------

   ERROR CORRECTION
---------------------------------------
|  add to PSS             |    -->
|  remove from PSS        |    -->
---------------------------------------
```

At this point, syntactic links should be created between the active word 6 (*occurred*) and the words of the left context. The subject writes in the first field the abbreviated names of syntactic relations for the new links (and for the link going left to right to the active word, also the number of its head word):

```
---------------------------------------
|  * -->  6         |  2  copulat
|  6 -->  2  was    |
|  6 -->  3  how    |  adverb
|  6 -->  5  words  |  predic
---------------------------------------
```

At the next step, these links are added to PSS, and the subject gets this dialogue file:

```
That was how the words occurred in the old ......
 1   That      <-.              predicative
 2   was       --'         --.  --
 3   how               <-.   |  adverbial
 4   the       <-.       |   |  determinative
 5   words     --' <-.   |   |  predicative
 6   occurred      --' --' <-'  copulative
 7   in
     the
     old

---------------------------------------
|  * -->  7       |
|  7 -->  2  was  |
---------------------------------------

   TENTATIVE LINKS
   ......
   ERROR CORRECTION
   ......
```

Then the subject moves on without creating new links until the last word 11 (*poem*) becomes active:

```
That was how the words occurred in the old Latin poem.
 1   That      <-.              predicative
 2   was       --'         --.  --
 3   how               <-.   |  adverbial
 4   the       <-.       |   |  determinative
 5   words     --' <-.   |   |  predicative
 6   occurred      --' --' <-'  copulative
 7   in                         --
 8   the                        --
 9   old                        --
10   Latin                      --
11   poem.
    [end of sentence]

----------------------------------------
|  * --> 11         |
| 11 -->  2  was    |
| 11 -->  7  in     |
| 11 -->  8  the    |
| 11 -->  9  old    |
| 11 --> 10  Latin  |
----------------------------------------

   TENTATIVE LINKS
   ......
   ERROR CORRECTION
   ......
```

The subject writes the relation names for the links between word 11 and the left context:

```
----------------------------------------
|  * --> 11         |  6  2-compl
| 11 -->  2  was    |
| 11 -->  7  in     |  prepos
| 11 -->  8  the    |  determ
| 11 -->  9  old    |  modif
| 11 --> 10  Latin  |  modif
----------------------------------------
```

These links are added to PSS, making PSS a complete dependency tree of the sentence. It is presented to the subject for the final check. Note that error correction is still possible at this point.

```
 1  That      <-.                  predicative
 2  was       --'             --.  --
 3  how               <-.     |    adverbial
 4  the       <-.       |     |    determinative
 5  words     --' <-.   |     |    predicative
 6  occurred      --' --' <-' --.  copulative
 7  in                        <-.  prepositional
 8  the               <-.     |    determinative
 9  old           <-.   |     |    modificative
10  Latin     <-.   |   |     |    modificative
11  poem.     --' --' --' --' <-'  2nd completive
    [complete structure]

   ERROR CORRECTION
----------------------------------------
|  add to PSS             |    -->
|  remove from PSS        |    -->
----------------------------------------
```

5 Increasing lookahead

Technically, the command to increase lookahead is given by typing any single character in the blank field on the right of the words "increase lookahead". It is incompatible with other commands, so no other changes should be made to the dialogue file. In the experiments with a zero default lookahead this command was used quite frequently – roughly speaking, "for every singular noun". The reason is that we often cannot correctly parse noun phrases until we get to the end of them. For example, consider the sentence

The temperature control device adjustment technique is described in the Appendix.

Here we have a chain of nouns connected with the compositive relation: *temperature* ← *control* ← *device* ← *adjustment* ← *technique*, and *the* is dominated by the rightmost noun *technique* (with the determinative relation). We cannot correctly attach *the* until we identify the last word of the chain. So it would be quite natural, when we are at step 2 (that is, when we are shown the segment *The temperature ...*), to repeatedly increase lookahead until the noun phrase comes to an end, and only then establish the syntactic links within the phrase moving from step 3 to step 6.

6 Tentative links

Another way to deal with uncertainty is to use tentative links. For example, with the sentence in Section 5 we might move from active word 2 (*temperature*) to 6 (*technique*), and at each step create a tentative determinative link from the current active word to the article *the*. As these links are mutually incompatible, we would keep them in the tentative pool rather than add to PSS. Simultaneously, at steps 3 to 6, we would create ordinary compositive links between the active word and the previous one. Then, when we come to the active word 7 (*is*), we would add the tentative link *the* ← *technique* to the PSS, and the processing of the sentence would continue.

It is clear, however, that in this case increasing lookahead is much more convenient. Nevertheless, there are situations of "long distance non-confidence" where it is quite natural to use tentative links. Consider the sentence

Weather forecast: clouds with sunny spells and occasional showers.

Its syntactic structure contains an explicative link between the heads of the two parts divided by the colon: *forecast* → *clouds* . In contrast, this sentence with two words added

Weather forecast: clouds with sunny spells and occasional showers are expected.

has an explicative link *forecast* → *are* . When presented with the initial segment *Weather forecast: clouds* with *clouds* as the active word (a few words of lookahead can also be shown), the subject must decide what to do with the explicative link *forecast* → *clouds.* Obviously, it would be too risky to create it as an ordinary link. The subject can repeatedly increase lookahead, but the number of increases needed may be large (of the same order as the length of the sentence). In this case the advisable course of action is to create a tentative explicative link *forecast* → *clouds* and add it to PSS. If it later proves to be incorrect it will be replaced with the correct one "free of charge".

On the whole, to avoid errors (in particular with garden-path sentences such as *The horse raced past the barn fell*), the subject should use tentative links and lookahead increases whenever straightforward use of ordinary links carries a real risk of error, however small.

7 What is the correct structure?

The task of the subjects in our experiments was to create syntactic structures for given sentences. The input sentences had no structures assigned to them in advance, so there was no gold standard for comparison. As the subjects were linguists with considerable experience of syntactic annotation, it was agreed that they themselves should decide what dependency tree is the correct structure for the given sentence. If any ordinary links in the current tree are incorrect or missing, the tree is repaired using the "error correction" field in the dialogue file. This can be done at any step up to N, or when the complete structure is presented after step N.

Sometimes the input sentence cannot be assigned a dependency tree – for example, this is true of many elliptical sentences. Such sentences are excluded from the experiment as soon as the subject discovers that they are unparseable. On the other hand, a sentence can have more than one correct syntactic structure. There are two main types of such situations: semantic indeterminacy and genuine semantic ambiguity.

Semantic indeterminacy [Ziering, Van der Plas, 2015] can be illustrated by the sentence *They mined the roads along the coast,* where the phrase *along the coast* may be attached either to the verb or to its object without essentially changing the meaning. In such cases the alternative syntactic structures are considered equally correct, and the subject is free to choose any of them as the target one.

Genuine semantic ambiguity is represented by N. Chomsky's sentence *Flying planes can be dangerous*, which has two semantically different dependency trees:

```
Flying    <-.      modificative          Flying    --. <-.  predicative
planes    --' <-.  predicative           planes    <-'   |  1st completive
can       --. --'  --                    can       --. --'  --
be        <-' --.  1st completive        be        <-' --.  1st completive
dangerous     <-'  copulative            dangerous     <-'  copulative
```

The subject has no reason to prefer one interpretation to the other (partly because sentences in the experiments are taken out of context). Moreover, at step 2 (and also at step 3 if the lookahead is 0 or 1) the subject doesn't know how long the sentence is going to be. The sentence may continue in a way that makes either of these readings the only correct one. Hence the right thing in this situation is to keep both options open by creating four tentative links: *planes → flying, can → planes, flying → planes* and *can → flying.* It should be said, however, that genuine ambiguity of this kind never really occurred in our experiments.

8 Experimental material

The sentences for the experiments were taken from the written subcorpus of the British National Corpus (BNC) [Burnard, 2007]. BNC covers British English of the late 20th century from a wide variety of genres. Its written subcorpus contains approximately 5 million sentences. For our experiments we selected sentences that satisfied the following additional requirements similar to those used in [Mityushin, Iomdin, 2019]:

(1) the number of words in the sentence is between 6 and 30;
(2) the first alphanumeric character is a capital letter;
(3) the last character is a small letter or full stop;
(4) the proportion of small letters among all alphanumeric characters is at least 90%.

The aim of these conditions was to restrict the experimental material to "ordinary declarative English sentences of moderate length". The number of sentences in the written part of BNC satisfying condition (1) is about 3.3 million, and the number of those satisfying all four conditions is about 2.6 million. The sentence length in this subset has the mean 17.0 and the standard deviation 6.7. Sentences for the experiments were selected from this subset using pseudorandom numbers.

9 Results and discussion

Three series of experiments were carried out for the default lookahead size equal to 0, 1 and 2, with 100 sentences processed in each series. The subjects in the experiments were the authors of this paper, who have an advanced command of English. The results are presented in Table 1.

Lookahead size	Total number of links in the trees	Number of lookahead increases	Number of created tentative links	Number of corrections
0	1621	373	28	4
1	1555	56	17	4
2	1686	17	7	0

Table 1. The results of the experiments.

The results demonstrate good improvement in performance for each additional word of lookahead. Although the number of corrections in the experiments was 8, the actual number of errors was 4, but each error needed two separate "correcting actions": removing the incorrect link from PSS and adding

the correct one. For the two-word lookahead, only 1.5 percent of links were accompanied with "signs of doubt" (increasing lookahead or creating a tentative link instead of ordinary one).

It is interesting to compare the present results with those for Russian [Mityushin, Iomdin, 2019]. In the Russian experiments, the average sentence length over 300 sentences was 17.6 words; in the English it was 17.2 words. The general setup of the experiments was the same, with one important difference: for Russian, the option of temporary increase of lookahead was not available. This makes strict quantitative comparison impossible. Another obstacle to strict comparison is a "greater weight" of Russian words. Analysis of parallel English–Russian corpora shows that English texts contain on average 20–30 percent more tokens than their Russian counterparts [Russian National Corpus, 2020]. The reason is the massive use of function words such as articles (absent in Russian), auxiliary verbs (infrequent in Russian) and adverbial particles (absent in Russian). This means that k Russian words as lookahead are on average more informative than k English words. Yet another important distinction is the richer morphological system of Russian and much wider use of grammatical agreement, which helps to identify syntactic links with greater confidence.

Nevertheless, the results can be compared on a qualitative level. For Russian, error correction was used 3 times (for zero lookahead), and the number of created tentative links was 75, 34 and 13 for 0, 1 and 2 words of lookahead respectively. We may say that the level of performance for English with 2 words of lookahead was somewhere between the levels for Russian with 1 and 2 words of lookahead.

10 Conclusion

The results of the experiments described in this paper, as well as those for Russian, may be regarded as arguments supporting the following general picture of text comprehension. Suppose that only ordinary links have been used to build the syntactic structure of the input sentence. This means that the structure was constructed in a strictly incremental way: links were added to it but never removed. In this case we can imagine the parsing process to develop like this: the reader/listener adds to the partial syntactic structure the links between the active word and the left context that satisfy the syntactic and semantic requirements, and later never returns to them. This strategy of immediately adding plausible links to the structure may be assumed to be used universally, even in the cases of doubt, with infrequent collisions (incompatibility of new links to be added and those already in the structure) being successfully resolved on the basis of information available at the point of collision. To make this strategy effective, natural language texts should have certain implicit properties, possibly connected with the specific character of human text generation.

Acknowledgements

This research was supported by a grant from the Ministry of Science and Higher Education of Russia No. 075-15-2020-793.

References

[1] *Apresjan Ju., Boguslavsky I., Iomdin L., Lazursky A., Pertsov N., Sannikov V., Tsinman L.* (1989), The Linguistics of the ETAP-2 System [Lingvisticheskoe obespechenie sistemy ETAP-2]. Nauka, Moscow. (in Russian)

[2] *Blackburn P., Meyer-Viol W.* (1994), Linguistics, logic and finite trees. Logic Journal of the IGPL. 2(1), pp. 3–29.

[3] *Burnard L.* (2007), Reference Guide for the British National Corpus (XML Edition), available at: http://www.natcorp.ox.ac. uk/docs/URG

[4] *Henderson J.* (2004), Lookahead in deterministic left-corner parsing. Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together, Barcelona, Spain, July 25–26, 2004, pp. 26–33.

[5] *Iomdin L., Petrochenkov V., Sizov V., Tsinman L.* (2012), ETAP parser: state of the art. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'yuternaya Lingvistika i Intellektualnye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2012"], Moscow, pp. 830–843.

[6] *Kempson R., Meyer-Viol W., Gabbay D.* (2001), Dynamic Syntax. The Flow of Language Understanding. Blackwell, Oxford.
[7] *Kempson R., Gregoromichelaki E., Howes C.* (eds) (2011), The Dynamics of Lexical Interfaces. CSLI Publications.
[8] *Kempson R., Gregoromichelaki E.* (2017), Action sequences instead of representational levels. Behavioral and Brain Sciences, Vol. 40, e296. DOI: https://doi.org/10.1017/S0140525X17000449
[9] *Mel'čuk I.* (1974), Towards a Theory of Meaning ⇔ Text Linguistic Models [Opyt Teorii Lingvisticheskikh Modelei "Smysl ⇔ Tekst"]. Nauka, Moscow. (in Russian)
[10] *Mel'čuk I.* (1988), Dependency Syntax: Theory and Practice. The SUNY Press, Albany, N.Y.
[11] *Mel'čuk I., Pertsov N.* (1987), Surface Syntax of English. A Formal Model within the Meaning–Text Framework. John Benjamins, Amsterdam.
[12] *Mityushin L., Iomdin L.* (2019), Experiments on human incremental parsing. Proceedings of the Fifth International Conference on Dependency Linguistics (Depling 2019), Paris, August 27–28, 2019, pp. 209–216.
[13] *Osborne T., Gerdes K.* (2019), The status of function words in dependency grammar: A critique of Universal Dependencies (UD). Glossa: a journal of general linguistics, 4(1):17, pp. 1–28.
[14] Russian National Corpus (2020), available at: http://www.ruscorpora.ru/new/search-para-en.html
[15] *Tugwell D.* (2006), Language modelling with dynamic syntax. International Conference on Text, Speech and Dialogue (TSD 2006), Brno, Czech Republic, pp. 285–292.
[16] *Ziering P., Van der Plas L.* (2015), One tree is not enough: cross-lingual accumulative structure transfer for semantic indeterminacy. Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, September 7–9, 2015, pp. 739–746.

Communication Failures in Everyday Conversations: a Case Study Based on the "Retrospective Commenting Method"

Arto Mustajoki
[1]HSE University, Moscow
Staraya Basmannaya, 21/4
105066, Moscow, Russia
[2]Helsinki University, Finland
arto.mustajoki@helsinki.fi

Natalia Cherkunova
HSE University,
St. Petersburg, 123A Kan.
Griboedova Emb. 190068,
St. Petersburg, Russia
nbcherkunova@edu.hse.ru

Tatiana Sherstinova
HSE University,
St. Petersburg, 123A Kan.
Griboedova Emb. 190068,
St. Petersburg, Russia
tsherstinova@hse.ru

Abstract

The paper deals with communication failures in everyday spoken discourse. The spontaneous character of oral speech is its basic property and becomes a prerequisite for the appearance of such a phenomenon as communicative failures. By communicative failures, we mean speech situations when the recipient of a speech message does not understand it correctly, i.e., in the way the speaker intended. The purpose of this pilot study is 1) to assess the total number of communication failures that occur with a person during a single day and 2) to determine the dependence of communication failure frequency on the communication settings and conditions. The main result of the study is a qualitative and quantitative assessment of communication failures during a subjects's day. The research is based on a special experiment based on 24-hour monitoring of the subject's speech and his subsequent retrospective commentary on all recorded data. Such an approach allows one to reduce the subjectivity inherent in much linguistic work. The research continues a series of studies devoted to the effectiveness of spoken communication and is important not only for understanding the fundamental processes of speech perception but is also crucial for the development of artificial intelligence systems involving human-computer speech dialogue systems and for speech technologies of the next generation.

Keywords: everyday speech communication; spoken Russian; dialogue structure; speech corpus; oral discourse; miscommunication

DOI: 10.28995/2075-7182-2021-20-514-523

Коммуникативные Неудачи в Повседневном Речевом Общении: Пилотное Исследование с Использованием Метода Ретроспективного Комментирования

Арто Мустайоки
[1]НИУ ВШЭ, Москва
Старая Басманная, 21/4,
105066, Москва, Россия
[2]Хельсинский университет,
Финляндия
arto.mustajoki@helsinki.fi

Наталья Черкунова
НИУ ВШЭ,
Санкт-Петербург
Наб. Канала Грибоедова
123А, 190068,
Санкт-Петербург, Россия
nbcherkunova@edu.hse.ru

Татьяна Шерстинова
НИУ ВШЭ,
Санкт-Петербург
Наб. Канала Грибоедова
123А, 190068,
Санкт-Петербург, Россия
tsherstinova@hse.ru

Аннотация

В статье рассматриваются коммуникативные неудачи в повседневном устном дискурсе. Спонтанность устной речи является ее базовым свойством и является предпосылкой возникновения такого явления, как коммуникативные неудачи. Под коммуникативными неудачами мы понимаем речевые ситуации, когда адресат речевого сообщения понимает его неправильно, т. е. не так, как задумал говорящий. Целями данного пилотного исследования является: 1) оценка общего количества коммуникативных неудач, которые происходят с человеком в течение одного дня и 2) определение зависимости частоты коммуникативных неудач от условий коммуникации. Главный результат исследования — количественная оценка коммуникативных неудач

информанта в течение дня и их качественная интерпретация. Предлагаемое исследование основано на специальном эксперименте, основанном на круглосуточном мониторинге речи испытуемого и его последующем ретроспективном комментарии ко всем записанным данным. Такой подход позволяет уйти от субъективизма, присущего многим лингвистическим исследованиям. Эта работа продолжает серию исследований, посвященных эффективности речевого общения, полученные результаты важны не только для понимания фундаментальных процессов восприятия речи, но также имеют решающее значение для разработки систем искусственного интеллекта, включающих системы речевого диалога человек-компьютер, а также для речевых технологий следующего поколения.

Ключевые слова: повседневная речевая коммуникация; русская разговорная речь; структура диалога; речевой корпус; устный дискурс; коммуникативные неудачи

1 Introduction

The research continues a series of studies devoted to the effectiveness of spoken communication. The spontaneous character of oral speech is its basic property and becomes a prerequisite for the appearance of such a phenomenon as communicative failures. By communicative failures, we mean speech situations when the recipient of a speech message does not understand it correctly, i.e., in the way the speaker intended.

The causes and risks of miscommunication have been widely discussed in the research literature (see e.g. [1; 2; 3; 4; 5; 6; 7; 8; 9; 10]). However, there is very little evidence about the frequency of communication failures.

If people are asked whether they have had any communication failures during a day, they usually say one or two or just none. One example of this is the experiments aimed at counting communication failures during a single day by a group of philology students at the National Research University Higher School of Economics, St. Petersburg. The students were asked to keep a "miscommunication diary" and register all communication failures they have. The number of communication failures turned out to be from 0 to 2 on average. A comparable experiment was conducted among students at Helsinki University and the results were similar. This means that we usually do not identify and register communication failures even in a situation where we should collect them. In everyday life, people remember only the most drastic communication failures which have had serious or amusing consequences.

There are some quantitative studies on very specific communication situations, e. g. in conversation concerning railway and air traffic control [11; 12] and health care contexts [13]. However, there is very little data on the frequency of communication failures in everyday settings, Ermakova's and Zemskaya's study [14] being an exception. On the basis of a large amount of authentic material, they arrived at the conclusion that communication failures are more frequent in everyday speech with family members and good friends than in conversation with strangers. The results sound paradoxical but there are good reasons for this [15; 16].

Ermakova's and Zemkaya's study is rich source of materials and has been carried out very thoroughly but is suffers from a methodological gap which is a major problem in research on miscommunication. If a researcher examines interaction between people as an observer or uses video or audio recording, s/he is unable to identify all the instances of misunderstandings. As shown e.g. in [17; 18; 19; 20], misunderstandings are often latent or covert. Recipients tend to apply the "let it pass" tactic [21]. There are various reasons for not asking for clarification in the case of misunderstanding: (1) the topic is not interesting for the recipient; (2) s/he thinks that s/he has already understood enough; (3) s/he believes that s/he will later understand what was said; (4) s/he does not want to interrupt the course of interaction; (5) s/he does not want to show her/his ignorance.

The purpose of the study presented in this paper was 1) to assess the total number of communication failures that occur with a person during a single day and 2) to determine the dependence of communication failure frequency on the communication settings and conditions.

2 The "Retrospective Commenting Method" and Methodology

In [22] a new methodological tool was presented to gain a deeper insight into the problems of communication. The "Retrospective Commenting Method" (RCM) aims to tackle the weakness of other methods by working afterwards on a recorded material with the informant. The new method is quite laborious but enables one to get more precise information on what really took place in a conversation.

The main stages of the methods are the following (see in detail [22]):

1) The first stage consists of recording all the communication the informant has had during one day. This takes place within the "One Day of Speech" project which is currently being conducted at Saint Petersburg University [23; 24; 25; 26]. This material is valuable as such, and much research has been done on the basis of "one day communication packages" of more than one hundred informants. Before the recording, the subject was asked to note the situations of miscommunication during the day of recording [22].
2) What is new in the RCM is the second stage, namely the way the recorded material is analysed. It takes place with the active participation of the informant. S/he (she in the pilot study) goes through the whole recorded material with the researcher, commenting on what really occurred in the communication settings she had during the day. Before listening to audio recording together with the researcher, the participant was told that she should note and comment the followings aspects of her recorded communication:
   - The main task was to distinguish and describe miscommunication situations and any other types of communicative failures (e.g., when something was understood in a different way from what was intended by the speaker, when the participant did not understand something from the speech of her interlocutor but chose to pretend that everything was fine, or when the informant believed that she was misunderstood, etc.).
   - Then, she should explain to the researcher any communication situations that a stranger could not understand correctly. It included description (clarification) of the communication context or some phrase/word meaning: (*Here, I am talking about…*), attribution of emotions (e.g., *At this moment I am very annoyed, but I try not to show it.*), explanation of her dialogue tactic (e.g., *I am speaking this way because...*), revealing hidden humour, irony, or language play (*Here, I am being ironic.*), etc.

   The researcher was listening to the recording with the participant, they discussed the recording together and this discussion was also recorded. The examination of one day's material took three days in the pilot study [22].

The total amount of audio material received was 22.5 hours: 8.5 hours of source material and 14 hours of recording of the commenting process.

Experiment participant: woman, 40 years old. Profession: actress, art critic, and university professor.

During the day, the informant interacts with:

- her mother;
- her daughter;
- her husband;
- a doctor at an outpatient clinic;
- her colleagues;
- her students;
- and occasional strangers.

3. Results: Daily Communication Failures in Numbers

The experiment shows that the retrospective commenting method enables one to go much deeper into decisions made by interlocutors than a normal examination of recorded material by the researcher her/himself alone. However, the job of informant is not always easy. What is most difficult to identify are those communicative failures in which there is no obvious conflict of interests and opinions. Another unclear situation occurs when the interlocutor stops communication (for example, leaves the room) or simply abruptly changes the topic of conversation. For the informant, being unable to read the interlocutor's thoughts in detail, it is difficult to identify the presence of a communication failure.

A further observation on the process of analysis of the material by the informant was that from the 19 communication failures, she identified only 13 instantly. The remaining 6 required the involvement of the assisting researcher.

The most surprising and important result was the total number of communication failures that occurred during the day, being as many as 19. Naturally, it is also interesting to see which categories they belong to. In analyzing them, we apply the following criteria of classification:

- place of communication and recipient;
- reasons for communication failures;
- resolved or unresolved communication failure.

Table 1 shows the distribution of communication failures, depending on the place of communication and recipient. According to the data, 10 (53%) of the total number are communication failures that occurred in the informant's home during communication with relatives. There were 5 failures at work and 4 in public places.

The main conclusion which can be made on the basis of this table is the great variety of reasons for communication failures, four being the highest number. There are slight differences in the numbers concerning domestic and external conditions, but one cannot see on that basis any reliable differences. Each type will be described in more detail below.

Place of communication	Participant	Total number
At home (53%)	Mother – 6 Daughter – 4	10
At work (26%)	Students – 3 Colleagues – 2	5
Other places (medical center, university) (21%)	Doctor – 1 Strangers – 3	4

Table 1: Distribution of communication failures depending on the place of communication and the interlocutors (recipients)

Table 2 presents the reasons for communication failures that occurred during the day of recordings divided into two groups: domestic and external.

Reasons	Domestic	External	Total number
Lack of interest in communication	2	2	4
Emotional effect	1	2	3
Insufficient volume (noise)	1	2	3
Distortion of information	1	1	2
Opinion imposition	2	0	2
Wrong interpretation	1	1	2
Incomplete information	1	1	2
Abrupt change of topic	0	1	1
Joke	0	1	1

Table 2: The main reasons for communication failures

Finally, Table 3 illustrates the ratio of resolved and unresolved communication failures.

Type	Percent
Resolved	61%
Unresolved	33%
Ambiguous	6%

Table 3: The ratio of resolved and unresolved communication failures

Unresolved communication failures are understood as those episodes of communication in which no attempt is made to clarify the situation (e. g., the interlocutor leaves the room or changes the topic). We may see that in our data one third of all communication failures remained unresolved.

4. Examples of Communication Failures and Their Settings

Let us have a look at some concrete examples of communication failures. As was shown, the most frequent reason for communication failures is a lack of interest in communication.

4.1. Lack of interest in communication

Case 1

- Participants in the communication: A — the informant; B — informant's mother.
- Setting: at home.
- Communication time: morning.
- Situational context: a dialogue about one of the informant's acquaintances.
- Transcript:

```
B: Eto pacient psikhiatora //
A: Da / da //
B: I ty ego podvodi k etomu //
A: Tak on voobshche // On prinimaet to, chto etot emu…
B: Tak / a pochemu on ne idet?
A: On / pr… prinimaet / net // Oni obshchalis' // On prinimaet to / chto tot emu skazal //
B: Da?
A: Da / na noch' tam dlya sna chto-to //
```

- Translation:

```
B: He is a psychiatric patient //
A: Yes / yes //
B: And you should hint it to him //
A: So he generally // He takes pills which the doctor…
B: So / why doesn't he go?
A: He / ta… takes / No // They communicated // He takes pills that / that the doctor told him //
B: Yes?
A: Yes / he takes something to sleep at night //
```

- Communication failure: resolved.
- Comment: Although a person is busy with something, family members often continue to communicate with her/him. Here the informant is looking for documents to visit the clinic and her mother tries to involve her in communication. As a result, the answers become short and rather meaningless. The informant apparently does not pay due attention to the dialogue. However, when the informant clearly explains his thought, the communication failure becomes resolved.

Case 2

- Participants in the communication: A – informant; B – her colleague.
- Setting: in the university.
- Communication time: afternoon.
- Situational context: the informant tries to relax in the teaching room. Her colleague enters the room and begins to tell a story about his relatives.
- Transcript:

```
B: I vot posle etogo on chetyre goda umiral //
A: Oj //
B: Rany vse pootkryvalis' / a serdce zdorovoe // On uzhe ego i pristrelit' prosil //
A: Von kakie organizmy na svezhem vozdukhe // Nu vot eto vse / gorodskie / navernoe / stol'ko ne zhivut //
```

```
B: Vot etot vot ded / kotoromu devyanosto shest' s polovinoj let / dedu
moemu / v devyanosto let doma rubil //
```

- Translation:

```
B: And after that he had been dying for four years //
A: Oh //
B: The wounds all opened / though his heart was healthy // He even
asked them to shoot him //
A: Fresh air produces healthy bodies // Well, that's all / city dwellers
/ probably / don't live so long //
B: My grandfather / who is ninety-six and a half years old / my grand-
father / at ninety years old he built houses //
```

- Communication failure: unresolved.
- Comment: In this case, the goals of the participants in the dialogue are different. The informant's interlocutor wants to get a story off his chest (so-called unburden speech, cf. [6, 27]). It is important for him to be heard, but he is not interested in a real dialogue and simply ignores the informant's reaction. This miscommunication is unsolved, because the informant ends the conversation, referring to the fact that it is time for her to go to her students for an exam.

4.2. Opinion imposition

Case 3

- Participants in the communication: A – informant; B – the informant's daughter.
- Setting: at home.
- Communication time: morning.
- Situational context: the informant's daughter wakes up. The informant is going to play a goose game they have invented with her daughter.
- Transcript:

```
A: Tak / nu davaj / kotenok // Ga-ga //
B: Maaa //
A: Chto? Chto snyat' khalat? Tak / rubashka …
B: Mam / a zachem ty snyala zapis'?
A: Davno vstala / davno umylas' // Ya kak-to segodnya rannyaya ptashka
//
B: Da ne // Pochemu ty / nu vot?
A: Chto pochemu?
B: Mam / a pochemu ty nu vot snyala vot etu zapis'?
A: Ono vse ravno zapisyvaet / v komnate mozhno vot tak //
```

- Translation:

```
A: So / come on / Kitten // Ga-ga //
B: Mum//
A: What? What?  Take off my robe? So / shirt …
B: Mum / why did you take the recorder off?
A: I got up a long time ago / washed my face a long time ago // I'm an
early bird today //
B: No // Why are you / this?
A: Why what?
B: Mum / why did you take this recorder off?
A: It is recording / it is possible to do this in the room //
```

- Communication failure: resolved
- Comment: Communication with children deserves special attention and an appropriate recipient design. Parents in most cases are not fully involved in children's games but respond with routine phrases. In this case, the informant tries to impose her opinion on the child (to predict what the

daughter wants from her). The situation is resolved when the child for the third time clearly formulates his question.

4.3. Emotional effect

Case 4

- Participants of communication: A – informant; B – a stranger.
- Setting: in the university.
- Communication time: afternoon.
- Situational context: a security guard at the university asks the informant to escort a man to the office he is looking for.
- Transcript:

```
A: A voobshche / eto – prepodavatel'skaya //
B: Davajte snachala nachnem s vas //
A: S menya ne nado nachinat'! A chto vy hotite? Chto vy hotite?
B: Tut takaya informaciya proshla // Tut teatr "Vstrecha%"[1] est'.
A: Teatr na vtorom etazhe // U Zimina%? U Zimina%? U kakogo mastera?
B: Ne znayu / ne znayu kakogo mastera //
A: Net / zdes' mastera kursov est' // Chtoby pryamo teatr byl…
B: Da tak i nazyvaetsya "Vstrecha%" //
```

- Translation:

```
A: Actually / this is a teacher’s room //
B: Let’s start with you first //
A: You don't have to start with me! What do you want? What do you want
from me?
B: There was such information // There should be a theatre "Vstrecha%"
here.
A: The theatre is on the 2nd floor // Zimin%'s one? Zimin%'s? Which
teacher do you need?
B: I don’t know / I don’t know which teacher //
A: No / there are teachers of course // I doubt about a theatre…
B: [Yes] that's what it's called "Vstrecha%" //
```

- Communication failure: not resolved.
- Comment: When talking with the stranger, the informant feels disturbed by his phrase “Let’s start with you first”, which she perceives as his misplaced attempt to flirt. For this reason, the informant becomes uncomfortable and reacts emotionally “You don’t have to start with me! What do you want? What do you want from me?”, as if suspecting her interlocutor might be insane or an agent. Such a reaction means that the stranger explains his position and the conversation proceeds. However, the intent of his disturbing remark, which causes miscommunication, remains unknown.

5. Conclusion

The study was based on a special experiment on 24-hour monitoring of the subject’s speech and his/her subsequent retrospective commentary on all recorded data. Such an approach, where a participant of interaction him/herself comments and describes the details of spoken interaction is unique in linguistic corpus studies and allows one to make research less subjective. The method used enables us to get not only qualitative but also some quantitative data on communication failures. To our knowledge, this is the first attempt to get a real picture of the frequency of communication failures in settings of everyday conversation. In this pilot study, we examined only one day’s interaction of a single person. Therefore,

[1] % — all proper names are anonymized.

one cannot make far-going conclusions on this basis. However, we got important hints about the frequency and forms of miscommunication in everyday communicative settings.

The main result of the study is a rather large number of communication failures during the subject's day — 19. If every person in the world meets on average at least 10 communication failures a day, it means that billions situations of miscommunication take place every single day. One can only image the damage caused by them. An unrealistic goal should be to try to free the world from failures in communication totally, but more research could give us tools to reduce their number. Even a small reduction of cases of miscommunication should make the world a better place to life.

A further important outcome of the study is a positive experience of using the retrospective commenting method. The method has, naturally, its own limitations but in comparison to traditional methods, it gives a much more reliable picture of the reality of a face-to-face everyday interaction. In the process of analyzing communication, difficulties arose with settings in which the situation had no outcome. In other words, communication is interrupted for some reason (for example, one of the communication participants leaves the room) or a completely new topic of dialogue is introduced into the communication. These situations need more attention during the commenting process. All in all, the method could provide a deeper insight into other big themes of interactional linguistics, e.g., a realization of the principles of politeness [28] and cooperation [29]. A further topic could be turn-taking, a central question in conversational analysis [30].

As to the concrete forms of communication failures, only very preliminary observations can be made. First, they are very diverse. Altogether eight different reasons were identified. This tells of the main characteristic of everyday interaction: it changes all the time and a setting is never repeated in the same form as it first took place.

A second observation confirms the claim that home is a rather "dangerous place" for communication (cf. [27; 14; 16]). This is a rather controversial discovery, because interaction with people you know well should be much easier than meeting with people without such a common ground. However, as shown in [31], domestic circumstances include some features which are favorable to communication failures. People are relaxed when spending their leisure time and do not always want to concentrate on interaction. They feel overconfident and take risks by relying on common ground. This results in the use of cryptic speech by the speaker and non-listening and overguessing by the recipient.

A third general remark concerns the reasons for communication failures. The study shows that they are very seldom purely linguistic, e. g. ambiguous constructions or vague meanings of words. More often, they come from circumstances and/or poor concentration on interaction by the communicants. Therefore, the main reason for communication failures, regardless of the communication conditions, seems to be insufficient involvement in the communication process. This violates the basic principle of successful interaction – cooperation.

Being the first pilot study of its kind, the results provide important building blocks for better understanding of the fundamental processes of spoken communication. In addition to that, the results should also be considered in the development of artificial intelligence systems involving human-computer speech dialogue systems and for speech technologies of the next generation.

Evidently, miscommunication in spoken interaction needs much more research. An important issue which remains totally unclear, is the borderline between sufficient understanding and non-sufficient understanding (cf. [32; 33]). The results obtained are important for a deeper understanding of the fundamental processes of spoken communication. They should be also taken into account when artificial intelligence systems involving human-computer interaction based on speech technologies are developed.

Acknowledgements

The experiment of longitudinal recording and its Retrospective Commenting Method which gave speech data for the given research was supported by the University of Helsinki. The methodology of longitudinal recording was approved during the creation of the ORD speech corpus, which is being created in St. Petersburg State University and was supported by several grants: the Russian Foundation for Humanities projects # 07-04-94515e/Ya (*Speech Corpus of Russian Everyday Communication "One Speaker's Day"*), #12-04-12017 (*Information System of Communication Scenarios of Russian Spontaneous Speech*). Significant extension of the corpus was achieved in the framework of the project

"*Everyday Russian Language in Different Social Groups*" supported by the Russian Scientific Foundation, project # 14-18-02070.

References

[1] Falkner W. Verstehen, Missverstehen und Missverständnisse: Untersuchungen an einem Korpus englischer und deutscher Beispiele. — Tübingen: Niemayer, 1997.
[2] Bazzanella C. and Damiano R. The interactional handling of misunderstanding in everyday conversations // Journal of Pragmatics. — 1999. — T. 31. — Vol. 6. — pp. 817-836.
[3] Dascal M. Introduction: Some questions about misunderstanding // Journal of Pragmatics. — 1999. — T. 6. — Vol. 31. — pp. 753-762.
[4] Tzanne A. Talking at Cross-Purposes: The Dynamics of Miscommunication. — Amsterdam/Philadelphia: John Benjamins. — 2000.
[5] House J., Kasper G., Ross S. Misunderstanding talk // Misunderstanding in social life. — Routledge, 2014. — pp. 9-29.
[6] Mustajoki A. A speaker-oriented multidimensional approach to risks and causes of miscommunication // Language and dialogue. — 2012. — T. 2. — Vol. 2. — pp. 216-243.
[7] Verdonik D. Between understanding and misunderstanding //Journal of Pragmatics. — 2010. — T. 42. —Vol. 5. — pp. 1364-1379.
[8] Roberts G., Langstein B., Galantucci B. (In) sensitivity to incoherence in human communication // Language & Communication. — 2016. — T. 47. — pp. 15-22.
[9] Bazzanella C. The complex process of mis/understanding spatial deixis in face-to-face interaction// Pragmática Sociocultural / Sociocultural Pragmatics — 2019. —T. 1. —Vol.7. — pp. 1-18.
[10] Honghui Z., Dongchun C. Understanding misunderstandings from socio-cognitive approach to pragmatics // International Journal of Language and Linguistics. —2019. — T. 5. —Vol. 7. — pp. 194-201.
[11] Gibson W. H. et al. A taxonomy of human communication errors and application to railway track maintenance // Cognition, Technology & Work. — 2006. — T. 8. — Vol. 1. — pp. 57-66.
[12] Gibson T., Megaw T., Donohoe L. Failures in pilot-controller communications and their implications for datalink // Engineering Psychology and Cognitive Ergonomics. — Routledge, 2016 — Vol. 5. — pp. 325-334.
[13] Docherty N.M., McCleery A., Divilbiss M., Schumann E.B., Moe A., Shakeel M. K. Effects of social cognitive impairment on speech disorder in schizophrenia // Schizophrenia Bulletin. — 2013. — T. 3. — Vol. 39. — pp. 608–616.
[14] Ermakova O.V., Zemskaya E.A. (1993), K postroyeniyu tipologii kommunikativnykh neudach [Towards the construction of a typology of communicative failures] // Russkiy yazyk v yego funktsionirovanii. Kommunikativno-pragmaticheskiy aspect [Russian language in its functioning. Communicative and pragmatic aspect], Moskow: Nauka, pp. 90-157.
[15] Mustajoki A. Why is miscommunication more common in everyday life than in lingua franca conversation // Current issues in intercultural pragmatics. — 2017. — pp. 55-74.
[16] Mustajoki A. (2011), Pochemu obshchenie na lingua franca udaetsia tak khorosho [Why interaction in a lingua franca is so successful] // Iazyki sosedej: mosty ili bar'ery? Problemy dvuiazychnoi kommunikatsii [The languages of the neighbors: bridges or barriers? Problems of bilingual communication], St Petersburg: Institut ligvisticheskikh issledovanii RAN, Evropeiskii universitet, pp. 10–31.
[17] Linell P. et al. Troubles with mutualities: Towards a dialogical theory of misunderstanding and miscommunication // Mutualities in dialogue. — 1995. — pp. 176-213.
[18] Hinnenkamp V. Constructing misunderstanding as a cultural event // Culture in Communication: Analyses of Intercultural Situations. — Amsterdam/Philadelphia: John Benjamins, 2001. — pp. 211-243,
[19] Hinnenkamp V. Misunderstandings: Interactional structure and strategic resources // Misunderstandings in Social Life: Discourse Approaches to Problematic Talk. — London etc.: Longman, 2003. — pp. 57-81.
[20] Pietikäinen K. S. Misunderstandings and ensuring understanding in private ELF talk // Applied Linguistics — 2016. — pp. 1–26.
[21] Firth A. The lingua franca factor // Intercultural Pragmatics. — 2016. — T.2. — Vol.6. — pp. 147–170.
[22] Mustajoki A., Sherstinova T. The "Retrospective commenting" method for longitudinal recordings of everyday speech // International Conference on Speech and Computer. — Springer, Cham, 2017. — pp. 710-718.
[23] Asinovsky A. et al. The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation // International Conference on Text, Speech and Dialogue. — Springer, Berlin, Heidelberg, 2009. — pp. 250-257.
[24] Sherstinova T. The structure of the ORD speech corpus of Russian everyday communication // International Conference on Text, Speech and Dialogue. — Springer, Berlin, Heidelberg, 2009. — pp. 258-265.

[25] Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic extension of the ORD corpus of Russian everyday speech //International Conference on Speech and Computer. — Springer, Cham, 2016. — pp. 659-666.
[26] T. Sherstinova. Studying Linguistic Variation and Communicative Diversity on the Basis of the "One Day of Speech" Corpus // Urban Voices: Studies in Sociolinguistics, Grammar and Pragmatics of Spoken Russian. — Peter Lang Int. Ac. Publ, 2019. — pp. 15-36.
[27] Mustajoki A., Sherstinova T., Tuomarla U. Types and functions of pseudo-dialogues // From Pragmatics to Dialogue. — 2017. — T. 31. — pp. 189–216.
[28] Leech G. N. Principles of Pragmatics London: Longman Group Ltd. — 1983.
[29] Grice H. P. Logic and conversation // Speech acts. — Brill, 1975. — C. 41-58.
[30] Sacks H., Schegloff E.A., Jefferson G.A. A simplest systematics for the organization of turn-taking in conversation // Language. —1974 — Vol. 50. — pp. 696-735.
[31] Mustajoki A., Bajkulova A. The risks of misunderstandings in family discourse: Home as a special space of interaction // Language and Dialogue. — 2020. — T. 10. — Vol. 3. — pp. 340-368.
[32] Gander A. J. Understanding in real-time communication: Micro-feedback and meaning repair in face-to-face and video-mediated intercultural interactions // URL: http://hdl. handle. net/2077/56223. — Gothenburg: BrandFactory, 2018.
[33] Gander A. J., Gander P. Micro-Feedback as Cues to Understanding in Communication //CLASP Papers in Computational Linguistics. – pp. 1-11.

RuSimScore: unsupervised scoring function for Russian sentence simplification quality

Mikhail Orzhenovskii
Saint Petersburg, Russia
orzhan057@gmail.com

Abstract

We propose an unsupervised complex scoring function (RuSimScore) to measure simplification quality of Russian sentences, and a model for text simplification based on this function. The function allows to score simplicity and original meaning preservation. First, filtered a noisy parallel corpus (machine translated WikiLarge) and extracted good simplification examples. After that, a pretrained language model was fine-tuned on these examples. We generate multiple outputs from the language model and select the best one according to the scoring function. The weights in the scoring function can be adjusted to balance between better content preservation and getting simpler sentences (controllable simplification).

Keywords: text simplification, pretrained language models, Russian
DOI: 10.28995/2075-7182-2021-20-524-532

RuSimScore: функция для оценки качества упрощения текста на русском языке

Орженовский М.В.
Санкт-Петербург, Россия
orzhan057@gmail.com

Аннотация

Мы предлагаем составную оценочную функцию (RuSimScore) для измерения качества упрощения текстов на русском языке, а также модель, построенную с помощью этой функции. Она позволяет оценить простоту результата и степень сохранения смысла исходного текста. Сначала из зашумленного корпуса (переведенный WikiLarge) отфильтровали примеры упрощения с достаточным качеством. Затем предобученная языковая модель была дообучена на этих примерах. С помощью этой языковой модели мы генерировали множество выходных предложений и выбирали лучшее на основе оценочной функции. Веса оценочной функции можно изменять (контролируемое упрощение), чтобы выбирать между лучшим сохранением смысла или более простыми выходными предложениями.

Ключевые слова: упрощение, предобученные языковые модели, русский язык

1 Introduction

RuSimpleSentEval (RSSE) competition[19] introduced the first text simplification dataset for Russian, which was collected on a crowd-sourcing platform. Another dataset suitable for Russian text simplification is WikiLarge, built from machine translated English Wikipedia and Simple English Wikipedia. WikiLarge has two issues: sometimes inaccurate alignment and errors introduced during machine translation. It has be be filtered in order to obtain high quality sentence pairs for training.

The idea of the proposed scoring function is to combine different aspects of simplification quality in a single-number metric which depends only on source (complex) and target (simplified) sentences, and does not require human labeling (like SARI). It is built from six different simple functions, which will be described later.[1]

[1]The code for training and running the model, as well as best model's weights will be published as open source at `https://github.com/orzhan/rusimscore`

We did not use the scoring function to directly calculate loss during the model training. Instead, we filtered the machine translated WikiLarge dataset (getting better result than with non-filtered one), extracting 15% of examples. Then we fine-tuned pretrained language model ruGPT-3 on the resulting dataset (sequence to sequence task can be converted to language modeling task by inserting special tokens into prompt). During inference we used the fine-tuned language model to generate multiple answers for each input sentence using nucleus sampling, and then selected the best answer with the scoring function.

2 Related work

Text simplification is often done with a sequence to sequence model trained on parallel corpus. Zhang and Lapata[21] use reinforcement learning of with a task-specific reward function. Martin et al.[3] trained a sequence to sequence model with controllable parameters. They also filtered sentences suitable for simplification from a very large corpus, based on cosine similarity of sentence level embeddings[16] .

Another approach is simplifying text in several consecutive steps. The task can be formalized as sequence labeling [15], generating a sequence of changes using a programmer-interpreter approach [7], or iteratively applying changes to a text [10].

Kuvshinova[12] solved sentence compression task for Russian with deletion based approach. This task is related to simplification.

Readability evaluation is language specific. For Russian, Ivanov et al.[11] identified text features that indicate complex texts. Laposhina et al.[2] explored readability formulas.

Language models can solve a wide range of language processing tasks including abstractive summarization [14] and paraphrase generation [20], these tasks are related to simplification.

Our work uses fine-tuned ruGPT3 language model [13] for text generation.

3 Model description

3.1 Scoring function

The idea of scoring function RuSimScore(c,s) is to estimate if a simple sentence s is a good simplification of the complex sentence c. Good simplification means that the target sentence is simple, and the meaning of the source sentence is preserved. Scoring function is calculated as multiplication of four simplicity scoring functions: lexical complexity score, dependency tree depth score, length score, reading ease score, and two content preservation scoring functions: cosine similarity score, named entity preservation score.

$$RuSimScore(c,s) = LS^{\alpha}(c,s)DD^{\beta}(c,s)LeS^{\gamma}(c,s)RS^{\delta}(c,s)SimS^{\epsilon}(c,s)NS^{\zeta}(c,s)$$

$$RuSimScore \in [0,1]$$

Where $\alpha, \beta, \gamma, \delta, \epsilon, \zeta$ are weights that allow to control importance of different aspects of similarity. In this work we selected the optimal weights which maximized SARI on the development dataset.

Cosine similarity score (SimS) is calculated as cosine similarity between sentence level embeddings of source and target sentences. We have chosen LASER embeddings[1].

Named entity preservation score (NS) aims to help with the weakness of the embeddings: if the language model modifies named entities in the text, the change in the embeddings may be very small, however the meaning of the text can become very different. To calculate named entity score, we extract all named entities from both source and target sentences using Natasha library and calculate how many entities from target are also present in source. While matching entities we only require that one of the entity's words is matched (so that entities themselves can also be simplified like: Оскар Александрович Энгберг $\rightarrow$ Энгберг). If count of entities is greater than 3, then 3 matching entities are considered enough and NS is set to 1.0.

$$NS = \frac{min(3, |NER(c) \cap NER(s)|)}{min(3, |NER(c) \cup NER(s)|)}$$

Lexical complexity score (LS) can judge if the words used in the target sentence are more common in language (which corresponds to higher usage frequency in corpus). Score is calculated as:

$$LS = 1 + \alpha_{LS}\frac{\sum_{i=1}^{N} log(f_i)}{N} + \beta_{LS} min(log(f_i))$$

where f_i - frequency of i-th word. Unknown words, named entities, pronouns and numbers are excluded from the calculation. Lexical complexity score consists of average log frequency with weight α_{LS} and most rare word's log frequency with weight βLS. Word frequency data was taken from [5][2].

Dependency tree depth score (DS) aims to measure syntactical complexity of the target sentence. The syntax tree of the sentence is created using Natasha library[3]. Dependency tree score DD = 1.0 for depth of 1 or 2, DD = 0.9 for depth 3, DD = 0.7 for depth 4 and DD = 0.5 for larger depths.

Length score (LeS) helps to choose target sentences than are shorter than the source one (in terms of word count), but not too short:

$LeS(c, s) = 0.5$ if $WC(s) > WC(c)$

$LeS(c, s) = 1 - \frac{WC(s)}{2WC(c)}$ if $WC(s) > 6$ and $WC(s) \leq WC(c)$

$LeS(c, s) = \frac{WC(s)}{6}$ if $WC(s) \leq 6$ and $WC(s) \leq WC(c)$

where $WC(x)$ is word count in sentence x.

Reading ease score (RS) is an implementation of Flesch reading ease with coefficients for Russian language [17], mapped into [0.5,1] range:

$$RS = 0.75 + 0.25\frac{max(-100, min(100, 206.835 - 1.52WPS - 65.14\frac{SC}{WC}))}{100}$$

where WPS is number of words per sentence, SC is syllable count and WC is word count.

See Table 1. for examples of scoring function values.

3.2 Generative model

We used ruGPT-3, a GPT-3 implementation by SberBank AI[4]. This language model has received high score on Russian SuperGLUE benchmark[18] and is capable of solving various tasks.

During fine-tuning, the training set is inputted into the LM in the following format:

<s>Original sentence <Simplify:> Target sentence </s>

During inference, for source sentence we provide the following prompt:

<s>Original sentence <Simplify:>

and expect LM to output the simplified sentence and </s> token.

We use top-p sampling[4] to increase fluency and variance of the generated sequences, and generate 10-100 sentences for each original sentence. After that we calculate the scoring function for each of the generated sentences and select the best ones.

Language models are initially trained on next token prediction objective. As a result, sometimes text continuation is generated instead of simplification. To counter this, we had to add checking if the generated sentence starts with a pronoun or a determiner.

We use HuggingFace Transformers implementation to fine-tune the model and generate text[9].

4 Datasets

Machine translated WikiLarge dataset was provided by the competition organizers. We selected 37,884 samples from 246,978 samples of WikiLarge which were filtered by following condition: ES $\geq$ 0.65, SimS $\geq$ 0.75, RS $\geq$ 0.6, LeS $\geq$ 0.55, LS $\geq$ 0.65, DS $\geq$ 0.5. These conditions were based on the statistics of the scoring functions on the RSSE dev dataset (so quality of selected samples was on the same level as quality of RSSE dev samples).

For examples of the selected and removed samples see Table 3.

[2] https://github.com/hermitdave/FrequencyWords

[3] https://github.com/natasha/natasha

[4] https://github.com/sberbank-ai/ru-gpts

Таблица 1: Examples of scoring function values

Original sentence: Положение стало угрожающим для царевича, когда Филипп женился в седьмой раз — на знатной македонянке Клеопатре.

Text	RuSimScore	DS	LS	LeS	RS	ES	SimS
Положение стало угрожающим для царевича, когда Филипп женился в седьмой раз — на знатной македонянке Клеопатре.	0.21	0.90	0.69	0.50	0.70	1.00	1.00
Наследник престола не был в восторге от этого брака. Он был счастлив жениться на красивой македонянке, но жениться на египтянке.	0.08	0.90	0.82	0.50	0.85	0.50	0.60
Вскоре после этого царевич Филипп женился на Клеопатре.	0.16	1.00	0.79	0.67	0.80	0.60	0.71
Филипп женился в седьмой раз, на македонянке Клеопатре.	0.32	0.90	0.81	0.67	0.84	1.00	0.86

Table 2: Datasets

Dataset	Reference samples	Simplification samples
RSSE dev	1000	3574
WikiLarge filtered	37884	37884
RSSE public test	1000	3521
RSSE hidden test	1126	N/A

Таблица 3: Selected and dropped samples from WikiLarge

Selected samples:
В Голландии они назывались Stadspijpers, в Германии Stadtpfeifer и в Италии Pifferi. → Их называли Stadtpfeifer в Германии и Pifferi в Италии.
Иногда могут появляться оттенки красного и оранжевого, заменяя или смешиваясь с желтым в зависимости от подвида. → Иногда могут появляться оттенки красного и оранжевого.
Dropped samples:
Женева - второй по численности населения город Швейцарии (после Цюриха) и самый густонаселенный город Романди (франкоговорящая часть Швейцарии). → Он окружен двумя горными цепями - Альпами и Юрой.
Оливковое масло также используется в мыловарении и в качестве лампового масла. → Оливковое масло - это растительное масло.

The final model was trained on both RSSE dev and WikiLarge filtered dataset, its hyperparameters were chosen based on results on RSSE public test dataset, and the final score was obtained on RSSE hidden test dataset.

5 Results and analysis

In the competition, submissions were scored based on SARI[6]. This metric includes F1 score of add, delete and keep operations on n-gram level.

The evaluation results of the proposed model and the benchmarks are shown in Table 4. Iterative deletion is our implementation that is not using a language model; instead it iteratively removes the parts of syntax tree if it increases RuSimScore (partial implementation of [10]). The result of official benchmark is taken from public test.

When more target sentence candidates are generated with the language model, the best candidates are better in terms on SARI. However, generating too many candidate sentences causes drop in score. The scoring function is selecting too short simplifications in this situation. See Table 5.

Different language model sizes are compared in Table 6. Medium model is slightly better then the large one, and both perform better than the small model.

Ablation study of the scoring function is displayed in Table 7. All of the six functions that are included in RuSimScore appear to be useful.

Examples of controllable simplification are shown in Table 8. We can see the effect of modifying the weights of the scoring function. For example, increased SimS weight leads to more accurate but more complex answer, and increased RS weight leads to less accurate but very simple result.

Table 4: Results

Model	Hidden test SARI
ruGPT3 on filtered WikiLarge + RuSimScore	**39.28**
ruGPT3 on filtered WikiLarge	38.68
Official benchmark (mBART)	30.15
Iterative deletion with RuSimScore	32.40
First half of source text	30.33
Source text unchanged	11.04

Table 5: Generated sentence count

Count	Hidden test SARI
100	39.28
30	**39.39**
10	39.16
1	38.68

Table 6: LM size dependency

Model	Hidden test SARI
ruGPT3-small	38.89
ruGPT3-medium	**39.34**
ruGPT3-large	39.28

Table 7: Ablation study of scoring function

Model	Hidden test SARI
Original RuSimScore	**39.28**
RuSimScore - SimS	37.22
RuSimScore - NS	38.94
RuSimScore - LS	39.03
RuSimScore - LeS	38.91
RuSimScore - DS	39.27
RuSimScore - RS	39.11

Таблица 8: Controllable simplification examples

Original sentence	Архимандрит Дионисий торопил ополчение поспешить к Москве и направил князю Трубецкому просьбу объединиться со Вторым ополчением.
Best model (balanced)	Архимандрит Дионисий сказал князю Трубецкому торопиться к Москве.
More accurate (increased SimS and NS weight)	Архимандрит Дионисий призвал ополчение поспешить к Москве и попросил князя Трубецкого объединиться с ними.
Simpler (increased RS, LS, LeS, DS weight)	Архимандрит Дионисий был в Москве и просил войска помочь ему

6 Error analysis

Neural sequence to sequence models sometimes generate factually inconsistent abstractive summaries[8]. Sentence simplification task is related to abstractive summarization, so we can expect similar issues in our model.

We performed a human evaluation of 100 randomly selected simplifications produced by the model. We asked the assessors to evaluate: 1) are the generated simplifications well formed Russian sentences and 2) are they factually consistent with the originals. We evaluated two models: one with default scoring function weights (according to highest SARI) and one with increased SimS and NS weights (to improve accuracy). The results of the evaluation (Table 9) show that 20-30% of the sentences have issues with factual consistency. Examples of the detected issues are shown in Table 10.

Table 9: Evaluation of 100 generated simplified sentences

	Original weights	Better accuracy
Well formed sentences	90	93
Not well formed sentences	10	7
Factually consistent	71	81
Issues with factual consistency, including:	29	19
Omitting significant details	15	7
Wrong entity reference	8	6
Incorrect replacement word used	6	6

Таблица 10: Examples of factual inconsistency

Original sentence	Влияние организованной преступности, наркосиндикатов, мафиозных структур, в особенности — в южных регионах — Кампании, Калабрии и на Сицилии.
Omitting significant details	На юге Кампания, Калабрия и Сицилия.
Original sentence	Лейси активно начал поединок, прессингуя и зажимая Джонса у канатов ринга, но Рой умело уходил от ударов и большинство блокировал.
Wrong entity reference	Рой атаковал Джонса и пытался его прессинговать, но Рой успешно отбивался.
Original sentence	Летописцы указывали, что кони уже не могли не ступать по трупам, так как не было чистого места.
Incorrect replacement word used	Летописцы писали, что кони уже не могли не пачкать трупы.

We also performed human evaluation of the scoring function. For 20 examples we provided 5 different simplifications and asked humans to order them from the most preferred one to the least preferred one. Spearman correlation between ranks of simplifications assigned by human and by the scoring function is 0.53 which indicates moderate correlation. Sample size was small, so this metric needs further evaluation on a larger scale.

7 Conclusion

The proposed model was able to produce higher results than the benchmarks, however the simplifications produced by the model may be inaccurate and therefore require manual verification. Selecting the best generated sentence with the scoring function improves the result, and scoring function is correlated with human perception. A simple iterative deletion approach, guided by the scoring function, was able to outperform the official benchmark. For future research, the scoring function can be used in different setup, for example fine-tuning with reinforcement learning.

References

[1] Artetxe Mikel, Schwenk Holger. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. — 2019. — 1812.10464.

[2] Automated Text Readability Assessment for Russian Second Language Learners / A. Laposhina, T. Veselovskaya, M. Lebedeva, O. Kupreshchenk // Proceedings of the international conference Dialogue 2018. — 2018.

[3] Martin Louis, Sagot Benoît, Éric de la Clergerie, Bordes Antoine. Controllable Sentence Simplification. — 2020. — 1910.02677.

[4] Holtzman Ari, Buys Jan, Du Li et al. The Curious Case of Neural Text Degeneration. — 2020. — 1904.09751.

[5] Dave Hermit. Github: Repository for Frequency Word List Generator and processed files. c2016. — 2016.

[6] EASSE: Easier Automatic Sentence Simplification Evaluation / Fernando Alva-Manchego, Louis Martin, Carolina Scarton, Lucia Specia // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 49–54. — Access mode: https://www.aclweb.org/anthology/D19-3009.

[7] EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing / Yue Dong, Zichao Li, Mehdi Rezagholizadeh, Jackie Chi Kit Cheung // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 3393–3402. — Access mode: https://www.aclweb.org/anthology/P19-1331.

[8] Cao Ziqiang, Wei Furu, Li Wenjie, Li Sujian. Faithful to the Original: Fact Aware Neural Abstractive Summarization. — 2017. — 1711.04434.

[9] Wolf Thomas, Debut Lysandre, Sanh Victor et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. — 2020. — 1910.03771.

[10] Iterative Edit-Based Unsupervised Sentence Simplification / Dhruv Kumar, Lili Mou, Lukasz Golab, Olga Vechtomova // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 7918–7928. — Access mode: https://www.aclweb.org/anthology/2020.acl-main.707.

[11] Ivanov V., Solnyshkina M., Solovyev V. Efficiency of Text Readability Features in Russian Academic Texts // Proceedings of the international conference Dialogue 2018. — 2018.

[12] Kuvshinova T. Sentence Compression for Russian: Dataset and Baselines // Proceedings of the international conference Dialogue 2020. — 2020.

[13] Brown Tom B., Mann Benjamin, Ryder Nick et al. Language Models are Few-Shot Learners. — 2020. — 2005.14165.

[14] Language Models are Unsupervised Multitask Learners / Alec Radford, Jeff Wu, Rewon Child et al. — 2019.

[15] Learning How to Simplify From Explicit Labeling of Complex-Simplified Text Pairs / Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold et al. // Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — Taipei, Taiwan : Asian Federation of Natural Language Processing, 2017. — Nov. — P. 295–305. — Access mode: https://www.aclweb.org/anthology/I17-1030.

[16] Martin Louis, Fan Angela, Éric de la Clergerie et al. Multilingual Unsupervised Sentence Simplification. — 2020. — 2005.00352.

[17] Oborneva Irina. Mathematical model of the estimation of educational texts[Matematicheskaja model' ocenki uchebnyh tekstov.] // Proceedings of the 15th international conference on Information Technologies in Education (ITO-2005) [Materialy XV Mezhdunarodnoj konferencii-vystavki Informacionnye tehnologii v obrazovanii (ITO-2005)]. — Moscow, 2005.

[18] Shavrina Tatiana, Fenogenova Alena, Emelyanov Anton et al. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. — 2020. — 2010.15925.

[19] Sakhovskiy Andrey; Izhevskaya Alexandra; Pestova Alena; Tutubalina Elena; Malykh Valentin; Smurov Ivan; Artemova Ekaterina. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[20] Witteveen Sam, Andrews Martin. Paraphrasing with Large Language Models // Proceedings of the 3rd Workshop on Neural Generation and Translation. — Hong Kong : Association for Computational Linguistics, 2019. — Nov. — P. 215–220. — Access mode: https://www.aclweb.org/anthology/D19-5623.

[21] Zhang Xingxing, Lapata Mirella. Sentence Simplification with Deep Reinforcement Learning // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — Copenhagen, Denmark : Association for Computational Linguistics, 2017. — Sep. — P. 584–594. — Access mode: https://www.aclweb.org/anthology/D17-1062.

RuShiftEval: a shared task on semantic shift detection for Russian

Pivovarova Lidia
University of Helsinki
Finland
lidia.pivovarova@helsinki.fi

Kutuzov Andrey
University of Oslo
Norway
andreku@ifi.uio.no

Abstract

We present the first shared task on diachronic word meaning change detection for the Russian. The participating systems were provided with three sub-corpora of the Russian National Corpus — corresponding to pre-Soviet, Soviet and post-Soviet periods respectively — and a set of approximately one hundred Russian nouns. The task was to rank those nouns according to the degrees of their meaning change between periods.

Although RuShiftEval is in many respects similar to the previous tasks organized for other languages, we introduced several novel decisions that allow for using novel methods. First, our manually annotated semantic change dataset is split in more than two time periods. Second, this is the first shared task on word meaning change which provided a training set.

The shared task received submissions from 14 teams. The results of RuShiftEval show that a training set could be utilized for word meaning shift detection: the four top-performing systems trained or fine-tuned their methods on the training set. Results also suggest that using linguistic knowledge could improve performance on this task. Finally, this is the first time that contextualized embedding architectures (XLM-R, BERT and ELMo) clearly outperform their static counterparts in the semantic change detection task.

Keywords: semantic change detection, Russian, shared task
DOI: 10.28995/2075-7182-2021-20-533-545

RuShiftEval: соревнование по детектированию семантических сдвигов в русском языке

Пивоварова Лидия
Университет Хельсинки
Финляндия
lidia.pivovarova@helsinki.fi

Кутузов Андрей
Университет Осло
Норвегия
andreku@ifi.uio.no

Аннотация

Мы представляем первую дорожку по автоматическому определению изменения значений слов для русского языка. Участники дорожки получили три подкорпуса НКРЯ - досоветский, советский и постсоветский - и список из около ста русских существительных. Задача состояла в ранжировании этих слов по степени семантического сдвига между этими периодами.

Наша дорожка во многих отношениях похожа на предыдущие подобные соревнования, которые организовывались для других языков. Однако мы предложили несколько нововведений, которые позволили участникам протестировать новые подходы к этой задаче. Во-первых, мы опубликовали новый датасет, в котором данные разбиты более чем на два периода. Во-вторых, это первая дорожка по автоматическому определению семантических сдвигов, где участникам был предоставлен тренировочный набор данных.

Дорожка получила более сотни решений от четырнадцати участников. Результаты соревнования продемонстрировали полезность тренировочных данных для определения семантических сдвигов: четыре лучших результа были продемонстрированы моделями, которые тренировались или донастраивались на тренировочных данных. Результаты так же демонстрируют, что использование априорных лингвистических знаний или сложных языковых моделей улучшают показатели в этой задаче.

Ключевые слова: диахронические семантические сдвиги, детектирование семантических изменений

1 Introduction

Words change their semantics over time as a result of combination of various processes that affect language simultaneously. Automatic detection and measuring the degree of meaning change could accelerate research in the history of language and also support a number of text analysis tasks such as information retrieval or media monitoring.

The RuShiftEval shared task is aimed at the comparison of various methods for detection of word meaning shift from diachronic corpora. Recently, two shared tasks for semantic change detection were organized: SemEval Task 1 for English, German, Swedish and Latin [17], and DIACR-Ita for Italian [2]. RuShiftEval is the first attempt to organize such an event with Russian data.

In many aspects, we follow the practices established during the previous shared tasks. However, we introduced several novelties: first, we deal with *three* time periods, namely pre-Soviet, Soviet and post-Soviet; second, we provided the participants with a *training dataset*, thus allowing for using supervised methods.

The shared task is collocated with Dialogue 2021, the 27th International Conference on Computational Linguistics and Intellectual Technologies. The test and development datasets used in RuShiftEval are now publicly available, as well as the evaluation code and the baseline.[1]

2 Related work

Automatic detection of word meaning change is a fast developing research area. The majority of modern approaches utilize distributional *word embeddings* to detect changes in word context over time. Overview of various approaches for this task could be found in the recent surveys [18, 4, 22].

To perform numerical evaluation, the problem is most commonly formulated as following: an input is *a corpus* split into several (usually two) time periods and *a set of words*; the task is *to rank* these words according to the degree of meaning change they have undergone between the periods. The performance is measured by rank correlation between a produced ranking and the gold manually created ranking. Alternatively, the task could be cast as binary classification of words into changed and not-changed classes. In this case, evaluation is also done as comparison against manual annotation.

Thus, manually annotated datasets are key components for development of lexical semantic change models. Since word meaning shift is a *lexicon-level phenomenon*, annotation should take into account many word usages from each periods, making it a time-consuming task. The most recent DUReL framework solves this by annotating pairs of sentences and then computing an averaged metric that generalizes these annotations [16]. We follow this approach in our shared task.

The first shared task on word meaning change detection was organized in 2020 as a part of SemEval conference (SemEval 2020 Task 1). The shared task [17] provided datasets for four languages — English, German, Swedish, and Latin — with several dozens manually annotated words for each language. The task included two subtasks, described above: binary classification and ranking. More than twenty teams participated in it. One of the main results of SemEval 2020 Task 1 was that type-base (static) embeddings are more suitable for *unsupervised* semantic shift detection than more recent contextualized embeddings currently dominating almost all other NLP tasks. Another important observation is a high variety across corpora: a method that yields the best performance for one corpus may not be the best for another one. Another shared task was organized for Italian [2], where the task was binary classification, and the results largely replicated those from the SemEval.

Although RuShiftEval is the first shared task on word meaning change for Russian, semantic shift detection methods have been previously applied to this language, e.g. in [10, 20]. This research is accelerated by publishing of time-specific sub-corpora of the Russian National Corpus (RNC), consisting of sentences from the texts created in the pre-Soviet, Soviet and post-Soviet time periods. Together they cover nearly full RNC.[2] It is important to note that the RuShiftEval organizers are fully aware that 1)

[1] https://github.com/akutuzov/rushifteval_public

[2] The sentence-shuffled version of the RNC split into 3 sub-corpora corresponding to the RuShiftEval time periods was made freely available specifically for this shared task (it is required to sign a license agreement to get access to the corpora): https://rusvectores.org/static/corpora/

the division of Russian language history into these particular periods is not the only possible option and the boundaries could be drawn differently; 2) the RNC itself is not fully representative of the history of Russian. However, some decisions had to be made with respect to the time bin boundaries; the division we chose is at least motivated with regards to historical events and yields sub-corpora of comparable sizes. In the same vein, no Russian corpus other than the RNC is available which is large enough, covers long enough time span, and provides the creation dates for the texts.

These diachronic sub-corpora of the RNC have previously been already used to create the *RuSemShift* dataset [14], which includes two subsets, each of 70 words, manually annotated and ranked according to their change from pre-Soviet to Soviet and from Soviet to post-Soviet times respectively. For the RuShiftEval data annotation, we used the same corpora and followed the same annotation procedure, so *RuSemShift* could be used as a training set by task participants. However, two parts of the *RuSemShift* dataset use different sets of words, while for the shared task we use the same list of words for *all three periods*, in principle allowing to study continuous word sense dynamic across time.

3 Task overview

The shared task focuses on three time periods, naturally stemming from the history of the Russian language and society. The boundary years of 1917 and 1991 were omitted from the annotation due to their transitioning nature:

1. pre-Soviet (1700-1916);
2. Soviet (1918-1990);
3. post-Soviet (1992-2016).

The RuShiftEval dataset consists of 111 Russian nouns (99 in the test set and 12 in the development set), manually annotated with the degrees of their meaning change in three time period pairs:

1. between pre-Soviet and Soviet periods (so called *RuSemShift1* score);
2. between Soviet and post-Soviet periods (so called *RuSemShift2* score);
3. between pre-Soviet and post-Soviet periods (so called *RuSemShift3* score).

We did not rely on any assumption on the dependencies of these three scores and annotated all pairs independently. Note that the resulting RuShiftEval dataset (about 30 000 human judgments in total) is described in more detail in a separate paper [9], so it is only briefly presented here. As per reviewers' suggestions, we provide the full list of target words with their change scores in the Appendix (although we strongly recommend to use the maintained version in our GitHub repository).

The annotation was conducted using crowd-sourcing (Yandex.Toloka platform). It followed the DuReL workflow described in [16]. An annotator had to read and score two sentences containing a target word and belonging to different time periods. The sentences were randomly sampled from the corresponding sub-corpora of the Russian National Corpus. The scores (from 1 to 4) grade semantic relatedness between the target word meanings in two sentences. The 1 score denotes 'the senses are unrelated', and the 4 score denotes 'the senses are identical'.

Then individual scores were accumulated into mean semantic relatedness between word usages from two different time periods; this measure is also known as COMPARE and was introduced in [16]. Basically, it reflects human judgments about such relatedness averaged across about 30 sentence pairs containing the target word. Thus, the lower is the score (the closer it is to 1), the stronger is the degree of semantic change. For each sentence pair, the score was in turn averaged across at least 3 human annotators.

As has been mentioned in Section 2, the *RuSemShift* dataset [14] could be used for training (or simply for sanity check in the *Practice* phase), and we encouraged participants to do this. To find out whether using training data actually helps semantic change detection was one of the purposes of the RuShiftEval shared task. We can now confirm that the answer is positive; see Section 5 for details.

We recommended the participants to use the RNC for their data-driven solutions, since this corpus has been used to annotate the data. They were free to employ any other linguistic sources, and some actually did; again, see Section 5. Submissions of the participants were processed, evaluated and ranked with the

help of Codalab platform.[3]

To help participants to start with the task, we also provided static word embeddings pre-trained on diachronic sub-corpora of the RNC, using the CBOW algorithm [11], with context window size 5 and vector size 300. Each model was published in two variants: trained on raw tokens and trained on lemmas with part of speech tags ('завод_NOUN', etc). These embeddings were used in the baseline solution, which was available as a part of the starting kit for the participants.

4 Evaluation workflow

The task was formulated as a ranking problem, similar to Subtask 2 of the SemEval 2020 Task 1 [17]: a set of Russian words should be ranked according to the strength of their meaning change. Thus, we did not make any binary decisions on whether a word has changed its meaning or not.

Importantly, it was one and the same set of words, for which the participants had to provide 3 semantic change scores per each word. The lower score meant a stronger change; the higher score meant a higher semantic similarity between word usages in different time periods, and thus a weaker change.

During the main *Evaluation* phase (February 22 - March 1, 2021), the participants were provided with a set of 99 target Russian words. For each word, they had to submit three non-negative values, corresponding to semantic change in the aforementioned time period pairs. These values were used to build 3 column-wise rankings: so called *RuSemShift1*, *RuSemShift2* and *RuSemShift3*. Since rank correlation was used as the evaluation metrics, the absolute numerical values of semantic change scores did not matter (only their relative ranks).

During the *Development* phase (February 1 - February 22, 2021), a small development set was provided (12 manually annotated Russian words), and the participants could submit their predictions to get a preliminary estimation of their system performance (no gold labels were openly published).

Before February 1, the shared task was in the *Practice* phase: the participants could submit predictions to the words from the *RuSemShift* test set [14]. This dataset was already public, so the true labels were known to everyone. This phase could be used to sanity check submission routines. There were only two time period pairs, each with its own set of words (this is how *RuSemShift* is built). We remind that in the *Development* and *Evaluation* phases, the participants had *one* set of words and *three* time period pairs.

Each participating team was able to submit up to 10 answers in the Evaluation phase, and up to 1000 answers in the Development phase. Submissions were evaluated using Spearman rank correlation between word ranking produced by a system and a gold ranking obtained in manual annotation. Thus, for each system we computed three correlations, for each of the time period pairs. The final ranking of the systems is based on averaging of the three scores.

5 Shared task results

In the Evaluation phase, we received submissions from 14 users (some of them in 4 different teams). Table 1 shows the performance of top submissions from each user or team (we give the name of the team by default or the name of the individual participant, if no team was associated with this submission). The teams are ranked by their average scores.

Some initial comments are due with regards to this table:

1. The baseline solution employed lemmatized diachronic embeddings trained on the Russian National Corpus[4] and the simple local neighborhood method from [5].
2. The differences between the first and the second best performing systems are not statistically significant according to the Fisher test; the differences between the second and the third systems are statistically significant at $p = 0.06$ for RuSemShift1 only. However, the differences between the top three systems and the rest of the submissions are all statistically significant.
3. Using median score instead of average score does not substantially change the ranking.

[3] https://competitions.codalab.org/competitions/28340

[4] These embeddings and diachronic corpora were available to all participants.

	Team	RuSemShift1	RuSemShift2	RuSemShift3	Mean	Type
1	**GlossReader**	0.781	**0.803**	**0.822**	**0.802**	token
2	**DeepMistake**	**0.798**	0.773	0.803	0.791	token
3	vanyatko	0.678	0.746	0.737	0.720	token
4	**aryzhova**	0.469	0.450	0.453	0.457	token
5	Discovery	0.455	0.410	0.494	0.453	token
6	**UWB**	0.362	0.354	0.533	0.417	type
7	dschlechtweg	0.419	0.373	0.383	0.392	type
8	jenskaiser	0.430	0.310	0.406	0.382	token
9	**SBX-HY**	0.388	0.281	0.439	0.369	type
	Baseline	0.314	0.302	0.381	0.332	type
10	svart	0.163	0.223	0.401	0.262	type
11	**BykovDmitrii**	0.274	0.202	0.307	0.261	token
12	fdzr	0.217	0.251	0.065	0.178	type

Table 1: Evaluation phase leaderboard (Spearman rank correlations). The Type column shows the type of the used distributional embeddings.

4. Bold names denote teams or individual participants who submitted papers with the description of their systems. For other participants, we rely on the contents of the 'Description' field in their Codalab submissions.
5. The DeepMistake team made several submissions of essentially the same system with varying hyperparameters; we show only the best one.
6. The SBX-HY team made a minor technical mistake, and their correlation scores were negative. Our opinion is that this does not undermine the developed system itself, so we show the absolute values in Table 1, and rank submissions accordingly.

5.1 Participating systems overview

Below, we give the descriptions of the participating systems. First, let us look at the submissions described in the submitted papers.

GlossReader [13] relied on the pretrained multilingual XLM-R language model [21]. On top of it, they trained a word sense disambiguation (WSD) system on English WSD datasets, using learned representations of sense definitions. Interestingly, this system shows excellent performance on Russian lexical semantic change data as well. Essentially, this participant reproduced the RuShiftEval annotation effort, replacing human judgments with the distances between XML-R contextualized embeddings of the target words. Additionally, a linear regression was trained on the *RuSemShift* dataset to convert vector distance values into relatedness scores (from 1 to 4).

DeepMistake [3] used the multilingual XLM-R as well, and also pre-trained on English WSD datasets, but without explicitly predicting senses. Similarly to **GlossReader**, they additionally fine-tuned this model on the *RuSemShift* using linear regression for mapping to relatedness scores.

aryzhova [15] tried both ruBERT [7] and ELMo contextualized embeddings.[5] Interestingly, in their experiments ELMo outperformed BERT. Note, however, that **aryzhova** system is different from **vanyatko** (described below) in that it does not fine-tune BERT or ELMo: instead, it calculates the average cosine similarity between target word embeddings (sometimes with the addition of the neighboring word

[5]The ELMo models for Russian were borrowed from the RusVectōrēs service.

tokens) in the sampled sentence pairs, reproducing the *APD* method from [8]. Another interesting experiment reported in the paper from this participant is using 'grammatical vectors' corresponding to the frequencies of 12 morphological forms of Russian nouns (6 cases and singular/plural forms). They report that the cosine similarities between such vectors calculated on different time bins improved the performance of relatedness score classifier (trained and evaluated on the *RuSemShift* dataset).

UWB [12] this team employed traditional 300-dimensional static word embedding (in particular, fastText). Orthogonal Procrustes and Canonical Correlation Analysis (CCA) were used for alignment, with CCA showing somewhat better results. The semantic change score was calculated as simple cosine similarity between word vectors across different time periods.

SBX-HY [6] again used static word embeddings, but in this case instead of post-training alignment, they relied on Temporal Referencing approach [19], successfully used for semantic change detection with other languages. In this approach, the target words are augmented with time period labels, and then one embedding model is trained on all available data. Hyper-parameters were selected based on the *RuSemShift* dataset. Interestingly, with the RuShiftEval data, Temporal Referencing barely managed to outperform the organizers' baseline, which is an interesting negative result.

BykovDmitrii [1] employed an interesting approach with lexical substitutes produced by the multilingual XLM-R as a masked language model. These substitutes were then clustered into senses and the divergence between clusters from different time periods was used as the semantic change score. This particular approach failed, but in the post-evaluation phase, the participant managed to significantly improve their result by skipping the clustering step and instead directly comparing bags of lexical substitutes (see more in their paper).

Now let us briefly describe the systems which did not submit papers, based on their descriptions in Codalab. **Vanyatko** employed the RuBERT model. They fine-tuned RuBERT with sentence pairs as inputs and relatedness scores (from 1 to 4) as outputs. Similar to **GlossReader** and **DeepMistake**, **vanyatko** tried to reproduce human annotation process. The **Discovery** team used BERT with ensemble of Average Pairwise Cosine Distance and Cosine Distance of averaged embeddings. **Dschlechtweg** trained regression on the labeled training examples from *RuSemShift* with SGNS embeddings. **Jenskaiser** also employed static SGNS embeddings and Temporal referencing or 'word injection (WI)'. They got results very similar to **SBX-HY**. Finally, **svart** used orthogonal Procrustes and cosine distances with the lemmatized word2vec embeddings provided by the organizers, and **fdzr** again relied on temporal referencing.

6 Discussion

We believe the results of the RuShiftEval are interesting for the lexical semantic change detection field in at least four aspects.

1 This is the first time the systems based on *contextualized embeddings* top the leaderboard. In both SemEval 2020 Task 1 [17] and DIACR-ITA [2], type embedding (or 'static' embedding) based architectures clearly won the rankings. But at the RuShiftEval, five top performing systems use pre-trained contextualized ('token-based') models: XLM-R, BERT and ELMo. In the previous work, the researchers in the field expressed doubts about the abilities of token embeddings with relation to semantic change detection. It seems that at least in the case of RuShiftEval, they are perfectly able to solve the task better than their static counterparts. However, the best performing teams introduced completely novel approaches to the problem, so the distinction between our results and results of the previous tasks lies in the difference between models rather than between embeddings themselves.

2 Surprisingly, the first and the second best submissions relied on the contextualized XLM-R model [21], which was not even specifically trained for processing Russian data. Its training corpus included texts in about 100 languages. Russian is well represented there but is far from being the largest in absolute size. The results of our shared task show that multilingual models like XLM-R can be very

successfully applied to semantic change detection for Russian (and arguably for many other languages): their transferability is extremely high.

Interestingly, at the SemEval 2020 Task 1, the attempts to use XLM-R did not end up very well: the system based on it ended up 7th in the Subtask 2 (closest to RuShiftEval), well below the type-based architectures. One of the reasons for this can be the next insightful outcome of RuShiftEval:

3 Using training data helps lexical semantic change detection. As already said, the *RuSemShift* dataset [14] was publicly available by the beginning of RuShiftEval, and the participants were free to use it as they saw fit. The annotation procedures were identical for *RuSemShift* and the shared task test sets. Thus, one of the aims of RuShiftEval was to find out whether using previously annotated data can improve the performance of semantic change ranking. As it turns out, it definitely can. Four top systems all train or fine-tune on *RuSemShift*. This was the first semantic change detection shared task to introduce such a setup. At the same time, using unsupervised methods with parameters fine tuning on the training set does not seem to be a productive strategy.

4 Finally, at least two participants (both in the top of the leaderboard) used explicit linguistic knowledge in addition to statistical distributional models. In particular, **GlossReader** (the winner of the task) fine-tuned their XLM-R model to select a definition (a gloss) from the WordNet, that is most appropriate for a particular target word occurrence [13]. Note that it was not the plain old classification: the model directly processed the definitions themselves as sequences of words. Another example is **aryzhova** who employed a linguistic intuition that semantic change is often linked to fluctuations in the frequency of different grammatical forms [15]. We believe using linguistic knowledge is an interesting direction for future development of the semantic change detection field.

It is important to note that the observations above are applicable only to the shared task setup used in RuShiftEval: that is, ranking words by the degree of semantic change estimated with the COMPARE measure calculated on human annotations conducted within the DUReL framework. Actually, many of the top-performing systems essentially reproduced the annotation process with large language models, which seems to be successful even though they could not know which particular sentences were sampled for manual annotation. With other evaluation setups, different approaches could be at the top. As an example, it is known that the COMPARE measure is much influenced by sense frequencies and can easily overlook changes occurring to rare senses — either their appearance or disappearance. If the systems were evaluated based on explicit senses they managed to detect, clustering-based approaches would arguably rank much higher.

7 Conclusion

In this paper, we summarized the outcome of RuShiftEval: the first shared task on lexical semantic change detection for Russian. The purpose of the shared task was twofold: first, to evaluate current state-of-the-art methods in semantic change detection on Russian data, and second, to explore the possibilities of *supervised* semantic change detection. This was ensured by the prior existence of *RuSemShift* dataset, annotated in exactly the same way as our testing data.

The results of the shared task show that training on existing semantic change data is indeed useful and can significantly boost evaluation scores. In absolute values, the correlations with human judgments achieved by the RuShiftEval participants are much higher than those demonstrated in the SemEval 2020 Task 1 across English, Latin, German and Swedish (the best system there yielded 0.527). Note that although *RuSemShift* (used as a training set) and RuShiftEval (used as a development and a test set) are annotated similarly, they are not splits of one and the same dataset. Thus, we believe this finding to be reliable and expect it to hold for other languages as well.

Another interesting outcome of RuShiftEval is the strong victory of contextualized (token-based) embedding architectures over static (type-based) ones. This is different from the results of previous shared tasks on semantic change detection, and we believe this means the community has finally learned how to properly use contextualized embeddings for this task. This is even more impressive considering the fact that the winning systems used the multilingual XLM-R instead of a Russian-specific model.

Despite these substantial findings, our shared task has just started to pave the way for studying approaches to automatic semantic change detection in Russian. Our evaluation setup (ranking by aggregated COMPARE score) cannot capture the entire spectrum of semantic change. This linguistic phenomenon is extremely complex, and we are hoping that future shared tasks will try to account for that.

Acknowledgments

The annotation effort for this shared task was supported by the Russian Science Foundation grant 20-18-00206. We are especially grateful to Valery Solovyev (Kazan Federal University). This work has been partially supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

References

[1] Arefyev Nikolay, Bykov Dmitrii. An Interpretable Approach to Lexical Semantic Change Detection with Lexical Substitution // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[2] DIACR-Ita@ EVALITA2020: Overview of the EVALITA2020 diachronic lexical semantics (DIACR-Ita) task / Pierpaolo Basile, Annalina Caputo, Tommaso Caselli et al. // Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org. — 2020.

[3] DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model / Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov et al. // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[4] Diachronic word embeddings and semantic shifts: a survey / Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, Erik Velldal // Proceedings of the 27th International Conference on Computational Linguistics. — 2018. — P. 1384–1397.

[5] Hamilton William L., Leskovec Jure, Jurafsky Dan. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. — Austin, Texas : Association for Computational Linguistics, 2016. — Nov. — P. 2116–2121. — Access mode: https://www.aclweb.org/anthology/D16-1229.

[6] Hengchen Simon, Viloria Kate, Indukaev Andrey. SBX-HY at RuShiftEval 2021: Doveriay, no proveriay // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[7] Kuratov Yury, Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2019. — Access mode: http://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf.

[8] Kutuzov Andrey, Giulianelli Mario. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection // Proceedings of the Fourteenth Workshop on Semantic Evaluation. — Barcelona (online) : International Committee for Computational Linguistics, 2020. — Dec. — P. 126–134. — Access mode: https://www.aclweb.org/anthology/2020.semeval-1.14.

[9] Kutuzov Andrey, Pivovarova Lidia. Three-part diachronic semantic change dataset for Russian // Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change. — online : Association for Computational Linguistics, 2021.

[10] Measuring Diachronic Evolution of Evaluative Adjectives with Word Embeddings: the Case for English, Norwegian, and Russian / Julia Rodina, Daria Bakshandaeva, Vadim Fomin et al. // Proceedings of the 1st International Workshop on Computational Approaches to Historical Language

Change. — Florence, Italy : Association for Computational Linguistics, 2019. — Aug. — P. 202–209. — Access mode: https://www.aclweb.org/anthology/W19-4725.

[11] Mikolov Tomas, Yih Wen-tau, Zweig Geoffrey. 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. — Atlanta, Georgia : Association for Computational Linguistics, 2013. — Jun. — P. 746–751. — Access mode: https://www.aclweb.org/anthology/N13-1090.

[12] Priban Pavel, Pražák Ondřej, Taylor Stephen. UWB@RuShiftEval: Measuring Semantic Difference as per-word Variation in Aligned Semantic Spaces // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[13] Rachinskiy Maxim, Arefyev Nikolay. Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[14] Rodina Julia, Kutuzov Andrey. RuSemShift: a dataset of historical lexical semantic change in Russian // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 1037–1047. — Access mode: https://www.aclweb.org/anthology/2020.coling-main.90.

[15] Ryzhova Anastasiia, Ryzhova Daria, Sochenkov Ilya. Detection of Semantic Changes in Russian Nouns with Distributional Models and Grammatical Features // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference. — 2021.

[16] Schlechtweg Dominik, Schulte im Walde Sabine, Eckmann Stefanie. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). — New Orleans, Louisiana : Association for Computational Linguistics, 2018. — Jun. — P. 169–174. — Access mode: https://www.aclweb.org/anthology/N18-2027.

[17] SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection / Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen et al. // Proceedings of the Fourteenth Workshop on Semantic Evaluation. — Barcelona (online) : International Committee for Computational Linguistics, 2020. — Dec. — P. 1–23. — Access mode: https://www.aclweb.org/anthology/2020.semeval-1.1.

[18] Tang Xuri. A state-of-the-art of semantic change computation // Natural Language Engineering. — 2018. — Vol. 24, no. 5. — P. 649–676.

[19] Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change / Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Florence, Italy : Association for Computational Linguistics, 2019.

[20] Tracing cultural diachronic semantic shifts in Russian using word embeddings: test sets and baselines / Vadim Fomin, Daria Bakshandaeva, Julia Rodina, Andrey Kutuzov // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2019. — P. 203–218. — Access mode: http://www.dialog-21.ru/media/4598/fominvplusetal-116.pdf.

[21] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 8440–8451. — Access mode: https://www.aclweb.org/anthology/2020.acl-main.747.

[22] A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains / Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, Sabine Schulte im Walde // Pro-

ceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 732–746. — Access mode: https://www.aclweb.org/anthology/P19-1072.

A RuShiftEval gold datasets

1. 1-2: change from the pre-Soviet to Soviet times;
2. 2-3: change from the Soviet to the post-Soviet times;
3. 1-3: change from the pre-Soviet to the post-Soviet times.

DEVELOPMENT SET					
WORD	**TRANSLITERATION**	**TRANSLATION**	**CHANGE SCORE**		
			1-2	**2-3**	**1-3**
верховье	verhovje	upper reaches	3.68	3.74	3.87
возраст	vozrast	age	3.47	3.69	3.58
завод	zavod	factory/breeding farm	3.22	3.65	3.52
закладка	zakladka	foundation/bookmark/hidden artifact	1.93	1.74	1.74
земля	zemlja	earth/land/soil	2.83	2.8	2.28
лох	loh	salmon/silver-berry/easy victim	1.07	2.94	1.04
помощник	pomoštšnik	assistant	3.38	3.56	3.28
пролетарий	proletarij	proletarian	3.4	3.58	3.44
промышленность	promyšlennost’	industry	3.24	3.51	3.47
радикал	radikal	radical	1.42	1.68	2.01
спутник	sputnik	fellow traveler/satellite/sputnik	2.96	1.81	1.94
четверть	tšetvert	quarter	2.25	2.96	3.07

TEST SET					
WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
авторитет	avtoritet	authority/prestige	3.23	2.95	2.84
амбиция	ambitsia	ambition	3.11	3.44	3.33
апостол	apostol	apostle/disciple	3.49	3.42	3.42
благодарность	blagodarnost’	gratitude/appreciation/thankfulness	3.23	3.56	3.65
блин	blin	pancake/damn	3.21	1.66	2.57
блондин	blondin	blonde (male)	3.94	3.92	3.95
брат	brat	brother	3.22	3.01	3.27
бригада	brigada	brigade/gang/team	2.8	2.71	3.08
веер	vejer	fan	2.55	2.43	2.44
век	vek	century/age	3.2	3.21	2.98
вызов	vyzov	call/challenge/summons	2.17	2.1	2.03
головка	golovka	(small) head	2.20	1.67	2.19
грех	greh	sin/fault	3.48	2.98	2.92
дух	duh	spirit/ghost/scent	2.32	1.63	1.88
дядька	djadka	uncle/man/(male) tutor	2.59	3.03	2.68
дядя	djadja	uncle/man	3.37	3.39	3.29
железо	železo	iron	2.2	2.56	2.40
жесть	žest	tin/horror	3.23	3.38	3.41
живот	život	stomach/belly/life	2.91	3.44	2.76
заблуждение	zabluždenije	delusion	3.5	3.62	3.55
издательство	izdatelstvo	publishing house	3.53	3.86	3.45
итальянец	italjanets	Italian	3.70	3.6	3.67
кабан	kaban	boar	3.6	3.32	3.30
карман	karman	pocket	3.46	3.47	3.56
крушение	krušenije	collapse	2.75	2.78	2.6
крыша	kryša	roof	3.57	3.0	2.82
кулиса	kulisa	wings	3.16	3.17	3.24
лечение	letsenije	cure	3.65	3.74	3.68
линейка	lineika	carriage/ruler/series of goods	1.87	1.37	1.22
лишение	lišenije	deprivation	2.94	2.07	2.33
локоть	lokot	elbow	3.27	3.41	3.73
любовник	ljubovnik	lover	3.45	3.71	3.65
любовь	ljubov	love	3.29	2.97	3.07
маньяк	manjak	maniac	3.08	3.01	3.11
монстр	monstr	monster	2.6	2.38	2.04
наволочка	navolotška	pillowcase	3.61	3.83	3.92
название	nazvanije	name/title	3.48	3.48	3.43

WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
наложение	naloženije	imposition	1.95	2.06	1.78
облако	oblako	cloud	3.17	3.0	3.16
обоснование	obosnovanije	grounds	3.74	3.5	3.58
огонь	ogon	fire	2.10	2.13	2.46
памятник	pamjatnik	monument	2.88	2.83	2.82
пафос	pafos	pathos	3.34	3.27	3.41
писк	pisk	squeak	3.21	3.0	2.53
план	plan	plan	2.67	2.27	2.54
поколение	pokolenie	generation	3.43	3.58	2.8
половинка	polovinka	half	2.51	2.75	2.62
полоса	polosa	stripe/ribbon/lane/runway	1.83	1.5	1.41
полость	polost	cavity/foot hide	2.23	1.88	2.56
полукруг	polukrug	semicircle	2.78	3.13	3.08
понедельник	ponedelnik	Monday	3.77	3.86	3.86
поставщик	postavštšik	supplier	3.56	3.44	3.25
поэзия	poezia	poetry	3.22	3.66	3.56
правда	pravda	truth/reality	3.13	2.94	2.96
предательство	predatelstvo	betrayal	3.67	3.48	3.8
прецедент	pretsedent	precedent	3.52	3.8	3.53
проникновение	proniknovenije	penetration	2.75	2.68	2.53
прорыв	proryv	breakthrough	2.08	2.05	2.05
путь	put’	way	2.41	2.04	2.3
размышление	razmyšlenije	reflection	3.52	3.55	3.62
ранец	ranets	backpack	3.6	3.53	3.38
расчет	rastšot	calculation/settlement	2.0	1.95	2.0
риторика	ritorika	rhetoric	3.06	2.95	2.93
роспись	rospis	mural/signature/list	1.43	2.98	1.57
сверстник	sverstnik	age-mate	3.86	3.86	3.82
связка	svjazka	ligament/vocal cords/mutual connection	2.33	1.96	1.77
собрат	sobrat	fellow	3.45	3.32	3.32
совершенство	soveršenstvo	perfection	2.95	3.16	3.08
советчик	sovettik	adviser	3.22	3.48	3.42
союзник	sojyznik	ally	3.66	3.47	3.75
список	spisok	list	3.28	3.31	3.05
ссылка	ssylka	exile/link	2.87	2.04	1.93
стена	stena	wall	3.1	3.16	3.32
стипендия	stipendia	scholarship	3.8	3.71	3.56

WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
стол	stol	table/diet	3.50	3.16	3.25
тачка	tachka	wheelbarrow/car	3.39	1.94	1.89
тупик	tupik	deadlock	3.17	2.83	3.14
увольнение	uvolnenie	furlough/layoff	3.21	3.53	3.32
углеводород	uglevodorod	hydrocarbon	3.68	3.31	3.2
удобство	udobstvo	convenience	2.43	2.42	2.51
уклад	uklad	setup	3.33	3.42	3.42
университет	universitet	university	3.54	3.7	3.72
установление	ustanovlenie	establishment	2.28	2.26	2.40
фаворит	favorit	favorite	3.15	2.53	2.84
формат	format	format	2.84	2.02	1.81
формула	formula	formula	2.81	2.26	2.57
хозяйка	hozjaika	hostess	3.25	3.22	3.42
хор	hor	choir	2.66	2.87	2.22
хрен	hren	horseradish/dick/old fart	1.8	2.26	1.6
цензура	tsenzura	censorship	3.49	3.46	3.45
центр	tsentr	center	2.14	1.83	1.87
цифра	tsifra	digit/number	2.96	2.87	3.19
частица	tšastitsa	part/particle	1.96	2.33	2.2
чек	tšek	check	2.37	1.95	2.65
штаб	štab	headquarters	3.63	3.38	3.5
эшелон	ešelon	echelon	2.92	2.28	2.33
юбилей	jubilei	anniversary/jubilee	3.68	3.7	3.78
ядро	jadro	cannonball/core/nucleus	1.55	1.91	1.47
ясли	jasli	nursery/manger	2.28	3.0	1.9

Semantics, Grammar and Prosody of parentheticals introduced by the Subordinator *kak* ‘as’

Podlesskaya V. I.
Russian State University for the Humanities, Moscow, Russia
vi_podlesskaya@il-rggu.ru

Pozhilov Ju. M.
Russian State University for the Humanities, Moscow, Russia
yuriko.pozhilov@yandex.ru

Abstract

Based on data from the multimedia subcorpus of the Russian National Corpus, the paper addresses syntactic, sematic and prosodic features of the particular type of quotations with the reporting frame headed by the subordinator *kak* ‘as’ (*kak skazal mne staryj rab pered tavernoj*…). Our data show mixed evidence regarding the parenthetical status of the construction. On the one hand, typically for parentheticals, its function is clearly pragmatized, since it expresses speaker’s attitude towards the quote. On the other hand, typical parentheticals have only loose syntactic connection with their “host”, while the *kak*-phrase is introduced by the subordinator and has the form of the standard adverbial clause. Further on, while typical parentheticals are characterized by grammatical and prosodic reduction, grammatical and prosodic restrictions operating in the *kak*-phrase are optional and context (e.g., word order) sensitive. The kind of data we present supports the approach to parenthesis that doesn’t favor either/or decisions, but rather is based on multifactorial analysis that considers the whole range of possible parameters and isolates their observed language-specific clusters.

Keywords: spoken discourse, Russian, prosody, parenthesis
DOI: 10.28995/2075-7182-2021-20-546-559

Семантика, грамматика и просодия вводно-союзных конструкций по данным мультимедийного подкорпуса НКРЯ

Подлесская В. И.
РГГУ, Москва, Россия
vi_podlesskaya@il-rggu.ru

Пожилов Ю. М.
РГГУ, Москва, Россия
yuriko.pozhilov@yandex.ru

Аннотация

В данной работе на материале мультимедийного подкорпуса НКРЯ рассмотрены синтаксические, семантические и просодические свойства авторской ремарки, вводящейся союзом «как» (*как сказал мне старый раб перед таверной…*). Наши данные показывают, что такие конструкции могут демонстрировать разную степень парентетичности. С одной стороны, они достаточно прагматикализованы, так как в их значении отражена позиция говорящего по отношению к цитате, что свойственно для парентез. С другой стороны, типичные парентезы теряют синтаксическую связь с «главной» клаузой, в отличие от *как*-фраз, которые вводятся союзом и имеют форму стандартной обстоятельной клаузы. Более того, в то время как стандартные парентезы характеризуются грамматической и просодической редукцией, в *как*-фразах грамматические и просодические изменения опциональны и контекстуально зависимы – в частности, они чувствительны к порядку слов. При работе с такими данными мы считаем нужным отказаться от бинарного противопоставления в пользу многофакторного анализа парентезы, позволяющего учитывать все многообразие исследуемой зоны.

Ключевые слова: устный дискурс, русский язык, просодия, парентеза

1 Постановка задачи

Предмет исследования в этой работе – конструкции, в которых авторская ремарка вводится в клаузе с союзом *как*[1]:

(1) *Как говорил Александр Иванович Герцен/ знакомство с иностранцем для русского человека/ это в некотором роде повышение в чине. [Владимир Меньшов, Марина Мареева. Зависть богов, к/ф (2000)]*[2]

Рассматриваемый нами класс конструкций относится к так называемым «вводно-союзным», согласно определению Е.В.Падучевой (1996: 321-334). В принципе, в вершине клауз, вводимых союзом *как* в данной функции, возможны глаголы разных семантических классов, например, глаголы мнения (*как считают британские ученые*), однако в данной работе мы ограничимся только глаголами речи (понимаемыми, впрочем, достаточно широко – как глаголы порождения текста, т.е. не только *говорить*, но и, например, *писать, учить, велеть* – в соответствии с номенклатурой семантических классов, принятой в НКРЯ). Мы постараемся показать, что *как*-ремарки на разных языковых уровнях – прагматическом, грамматическом и просодическом - демонстрируют симптомы парентезы, т.е. редукции, или ослабления их вклада в пропозициональное и истинностное значение высказывания; ср. сходные понятия «десентенциализации» (desententialization, Lehmann 1988), или «понижения в ранге» (deranking, Cristofaro 2003), в русистике понятие редукции успешно применяется к анализу вводных единиц Г.И.Кустовой (2020:18).

Ядро наших данных составляют примеры из Мультимедийного подкорпуса НКРЯ (МУРКО). Для инструментального изучения просодии из видеофайлов выдачи МУРКО извлекались аудиофайлы и подавались на вход анализатора PRAAT (Boersma, Weenink 2021). При анализе грамматики и семантики мы прибегаем также к данным из других подкорпусов НКРЯ, прежде всего, устного.

Изложение будет построено так: в разделе 2 мы остановимся на семантике конструкций с *как*-ремаркой, в разделе 3 обсудим их структурные свойства. В разделе 4 мы продемонстрируем основные паттерны, которым следует просодическая реализация конструкций с *как*-ремаркой. Эти три раздела не герметичны: с одной стороны, многие грамматические свойства этих конструкций объясняются их семантикой, а с другой – просодия неразрывно связана и с грамматикой, и семантикой – так, локализация и направление движения тона во фразовых акцентах чувствительны к порядку слов и референциальному статусу именных групп, обозначающих локуторов. В разделе 5 мы подведем итоги и попытаемся ввести полученные результаты в общий контекст исследований парентезы как лингвистического феномена.

2 Семантика конструкций с *как*-ремаркой

Как было убедительно показано Е.В.Падучевой (1996: 327-328), «вводно-союзная конструкция … используется для передачи чужого мнения, которое служит говорящему основанием для его собственного суждения», или для передачи чужого способа выражения, причем в этом случае «говорящий отстраняется от способа выражения, хотя присоединяется к самой мысли», а автор цитаты обычно бывает «авторитетным» (*как указывает академик Виноградов*). Наши корпусные данные показывают, что ссылка на авторитет, как в примере (1), – это частая, но не единственно возможная функция таких конструкций. Другой распространенный случай – это отсылка к высказыванию локуторов или иных, третьих лиц – участников беседы:

(2) *Я сам из Челябинска/ вот как Антон сказал/ из благополучного Челябинска [Ю.Б. Черкасов. Выступление на первомайском митинге в Новосинеглазово (2017)]*

[1] Исследование поддержано грантом РНФ 17-18-01184

[2] Примеры даются в том графическом формате, в котором они приводятся в соответствующем подкорпусе НКРЯ, за исключением случаев, где мы добавляем просодическую разметку. Эти случаи отдельно оговариваются.

Если в *как*-фразе вводится высказывание от первого лица, то возникает дополнительный оттенок смысла – говорящий оправдывает возможную нестандартность цитаты своим авторским приоритетом, а также тем, что это цитата вводится как уже апробированная:

(3) *Ой/ ну это тяжело всё. Я понимаю/ что я в своём возрасте щас открываю/ как я говорю/ хайло. [Валентина, жен, начальник отдела кадров] Ну это да! [Разговоры в офисе (2008) // Из коллекции НКРЯ]*

(4) *Сейчас приезжаешь куда-нибудь в Америку/ извините за снобизм/ как я говорю своим студентам/ в Гарварде идёшь там по стеллажам в хранении — наслаждение одно ходить по стеллажам [Михаил Булгаков. Программа "Гордон" (НТВ) (2003)]*

Как-фраза может отсылать и к массовому узусу, это особенно часто случается, если она строится с использованием возвратного пассива, безличного, неопределенно-личного или иного «анонимизирующего» формата (*как говорят у нас в Одессе, как поётся в известной песне*)[3]. Инвариантом отсылки к высказыванию авторитетного источника, отсылки к высказыванию участника беседы и к массовому узусу можно считать отсылку к такому высказыванию, которое, по мнению говорящего, имеется не только в его собственной базе знаний, но и в базе знаний иных членов языкового коллектива.

3 Грамматика конструкций с *как*-ремаркой

Базовое значение конструкции – предъявление некоторого выказывания как уже имеющегося в общем поле знаний – помогает объяснить некоторые грамматические ограничения на *как*-ремарку.

3.1 Ограничения на форму глагола

Имеются ограничения на грамматическую форму глагола речи в ремарке.

Это, во-первых, ограничения на грамматическое время. Глаголы совершенного вида имеют только форму прошедшего времени, глаголы несовершенного вида могут использоваться и в презенсе, передавая узуальное значение, значение одновременности с событиями главной клаузы (происходившими в прошлом) или значение praesens historicum, (5) (помимо того, что часть вхождений в форме настоящего времени обеспечивается упоминавшимися выше «анонимизирующими» формами):

(5) *Закончилось все тем, что в 1742 году «князя-черепаху» свалил апоплексический удар, и он скончался, как пишет современник, «от семейных неприятностей». [И. Грачева. Твердость металла и нежность цветка // «Наука и жизнь», 2007]*

Использование будущего времени оказывается практически заблокированным. Имеющиеся единичные примеры могут интерпретироваться как будущее в прошедшем:

(6) *То есть было безлюбовное отношение к земле/ не моя земля. Сегодня попользовался и отдам. Как потом скажет один русский философ/ военнопленная земля. [Юрий Пивоваров. Традиции русской государственности и современность. Проект Academia (ГТРК Культура) (2010)]*

Как-ремарки употребляются и во «вневременных» контекстах, например, в контекстах со снятой утвердительностью. В этих случаях типично употребление глагола речи в сослагательном наклонении[4]:

[3] «Анонимизирующие» конструкции и, в особенности, *как*-фразы с некоторыми глаголами речи в форме возвратного пассива склонны к прагматизации вплоть до превращения в дискурсивный маркер с хезитативным или аппроксимативным значением. Мы благодарны рецензенту, обратившему наше внимание на то, что в качестве такого дискурсивного маркера употребляется конструкция *как говорится.*

[4] Рецензент справедливо заметил, что в подобных контекстах допустимо и будущее время глагола. В примерах, которые приводит рецензент, снятая утвердительность подкрепляется и «анонимизирующими» местоимениями *некий, всякий*: *Но почему-то мне показалось сегодня, что этот розовато-золотой, дымчато-голубой зимний день требует*

(7) *И работа его над этим текстом закончилась — ну/ "автор умер"/ как сказал бы Ролан Барт. [Рождение художественного текста. Программа "Гордон" (НТВ) (2003)]*

Во-вторых, как показали проницательные наблюдения Е.В.Падучевой (1996: 327-328) и Г.И.Кустовой (2020:18) в *как*-ремарке – как и во многих вводных фрагментах – не допускается общее отрицание (**как не говорил Александр Иванович Герцен*).

В-третьих, *как*-ремарка, не являясь по своему синтаксическому статусу главной клаузой, не способна быть самостоятельным носителем иллокутивной силы. Более того, использование *как*-ремарки затруднено в составе вопросов и директивов, хотя и не запрещено полностью:

(8) *И неужели, как говорила Лазура, жизнь есть вечная перемена различных образов? [Н. П. Вагнер. Сказки Кота-Мурлыки (1872)]*

(9) *Так шо/ козаки/ вручить шашку Игорю Владимировичу/ а? Любо! Так шо/ как говорится/ прими в знак любви и уважения. [Леонид Барац, Олег Фомин, Ростислав Хаит, Сергей Петрейков. День выборов, к/ф (2007)]*

3.2 Ограничения на порядок слов

Важнейшим структурным ограничением, действующим в *как*-ремарке, является ограничение на порядок слов: если подлежащее ремарки (субъект речи = автор цитаты) выражено полной именной группой, то предпочтительна инверсия подлежащего и сказуемого. Это правило действует независимо от того, расположена ли ремарка до, после или внутри цитаты:

(10) *Как мне сказал знакомый астроном/ количество нейронных связей в мозгу больше/ чем звёзд во Вселенной [Татьяна Черниговская. Как мы мыслим. Разноязычие и кибернетика мозга. Лекции Полит.ру (2009)]*

(11) *В этом залог нашей победы. Как говорили шведы под Полтавой. [Владен Бахнов, Леонид Гайдай. Спортлото-82, к/ф (1982)]*

(12) *Вообще/ э/ мир/ как говорили греки/ начался из Хаоса/ но потом родил Космос. [Мир как вакуум. Программа "Гордон" (НТВ) (2001)]*

Если же подлежащее ремарки – местоимение, то сохраняется прямой порядок[5], ср. невозможность инверсии **как говорил я* в следующем примере:

(13) *Ведь Солнце/ как я говорил/ не единственная/ так сказать/ звезда/ есть вещество от других звёзд/ есть межпланетная плазма. [Лев Зеленый. Солнечная империя. Проект Academia (ГТРК Культура) (2010)]*

Данное различие естественно связать со степенью актуализованности референта-говорящего. Это предположение получает дополнительное подтверждение, если посмотреть на те редкие контексты, в которых прямой порядок облегчен и для полных именных групп: это, в частности, те *как*-ремарки, где референт-говорящий актуализован как один из участников беседы, см. (14) или как один из членов ассоциированного множества – например, входит в составляющую с контрастивным значением, см. (15), где контрастивную интерпретацию обеспечивает частица *ещё* (контраст в (15) усилен еще и самоисправлением говорящего, заменяющего инверсивный порядок на прямой *– как говорил… как еще классики марксизма говорили*):

от меня, как скажет некий политик через двадцать лет, «симметричного ответа»; купец, как скажет всякий, кто имел с ним дело, жил обманом…

[5] В обследованных нами данных МУРКО это правило практически не имеет исключений, однако, как верно указал рецензент, инверсия местоимения и глагола возможна: если в *как*-фразе имеется тяжелая группа, которая может приобретать рематический статус, она перемещается в крайнюю правую позицию, а местоимение помещается после глагола перед этой группой: *Нет, нет, она не забыла его, как говорил он ночью в клинике бедному Ивану*. По-видимому, такие построения не очень свойственны устной речи, и этим объясняется отсутствие подобных примеров в нашей выборке.

(14) *Это соответствует ну ультрафиолету/ как Володя говорил/ это пятьдесят тысяч градусов по другой шкале/ в другом формате. [Динамическая нестабильность воды. Программа "Гордон" (НТВ) (2003)]*

(15) *Вот по сути дела в этот момент оказалось/ что люди/ создающие этот продукт/ являются не рабочими/ которые должны прийти/ как говорил… как ещё классики марксизма говорили/ прийти и наняться к капиталисту/ а он является человеком/ который создаёт от начала до конца весь продукт. [Класс интеллектуалов. Программа "Гордон" (НТВ) (2003)]*

Более глубокий теоретический и типологический анализ инверсии в *как*-ремарке выходит за рамки данной работы, однако мы считаем уместным указать на следующие три обстоятельства, демонстрирующие системность обнаруженного нами явления. Первое. Инверсия в клаузах, вводимых сравнительными союзами, наблюдается в целом ряде языков, причем, как правило, такие клаузы демонстрируют те или иные симптомы парентезы, ср. так называемые as-parentheticals в английском языке *Mary was reading Moby-Dick, as could have been her brother*; данные английского, датского, тайского, испанского, ирландского и др. приводятся, inter alia, в Potts 2002, LaCara 2016. Второе. Инверсия в авторской ремарке наблюдается и в стандартных конструкциях с прямой речью, если ремарка вынесена в постпозицию к цитате, ср. *«Сегодня занятия не будет!» – сказал Петя.* Более того, в отличие от *как*-ремарки, в постпозитивной ремарке в формате главной клаузы инверсия сохраняется, даже если подлежащее местоимение, ср. *«Сегодня занятия не будет!» – сказал он,* но не **«Сегодня занятия не будет!» – он сказал*. И опять же, препозиция сказуемого в таком контексте наблюдается не только в русском языке, ср. англ. *‘You’re the Best Bear in All the World,’ said Christopher Robin soothingly*. И, наконец, третье. В русском языке имеется целый арсенал конструкций с выдвижением глагола влево. В работах Т.Е.Янко (2001:197-229, inter alia) детально проанализированы коммуникативные стратегии, которые кодируются с помощью различных вариантов такого выдвижения. Как кажется, с *как*-ремаркой также ассоциированы определенные ограничения на коммуникативную структуру. В частности, в конструкциях с *как*-ремаркой рематический статус имеет сама цитата, тогда как ремарка является либо темой, либо парентетической составляющей. В следующем разделе мы продемонстрируем некоторые просодические подтверждения этому тезису.

4 Просодия конструкций с *как*-ремаркой

В предыдущих разделах мы увидели ряд симптомов семантической и грамматической редукции *как*-ремарки, или, иначе говоря, по сравнению с функционально близкими главными клаузами, здесь налицо ограничения семантических и грамматических свобод: *как*-ремарка может вводить не любую цитату, а только ссылку на текст, ассоциированный с эпистемической оценкой говорящего (предъявляемый как имеющийся в общем поле знаний), в *как*-ремарке невозможно отрицание, ограничены формы глагольных времен и типы иллокуций, действуют особые правила порядка слов. В этом разделе мы покажем, что в *как*-ремарке наблюдается и просодическая редукция, однако с этой точки зрения *как*-ремарки неоднородны: степень их просодической редукции тесно связана с их позицией в синтаксической структуре предложения и, шире, с их позицией в локальной дискурсивной структуре. Рассмотрим прототипические паттерны реализации конструкций с *как*-ремаркой.

4.1 *Как*-ремарка перед цитатой

Рассмотрим два примера препозитивной *как*-ремарки – в (16) автор выражен полной именной группой, в (17) – местоимением. Для каждого примера приведем сначала его полную графическую форму, как она дана в МУРКО, далее – транскрипцию части примера с разметкой движения тонов и локализацией фразовых акцентов[6] и интонограмму в формате анализатора PRAAT:

[6] Для указания на направление движения тона иконически используются знаки «/», «\» и «–». Ударный слог слова – носителя рематического акцента подчеркивается. О других деталях используемой системы дискурсивной транскрипции см. Кибрик, Подлесская (ред.) 2009. Напомним, что в той версии примера, которая дается по МУРКО, знаки «/», «\» имеют другую интерпретацию – там они используются для сегментации на межпаузальные фрагменты.

(16) *Как говорил Козьма Прутков/ «зри в корень». [Теория асимметрии мозга. Программа "Тордон" (НТВ) (2003)]*

Как говорил (…0.6) Козьма /Прутк<u>о</u>в,
(…0.5)
«/зри в \к<u>о</u>рень».

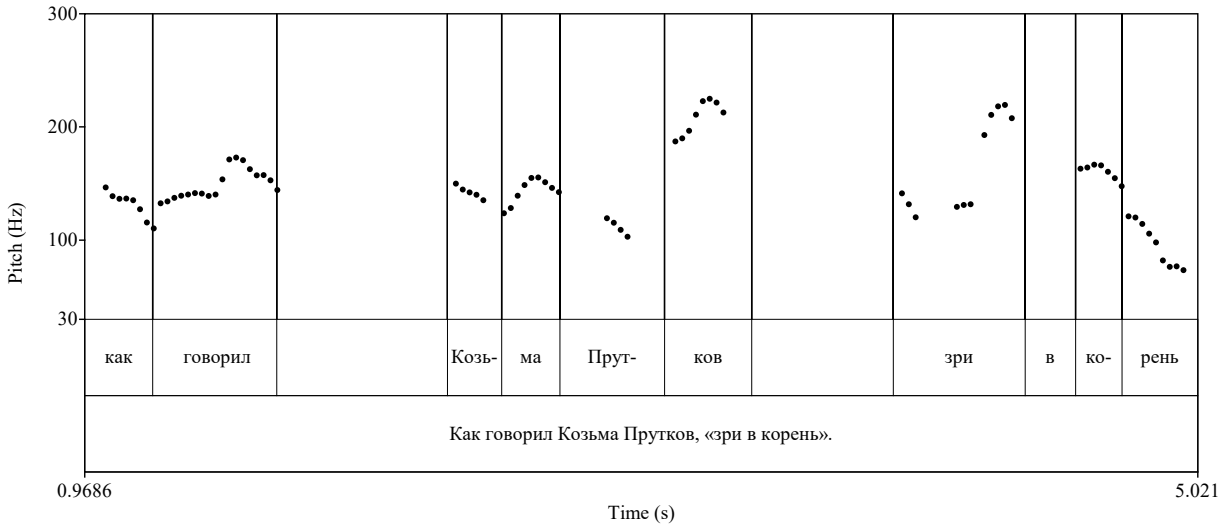


Рисунок 1. Интонограмма к примеру (16)

(17) *И особый признак/ на котором бы хотелось остановиться/ это тип значения. Как мы уже сказали чуть раньше/ конечно нам бы хотелось прослеживать модели переносов в разных языках. Однако что мы делали в русской базе? Для каждого значения мы указывали значение/ от которого образовано данное/ и тип перехода. [М. Кюсева. Доклад на конференции «Диалог 2013» (2013) // Из коллекции НКРЯ]*

это тип \знач<u>е</u>ния.
(..0.4)
Как мы уже \сказали чуть /р<u>а</u>ньше,
(..0.1)
конечно нам бы \хот<u>е</u>лось прослеживать модели /переносов в разных \ языках.

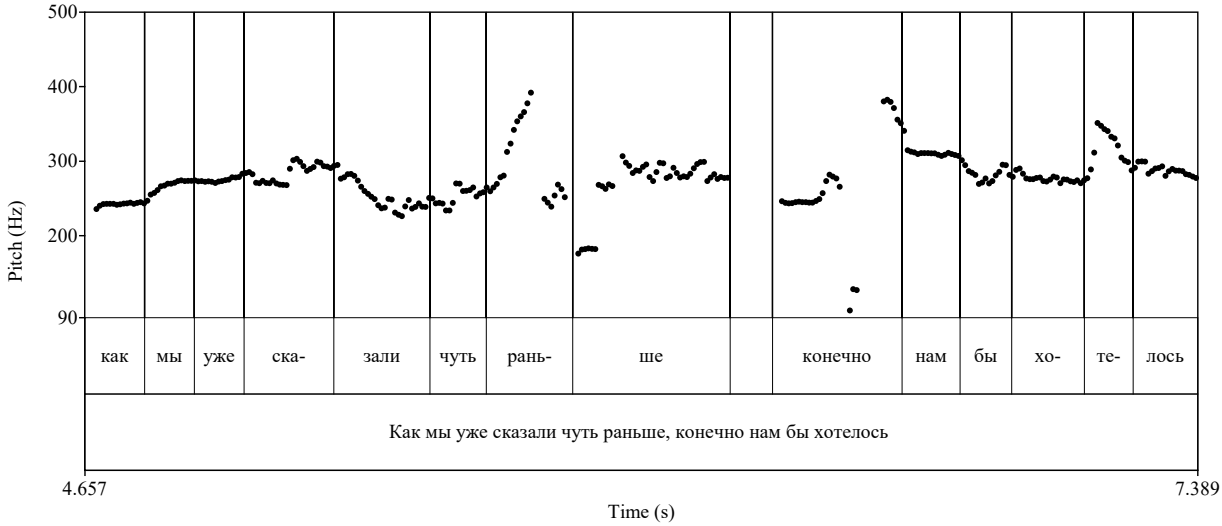


Рисунок 2. Интонограмма к примеру (17)

Данный паттерн используется в абсолютном начале эпизода или после явно выраженной завершенности предшествующего сегмента. И в (16), и в (17) фразовый акцент в *как*-фразе расположен на последнем слове. В (16) – это подлежащее (=автор цитаты), здесь действует упоминавшееся выше правило препозиции глагола при подлежащем, выраженным полной именной группой. В (17) подлежащее – личное местоимение, инверсии нет, фразовый акцент локализуется на заключительном обстоятельстве времени. В обоих случаях фразовый акцент реализуется с выраженным подъемом и кумулятивно кодирует рему и незавершенность, т.е. проецирует продолжение дискурса. Эта же просодическая конфигурация может манифестировать и омонимичную коммуникативную структуру, в которой *Козьма Прутков* и *чуть раньше* являются носителями узкого (контрастного) фокуса (‘Козьма Прутков, а не кто-то другой’; ‘раньше, а не сейчас’). При контрасте акцент может сдвигаться внутрь *как*-фразы с сохранением того же направления движения тона – в данном случае, подъема тона по типу ИК3 в терминах интонационных конструкций (Брызгунова 1982):

(17)*’ Как мы уже чуть /р<u>а</u>ньше сказали….*

Если инверсии нет, и у глагола нет распространителей, то типичным акцентоносителем становится финальный глагол, так, у примера (17) легко реконструируется и вариант (17)”:

(17)*” Как мы уже /сказ<u>а</u>ли….*

Препозитивные *как*-фразы с финальным акцентированным глаголом встречаются в МУРКО массово, что не удивительно, учитывая, что в корпусе широко представлены записи докладов и лекций «от первого лица»:

(18) *Как я уже сказал/ сочинения лингвистов-любителей чрезвычайно однообразны. [Андрей Зализняк. Что такое любительская лингвистика? Лекции Полит.ру (2010)]*

Препозитивные *как*-фразы в рассмотренных выше примерах не демонстрируют симптомов просодической редукции, для них характерны выраженные акценты: на рисунках 1 и 2 видно, что подъем тона происходит в существенном интервале – от 100 Hz до 250 Hz в (16) и от 250 Hz до 400 Hz в (17); по перцептивной оценке, нет увеличения темпа, снижения громкости. Кроме того, препозитивные *как*-фразы могут иметь внутреннюю коммуникативно-просодическую структуру и, в частности, иметь дополнительные акценты, помимо главного фразового: так, в примере (17) имеется дополнительный акцент на слове *сказали*, он реализуется с падающим тоном, адаптированным к финальному подъему. В целом, препозитивные *как*-фразы ведут себя, с точки зрения просодии так же, как прототипические препозитивные обстоятельственные клаузы, для которых характерен тематический статус в составе полипредикативной конструкции (о возможной интерпретации коммуникативного статуса нефинальных клауз в иллокутивной цепочке, см. inter alia, Коротаев 2018).

Как мы увидим ниже, правила локализации актентоносителя в *как*-фразе едины и не зависят от расположения *как*-фразы относительно цитаты, а вот характер движения тона во фразовом акценте и степень просодической редукции *как*-фразы, наоборот, весьма чувствительны к ее расположению относительно цитаты.

4.2 *Как*-ремарка внутри цитаты

Рассмотрим пример (19):

(19) *Чтобы выполнить свои обещания/ он начинает воевать с Новгородом/ он устраивает жестокий разгром Торжка. То есть тверской князь вёл политику/ как мы бы сказали/ не по средствам/ не по карману. [Николай Борисов. Возвышение Москвы в 14-15 веках. Проект Academia (ГТРК Культура) (2010)]*

То есть \тверской /князь (…0.8) /вёл политику —

(…0.5)
как –мы бы сказали,
— /не по \средствам,
(..0.2)
/не по \карману.

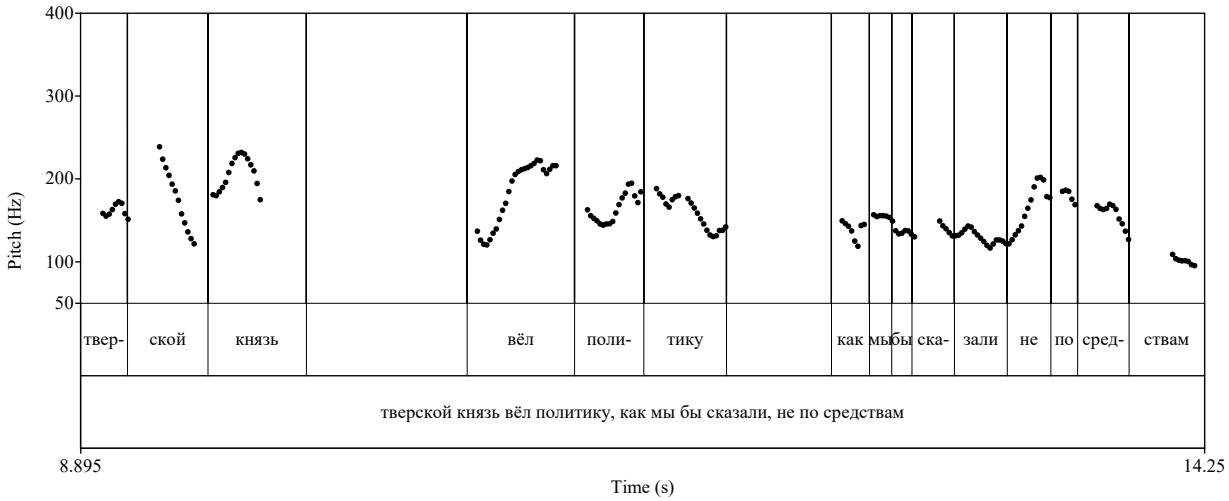


Рисунок 3. Интонограмма к примеру (19)

В данном примере *как*-фраза оказывается во вставке между последовательностью тем (первая тема *тверской князь*, *вторая – вёл политику*) и последовательностью рем (*не по средствам, не по карману*). *Как*-фраза демонстрирует характерные признаки просодической редукции – перцептивно на слове *мы* можно уловить акцент, но он крайне слабый, движение тона – ровное, вся фраза – в тональном диапазоне существенно более узком, чем фрагменты цитаты слева и справа, произносится в более высоком темпе. Если считать акцент на слове *мы* коммуникативно значимым, то его можно интерпретировать как контрастный ('это мы, в наше время, так говорим, в то время так бы не сказали')

Однако продемонстрированный выше паттерн – не единственно возможный вариант манифестации *как*-фразы, разрывающей клаузу. Широко представлен в корпусе и другой вариант, при котором в *как*-фразе просодической парентезы нет, см. (20):

(20) *Но второй глобальный элемент этого синтеза — германцы/ на которых мы завершали прошлый раз наш разговор/ не были христианами. Конечно/ ни в коей мере/ это были чистые язычники/ у них были свои боги и больше всего связаны с силами природы — это были настоящие/ ну как римляне бы сказали/ дикари/ варвары. [Наталия Басовская. Зарождение средневековой цивилизации Западной Европы. Проект Academia (ГТРК Культура) (2010)]*

/больше всего связаны с силами /–прир<u>о-о</u>ды,,,
(..0.3)
это были /настоящие —
(…0.5)
(ə 0.5) /н<u>у</u>,
(ə? 0.5)
(..0.3)
как /р<u>и</u>мляне бы сказали,
— /–дикар<u>и</u>,,
(…0.7)
/–в<u>а</u>рвары,,,

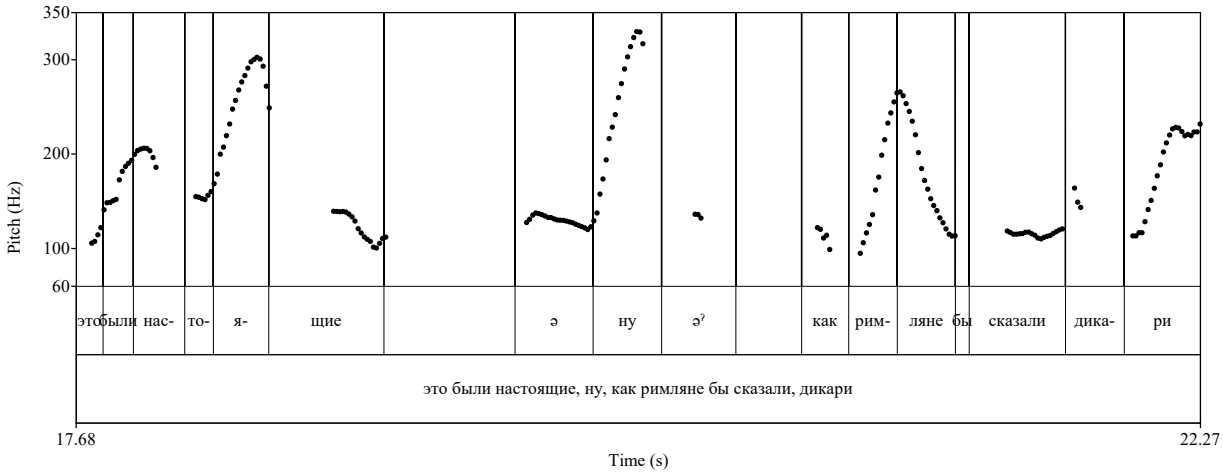


Рисунок 4. Интонограмма к примеру (20)

В примере (20) в *как*-фразе главный фразовый акцент – на слове *римляне*; инверсии нет, что объясняется контрастом – римляне противопоставлены современным интерпретаторам, которые не употребляют столь явно оценочных номинаций. *Как*-фраза «разбивает» сильную внутриклаузальную синтаксическую связь, располагаясь между согласованным определением и определяемым. Однако несмотря на то, что фраза структурно является вставкой, просодических симптомов парентезы не наблюдается: фразовый акцент реализуется с характерным для данной говорящей выраженным подъемом по типу ИК3, разница между нижней и верхней точкой составляет около 180 Hz, нет ни ускорения, ни снижения громкости, слова выразительно артикулированы. Как кажется, мы имеем здесь дело с особой коммуникативной стратегией: *как*-фраза намеренно реализуется как просодически незавершенная, проецирующая продолжение, тем самым говорящий демонстрирует, что ядром цитаты (или, сферой действия цитации) является не вся обрамляющая ставку структура, а только та ее часть, которая расположена после *как*-фразы. В примере (20) имеется дополнительное подтверждение тому, что ядро цитаты расположено после *как*-ремарки: слова *дикари* и *варвары* реализуются с выраженной эмфазой – с усилением громкости и особым типом фонации.

4.3 *Как*-ремарка после цитаты

Постпозиция *как*-ремарки может быть продемонстрирована следующими двумя примерами:

(21) *Такие слова/ товарищи/ засоряют наш язык/ наш великий/ могучий/ прекрасный язык/ как сказал Некрасов. [Анатолий Рыбаков, Николай Калинин. Кортик, к/ф (1973)]*

наш \\великий \\могучий \\прекрасный /язык,
(..0.3)
как (..0.1) /сказал \Некрасов.

Рисунок 5. Интонограмма к примеру (21)

(22) *но мне хотелось бы щас действительно вернуться к физике. Вот в области магнитных нано-структур/ в области суперпарамагнетизма имеет много интересных квантовых эффектов/ где встречаются квантовые и классические закономерности/ как мы сказали. [Суперпарамагнетизм. Программа "Гордон" (НТВ) (2003)]*

где \\встреч<u>а</u>ются (..0.2) квантовые классические \закономерности ,
как мы \сказ<u>а</u>ли.

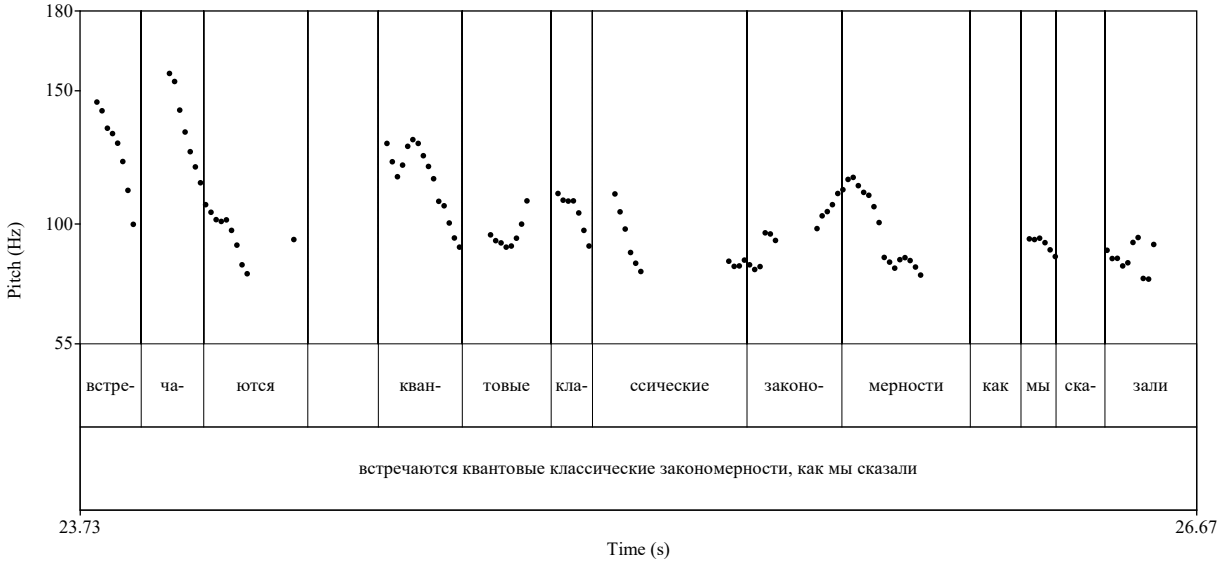


Рисунок 6. Интонограмма к примеру (22)

В (21) постпозитивная *как*-ремарка не парентетична: носитель фразового акцента – слово *Некрасов* – произносится с выраженным падением; кроме того, во фразе есть еще один акцент – подъем на слове *сказал*, это движение тона адаптировано к финальному рематическому падению на *Некрасов*. В цитате, предшествующей *как*-ремарке, скандирующие падения по типу ИК2 и на слове *язык* – финальный подъем в очень высокий регистр, проецирующий продолжение дискурса. В (22) напротив, цитата не проецирует продолжения, фразовый акцент – на слове *встречаются* имеет верификативную функцию и реализуется с резким падением по типу ИК2. Постпрематическая часть достаточно длинная, и, возможно, поэтому появляется дополнительный падающий тон на финальном слове *закономерности*, что подтверждает просодическую завершенность цитаты. В

этой ситуации, *как*-ремарка ожидаемо появляется в редуцированном виде – с очень слабым падением на реме *сказали*, в суженом тональном диапазоне и со сниженной громкостью. Таким образом, в постпозиции к цитате парентеза в авторской ремарке возможна, но не обязательна.

Разумеется, мы вынужденно рассмотрели лишь самые базовые просодические конфигурации с бинарной структурой «цитата + ремарка». В реальном дискурсе реализации *как*-ремарки чувствительны не только к структуре самой цитационной конструкции, но и к её внешним синтаксическим и, шире, дискурсивным связям. Это дискурсивное разнообразие осталось за пределами данной статьи, однако приведем в порядке иллюстрации один красноречивый пример. Как мы уже писали выше, базовое значение вводно-союзной конструкции – дать отсылку к тексту, который имеется в общей базе знаний коммуникантов; при этом сама эта отсылка становится фоном, на котором говорящий формулирует собственное суждение. В результате, в живой речи часто оказывается, что во вставку помещается не только *как*-ремарка, но и целиком вся конструкция «цитата + ремарка». Так, в примере (23) во вставке оказывается цитационная конструкция *Как говорил Блок, дворяне все родня друг другу*: она размещается внутри цепочки характеристик поэта – *дружен с декабристами, связан родственными связями, не участвовал в декабристском движении*. Единство этой цепочки поддерживается семантической однородностью элементов и их просодической однообразностью – все они реализуются как нефинальные элементы списка (с подъемом тона по типу ИК6 или ИК3). Вставочный характер конструкции приводит к тому, что вся она произносится в сниженном регистре, в диапазоне, существенно суженном по сравнению с основным текстом. При этом препозитивная *как*-ремарка вовсе лишается акцента, что, как мы видели выше, препозитивным ремаркам в общем случае не свойственно:

(23) *Был дружен с декабристами. И связан родственными связями. Как говорил Блок/ дворяне все родня друг другу. И не участвовал в декабристском движении. [Юрий Пивоваров. Русская история в зеркале русской мысли. Проект Academia (ГТРК Культура) (2010)]*

был \дружен с /\декабристами,,,
и связан родственными /→связями,,,
(ц 0.4)
(Как говорил Блок дворяне /все² (..0.2) родня друг \другу.)
(ц 0.5)
и-ии не \участвовал в декабристском /движении,

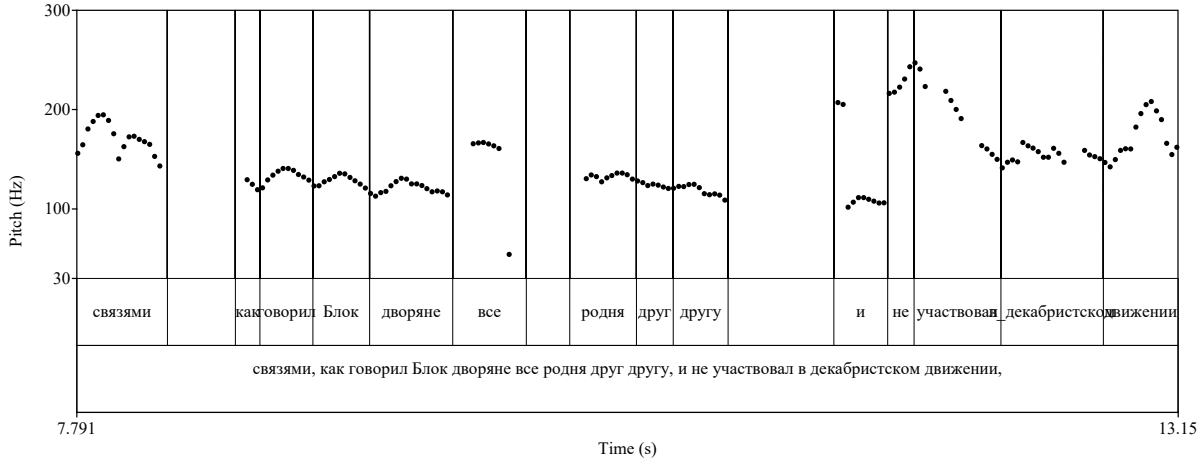


Рисунок 7. Интонограмма к примеру (23)

Сказанное позволяет сделать вывод, что просодическое поведение *как*-ремарки может быть весьма разнообразным и в зависимости от структурных и дискурсивных условий демонстрировать разную степень просодической редукции – от прототипической парентезы с отсутствием фразовых акцентов, узким тональным диапазоном, высоким темпом и т.п., до полного отсутствия парентетических проявлений.

5 Итоги

В богатой литературе по парентезе (см. inter alia Dehé 2010, Schneider 2015 и подробный обзор публикаций там) слова, конструкции и предложения квалифицируются как парентетические (вводные), прежде всего, не по форме, а по функции. Это функция плохо формализуется, но, так или иначе, сводится к признанию их вторичности по отношению к «опорному» высказыванию: они являются вспомогательной, дополнительной ремаркой («interpolated into the current string of the utterance», Dehé 2010:307); их роль в пропозициональном содержании ослаблена (Янко 2003:331), они выведены за пределы текущего речевого акта и фокуса внимания говорящего («Parenthesis enables the speaker to put an item outside the ongoing speech act and, thus, to distance it from the focus of attention of the addressee», Schneider 2015: 281).

Эта функция может манифестироваться в целом ряде эмпирически проверяемых симптомов, некоторые из которых приводятся ниже:

a. Парентеза разрывает текущее высказывание на два фрагмента, между которыми сохраняется тесная связь – синтаксическая или риторическая (дискурсивная)
b. Парентетический фрагмент не формирует узла в синтаксической структуре опорного высказывания.
c. Парентетический фрагмент имеет регуляторный статус (Chafe 1994: 63ff), т.е. используется не для выражения пропозиционального содержания, а для организации и регулирования речевого потока, выражения оценочных значений, привлечения внимания слушающего и т.п.
d. Парентетический фрагмент просодически выделен из опорного высказывания: его границы просодически маркированы, текущие просодические параметры (тоновый диапазон, темп, громкость) отличаются от таковых в опорном высказывании, он не имеет внутреннего коммуникативно-просодического членения.
e. В парентетическом фрагменте имеются ограничения на лексическое многообразие и морфосинтаксис.

Легко убедиться, что каждый их этих признаков в отдельности не проходит теста на необходимость и достаточность. Так, «освященные лингвистической традицией» вводные выражения типа *к счастью* легко размещаются не только внутри опорного высказывания, но на его левой или правой периферии. Или другой пример – регуляторный дискурсивный маркер *вот* в значении возврата к временно прерванному эпизоду: согласно (Кибрик, Подлесская (ред.) 2009: 146-152), этот маркер допускает не только просодически автономное употребление – ожидаемое для вводного слова, но и полную интеграцию в опорное высказывание – атоническое произнесение, отсутствие просодических границ с материалом опорного высказывания (ср. также наблюдения о возможной просодической интеграции парентетических фрагментов в английском языке в Dehé 2007). Статус вводных признается за предложениями, которые присоединяются сочинительными и сравнительными союзами, т.е. имеют не типичную для вводных конструкций эксплицитно маркированную коннектором связь с опорным высказыванием, ср. (24) со вводной клаузой, вводимой союзом *а* (цит. по Подлесская 2018), а также английские *and*-parentheticals (Kavalova 2007) и упоминавшиеся выше *as*-parentheticals (Potts 2002, LaCara 2016):

(24) FS_02-f_Sp (корпус «Веселые истории из жизни», spokencorpora.ru)
38. ··(0.40) И тут он значит /оборачивается,
39. смотрит на одну из кадок с этими /цветами,
40. (а там \кактусов было много,)
41. ···(0.52) и /говорит,
42. что мол одного кактуса не \хватает.

Рассмотренный нами класс вводно-союзных конструкций с глаголом речи тоже показателен в этом отношении – эти конструкции демонстрируют весьма пестрое проявление симптомов парентезы, они:

a) могут занимать любую позицию относительно опорной клаузы. причем *как*-фразы с идентичным лексическим наполнением и грамматической структурой могут менять свойства в зависимости от их позиции относительно цитаты;
b) вводятся сравнительным союзом, т.е. встроены в синтаксическую структуру предложения;
c) имеют прагматически вспомогательную функцию, отсылая к источнику цитаты;
d) неоднородны в просодическом отношении и, в частности, могут произноситься как атонически, так и с выраженной акцентуацией;
e) подвержены частичным ограничениям на порядок слов и грамматическую форму.

Из всего перечисленного только пункты c. и e. безусловно ассоциированы с прототипом парентезы, остальные – плохо вписываются в этот прототип. Эти результаты позволяют предположить, что наиболее плодотворным подходом к описанию парентезы является многофакторный анализ, позволяющий учитывать все многообразие исследуемой зоны и выделять лингвоспецифические кластеры значений релевантных параметров. Таким образом, наше исследование, с одной стороны, имеет дескриптивное значение: на основе корпусных данных оно уточняет наше представление о важном сегменте русской грамматики. С другой стороны, оно позволяет ввести данные русского языка в оборот исследований по кросс-языковой вариативности парентетических конструкций и, тем самым, может иметь теоретическое и типологическое значение.

Литература

[1] Boersma, Paul & Weenink, David. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, 2021. Access mode: http://www.praat.org/
[2] Bryzgunova E. A. Intonation [Intonatsiya], Russian Grammar [Russkaja grammatika]. Vol. 1, Nauka, Moscow, 1982. — P. 98–118.
[3] Chafe, Wallace. Discourse, consciousness, and time. Chicago: University of Chicago Press, 1994.
[4] Cristofaro, SoniaSubordination. (Oxford Studies in Typology and Linguistic Theory.) Oxford: Oxford University Press, 2003.
[5] Dehé, Nicole. Parentheticals // Cummings, Louise (ed.) The Routledge Pragmatics Encyclopedia. Routledge, 2010. — P. 307-308.
[6] Dehé, Nicole. The relation between syntactic and prosodic parenthesis // Dehé, N., Kavalova, Y. (eds.) Parentheticals (Linguistik aktuell/ Linguisticstoday 106), Amsterdam: John Benjamins, 2007. — P. 261-284.
[7] Janko T. E. Kommunikativnye strategii russkoj rechi [Communicative strategies in spoken Russian]. Moskva: Jazyki Slavjanskix Kul'tur, 2001.
[8] Kibrik A. A., Podlesskaya V. I. [Eds.] Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki Slavjanskix Kul'tur, 2009.
[9] Kavalova, Yordanka. And-parenthetical clauses. In Nicole Dehé, Yordanka Kavalova (eds.). Parentheticals. Amsterdam - Philadelphia: Benjamins, 2007. — P. 145–172.
[10] Korotaev, N.A. How Intonation Structures Spoken Narratives: Non-final Phase Contexts // Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue" (2018). Issue 17, 2018. — P. 342-355.
[11] Kustova, Galina. Sistemnye svjazi prisoedinitel'nyx i parentetičeskix konstrukcij [Systemic relations in parcellated and parenthetical constructions] // International conference "Problematic questions in Functional Grammar": In commemoration of A.V.Bondarko 90-th anniversary. Abstracts. St.Petersburg: Institute for Linguistic Studies, Russian Academy of Sciences, 2020. — P. 18.
[12] LaCara, Nicholas J. Anaphora, Inversion, and Focus". Doctoral Dissertations. 746. – 2016 – Access mode: https://scholarworks.umass.edu/dissertations_2/746
[13] Lehmann Christian. Towards a typology of clause linkage // Haiman, John and Thompson, Sandra A. (eds.) Clause combining in grammar and discourse. Amsterdam: John Benjamins, 1988. — P. 181-225
[14] Paducheva E.V. Semanticheskie issledovaniya [Semantic studies]: Semantika vremeni i vida v russkom yazyke [Semantics of tense and aspect in Russian]. Semantika narrativa [Semantics of narrative]. M: Shk. «Yazyki russkoj kul'tury», 1996
[15] Podlesskaya V. "A u nas v kvartire gaz. A u vas?": the Russian Conjunction A viewed through the Prism of Prosodically Annotated Corpus Data // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" Issue 17 (23), Moscow: RSUH, 2018. — P. 601-618.

[16] Potts, Christopher. The syntax and semantics of *as*-parentheticals // Natural Language & Linguistic Theory 20, 2002. — P. 623–689.
[17] Schneider, Stefan..Parenthesis: Fundamental features, meanings, discourse functions and ellipsis // Kluck, Marlies; Ott, Dennis; Mark de Vries (eds.) Parenthesis and Ellipsis [Studies in Generative Grammar, Volume 121], De Gruyter Mouton, 2015. — P. 277-300.
[18] Брызгунова Е. А. Интонация, Русская грамматика, том 1, М.: Наука, 1982. — P. 103–118.
[19] Кибрик А.А., Подлесская В.И. (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 2009.
[20] Коротаев Н. А. Интонационная структура устного рассказа в контексте незавершенности // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог», Вып. 17(24), 2018. — P. 342-355.
[21] Кустова Г.И. Системные связи присоединительных и парентетических конструкций // Международная конференция «Актуальные проблемы функциональной грамматики», посвященная 90-летию со дня рождения А. В. Бондарко. Тезисы докладов. СПб: Институт лингвистических исследований РАН, 2020.
[22] Падучева Е.В. Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. М: Шк. "Языки русской культуры", 1996
[23] Подлесская В.И. «А у нас в квартире газ! А у вас?»: конструкции с союзом А по данным просодически размеченного корпуса // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 17 (23), М.: Изд-во РГГУ, 2018. — P. 601-618.
[24] Янко Т. Е. Коммуникативные стратегии русской речи. Москва: Языки славянских культур, 2001.

SemSketches-2021: experimenting with the machine processing of the pilot semantic sketches corpus

Maria Ponomareva[♠,♡] **Maria Petrova**[♠] **Julia Detkova**[♠] **Oleg Serikov**[♡,♢] **Maria Yarova**[♣]

[♠]ABBYY, Moscow, Russia

[♡]National Research University Higher School of Economics, Moscow, Russia

[♣]Moscow Institute of Physics and Technology, Moscow, Russia

[♢]Deeppavlov MIPT, Moscow, Russia

Abstract

The paper deals with elaborating different approaches to the machine processing of semantic sketches. It presents the pilot open corpus of semantic sketches. Different aspects of creating the sketches are discussed, as well as the tasks that the sketches can help to solve. Special attention is paid to the creation of the machine processing tools for the corpus. For this purpose, the SemSketches-2021 Shared Task was organized. The participants were given the anonymous sketches and a set of contexts containing the necessary predicates. During the Task, one had to assign the proper contexts to the corresponding sketches.

Key words: word sketches, semantic sketches, frame semantics, semantic role labeling, corpus lexicography

DOI: 10.28995/2075-7182-2021-20-560-570

SemSketches-2021: опыт автоматической обработки пилотного корпуса семантических скетчей

Мария Пономарева	**Мария Петрова**	**Юлия Деткова**	**Олег Сериков**	**Мария Ярова**
ABBYY, ВШЭ	ABBYY	ABBYY	ВШЭ, Deeppavlov	МФТИ
Москва	Москва	Москва	Москва	Москва

Аннотация

Статья посвящена различным подходам к автоматической обработке семантических скетчей. В статье представлен первый открытый корпус семантических скетчей для русского языка. На примере данного корпуса рассматриваются особенности семантических скетчей и проблемы, возникающие при их построении, обсуждаются задачи, которые могут решаться с привлечением скетчей, а также дальнейшие перспективы использования скетчей. Особое внимание уделяется возможности создания инструментов автоматической обработки корпуса. В качестве эксперимента по созданию подобных инструментов авторами проведено соревнование SemSketches-2021, в рамках которого участникам предлагалась задача по работе с корпусом скетчей, где требовалось соотнести анонимизированные скетчи с рядом контекстов для соответствующих предикатов.

Ключевые слова: скетчи слов, семантические скетчи, семантика фреймов, разметка семантических ролей, корпусная лексикография

1 Introduction

The current paper continues the work on the semantic sketches which were first presented at the Dialogue-2020 conference.

The idea of the semantic sketch was introduced in [7]. The semantic sketch is a special representation of a word's compatibility where all semantic links of the word are grouped according to their semantic relations with the core they depend on. All possible semantic dependencies are statistically ranged: first, the frequency of the collocation between the parent and the child is taken into account; second, the frequency of the semantic role for the given core (for instance, the frequency of the Agent, Locative, Object, or Time).

The most frequent collocations form the semantic sketch of the word. In [7], the authors focused on the creation of the semantic sketches and on testing the semantic mark-up used for the sketches. Namely,

they measured the correctness of the predicate's choice in a set of sentences and the choice of the proper semantic roles for the predicates' dependencies.

In the present work, the focus has been made on building the pilot corpus of the semantic sketches themselves, **the SemSketches corpus**. The corpus is aimed at achieving several purposes:

1. to evaluate how representative the sketches are,
2. to elaborate some tools for processing the sketches,
3. to specify what kind of tasks the semantic sketches can help to solve, as our further plan is to integrate the sketches into the General Internet-Corpus of Russian (GICR, [4], [3]),
4. to analyze what kind of mistakes we happen to face while creating the sketches.

The idea to represent a word's meaning in the form of the semantic sketch is closely related to the main idea of distributional semantics according to which the meaning of the word can be represented through its lexical co-occurrence. The famous formula for the idea given in [10] says: "You shall know a word by the company it keeps".

Over the past few years, vector representations have become a traditional method of representing the word's semantics. The static embeddings such as word2vec [8] and FastText [9] as well as the dynamic embeddings that followed, such as ELMo [5], ULMFit [13], and BERT [2], have completely changed the NLP field. However, quality evaluations of the vector representations pose a challenge, as their serious drawback is that one can neither assess nor interpret them directly.

Whereas the vector is a numeric meaning representation, appropriate for computers, the semantic sketch can be considered its human-interpretable counterpart.

As an experiment on processing the sketches automatically, we have introduced **the SemSketches Shared Task**. One of its goals is to connect these two methods of semantic representation.

The Shared Task suggested the following problem. Participants were given the corpus of the semantic sketches with the core predicates unknown, that is, the semantic roles of the dependencies and the word-fillers of the roles were given, but not the predicates they were attached to. We have presented a set of such anonymous sketches and a list of contexts containing the predicates. The task was to create a tool that assigns the sketch to the corresponding contexts.

For most sketches, the task did not seem difficult for a human, as some of the examples will demonstrate below, but it turned out to be rather complicated for the computer, as the results of the competition showed. The corpus and the Shared Task results are available at the SemSketches github[1].

2 What is a semantic sketch

There is no need to underline the importance of using text corpora for various purposes nowadays. The size of the corpora is growing quickly. On the one hand, it gives the users more opportunities and allows one to receive more representative data. On the other hand, with a bigger corpus, more sophisticated tools are demanded to process the results of the search queries.

One of the methods to describe the word's compatibility is to present it in the form of a syntactic sketch [22]. The syntactic sketch is a lexicographical profile of a word, where word dependencies are classified by their grammatical roles and ranged by the statistics of their compatibility with the core. The syntactic sketches were first introduced in the Sketch Engine project[2] and over the past years have become widely used in lexicography, language teaching, multilingual corpora creation, various translation resources, and in a number of other areas.

The evident advantage of the syntactic sketch is its vividness: it shows simultaneously all of the most frequent syntactic dependencies of a word and arranges them in a table according to the roles. At the same time, the syntactic sketches have one strong limitation: the grammatical information they are based on does not allow one to take lexical homonymy into account, which complicates the interpretation of the obtained results.

In order to solve this problem, an attempt was made to create the semantic sketches [7], where the representation of a word's compatibility is supplemented with semantic relations between words (each

[1] https://github.com/dialogue-evaluation/SemSketches

[2] www.sketchengine.eu

relation is marked not only with a syntactic, but with a semantic role as well) and semantic classes of words (which mark the specific semantic meaning of a word in a context).

Therefore, the semantic sketch is understood as a generalized lexicographic portrait of a word, which includes the most frequent semantic dependencies of the verb. In other words, it is a way of representing the compatibility of words, where the description of each word includes a set of its most frequent semantic dependencies classified according to their semantic roles. For each role a number of relevant "fillers" (words and phrases) are given, and the fillers are ranked according to the frequency of their compatibility with the core. Each sketch illustrates a word with a certain meaning.

The semantic sketches are built with the help of the Compreno parser [24]. Unlike other parsers, the Compreno suggests full semantic mark-up, namely, it deals not only with the actant semantic dependencies of the predicates, but with the adjuncts, modifiers, and other dependencies as well [18]. It makes the sketches an important tool for dealing with the semantic role labeling (SRL) problem which has attracted many researchers recently.

Despite high interest in the problem ([12], [11], [17], [15], [6], [16], [25]), until the current moment no research among the SRL papers has been presented (or, at least, we have not seen any), where all semantic roles are taken into account. Most works focus on the actant dependencies only, such as Agent, Object, or Experiencer. In the meantime, for many predicates, circumstantial dependencies are enough frequent and significant to get into the predicate's sketch together with its actants, and, moreover, in some cases, help to identify the core even better than the actants do.

The sketches are illustrated in the two examples below, the first one — for the verb «страдать:SUFFERING_AND_TORMENT» 'to suffer' (Figure 1) and the second one — for the verb «готовить:TO_PREPARE_FOOD_SUBSTANCE» 'to prepare food, to cook' (Figure 2):

Experiencer	DegreeIntensity	Cause_From	Time	Modality	Cause
моя душа my soul	ужасно terribly	от одиночества from loneliness	хронически chronically	по-настоящему truly	оттого therefore
герой character	неимоверно appallingly	от голода from hunger	всю жизнь all their life	должно быть must be	из-за нашей любви because of our love
тело body	больше more	от отсутствия свободы from lack of freedom	в детстве in childhood	явно clearly	по собственной вине through one's own fault
народ nation	нестерпимо unbearably	от холода from cold	в юном возрасте at a young age	по-видимому apparently	потому because of
люди people	бесконечно endlessly	от жажды from thirst	потом after	несомненно certainly	поэтому that's why
дети children	безмерно immensely	от недостатка времени from lack of time	вечно forever	вроде бы seem to be	
мирное население civilians	меньше less	от любви from love	нередко often	действительно really	
женщины women		от нехватки дров from lack of firewood	раньше earlier	на самом деле actually	

Figure 1: the sketch for the verb «страдать:SUFFERING_AND_TORMENT» ('to suffer'). Here the elements of the sketch are given with their rough translations.

The participants of the Shared Task got the same representations, but did not get the titles of the sketches. However, as the pictures above demonstrate, it does not seem difficult for a human to guess the proper predicates for the sketches, which allows us to regard the sketches as representative illustrations for the verb's compatibility.

Object	Time	Agent	Locative	Ch_Evaluation	Quantifier
ужин supper	заранее in advance	повар chef	на примусе on the primus stove	отменно perfectly	сам himself
обед dinner	загодя in advance	хозяйка housewife	на плите on the stove	классно great	одна by herself
еду food	свеже freshly	бабушка grandmother	на кухне in the kitchen	превосходно superbly	
завтрак breakfast	завтра tomorrow	кухарка cook	на плитке on the portable stove	замечательно wonderfully	
блюда dishes	впрок for the future	жена wife	на пару in the steam	великолепно excellently	
пищу meal	к празднику октября for the October holiday	повариха cook	на керосинках on the kerosene stove	неплохо nicely	
кофе coffee	в субботу on Saturday	мама mother	в русской печи in the Russian stove	прекрасно perfectly	
салат salad	по очереди by turn	временщик temporary worker	на костре on the fire	здорово excellently	

Figure 2: the sketch for the verb «готовить:TO_PREPARE_FOOD_SUBSTANCE» (‘to prepare food, to cook’). Here the elements of the sketch are given with their rough translations.

3 The SemSketches Shared Task

To explore the semantic sketches as far as their quality and representativeness are concerned, we have created the pilot corpus of Russian semantic sketches and made it the basis for the SemSketches Shared Task. The problem was formulated as follows: *given a set of anonymized sketches and a set of contexts for different predicates, one should match each predicate in its context to a relevant sketch.*

The second part of the competition data is the set of the contexts given for different predicates. In the case of ambiguous predicates, the WSD problem can be stated.

3.1 Data preparation

Sketches

The sketches were built on the texts from the Magazine Hall of the GICR.

Although the parser gives us the full semantic mark-up, we have implemented some restrictions for the present research. As in [7], the authors have taken only verbal cores and their subtrees: all verbs are marked with semantic classes (denoting their meanings) and the semantic roles for their direct dependencies.

We did not mark the dependencies of the non-vebal cores, the dependencies of the ellipted verbs and the ellipted groups themselves, as well as the syntactically moved groups. In addition, we have introduced some additional restrictions for the purpose of the current competition, namely, we have excluded pronouns and personal nouns, as they complicate the work with the anonymized sketches.

For the current corpus, we have chosen only verbs which have at least two meanings, as it makes the task of defining proper sketches more interesting, on the one hand, and, on the other hand, contributes to solving the WSD problem. It means that each verb chosen entered at least two semantic classes.

The number of such verbs for the Russian language turned out to be more than ten thousand. Then we chose a subset of the list through selecting the verbs by the following principles.

At the beginning, we have ranged the sample so that the verbs with the most frequent meanings came first: for instance, the verb *рубить* meaning TO_HACK (***рубить*** *дерево* — 'to hack a tree') is sufficiently frequent, while the same verb meaning TO_KNOW_ABOUT (***рубить*** *в математике* — 'to understand mathematics well') is rather marginal and has thus been positioned at the end of our list. The frequency of different meanings has been obtained with the help of the Compreno parser.

Next, we have collected the verbs' sketches taking into account the number of the relations the verb has in the corpus. Namely, we have collected all the semantic dependencies for each meaning of each verb in our marked-up corpus, and if the number of the dependent nodes exceeded the threshold of 2000, the predicate in the certain value was selected for inclusion in the final set. During this procedure, all dependencies were taken into account — both different and repeated, in order not to lose any frequent predicates with limited lexical compatibility. At the same time, the threshold was rather high to preserve the quality of the sketches.

At last, the final number of sketches in the pilot corpus became 915. Due to the exclusion of rare meanings, some verbs kept only one meaning in the sample, that is, the terminal verb list contained both polysemantic verbs with several meanings in the sample and polysemantic verbs which entered in our sample only in one (the most frequent) meaning.

The next step was to analyze the correctness of the sketches, namely, to check whether the semantic dependencies and the fillers of the dependencies that got into the sketch really refer to the verb in the given meaning. The errors check was performed for a subsample of the corpus which formed the manual Dev data (see below).

Most errors refer to situations where the more frequent homonym influences the less frequent one. For instance, the verb *писать* meaning 'to paint' (***писать*** *портрет с кого-л.* — 'to paint smb.'s picture') is less frequent than *писать* meaning 'to write' (***писать*** *письмо* — 'to write a letter'), so the sketch for the *писать* — 'to paint' contains some incorrect examples in the Object dependency — such as 'to write letters'.

The reason is that when building the semantic structures for the sentences the sketch is based on, the structure with the incorrect but more frequent homonym gets a higher evaluation due to the high statistics of the more frequent verb.

Another error can be illustrated with the sketch «готовить:TO_PREPARE_MEDICINE _OR_FOOD» 'to cook'. It contains combinations like ***готовить*** *резервную копию* — 'to prepare a reserve copy'. Here the problem is that the compatibility of 'copy' with the verbs depends not on the 'copy' itself but on the semantics of the noun following it, that is, 'the copy of the cake' is also possible.

As an instance of the sketch with the incorrect semantic dependency, let us take the sketch «выходить:идти:TO_WALK» 'go out' on the Figure 3. The sketch contains the Agent_Metaphoric slot which must be definitely referred to another meaning, and the Purpose_Goal slot contains the incorrect filler *на связь* (*выйти* ***на связь*** means 'to get in touch', and here another homonym of the verb *выйти* is supposed to be):

The main reasons for the mistakes in the sketches are the incorrect influence of the statistics, certain inaccuracies of the semantic models in the parser, and the impossibility of distinguishing between the homonyms due to the closeness of their meanings or lack of distinguishing context in the sentences.

Contexts

Every meaning from the chosen set is illustrated with contexts. A context is a sentence with one target predicate highlighted. No additional mark-up is presented. Each meaning corresponds to several dozens of contexts with the target words having this meaning. The contexts were collected from news, fiction, publicistic texts, being close by genre to those presented in Magazine Hall. It is important that the contexts do not overlap with the corpus which the sketches were built on. The excerpt from the contexts is given in Table 1.

Locative_FinalPoint	Locative_InitialPoint	Time	Agent	Agent_Metaphoric	Purpose_Goal
на улицу outside	из дома out of the house	утром in the morning	люди people	книга book	покурить for a smoke
во двор into the yard	из комнаты out of the room	только что just now	женщина woman	второе издание second edition	погулять for a walk
в коридор into the corridor	из дому out of the house	через минуту in a minute	мужчина man	срок deadline	на волю to the liberty
на сцену on the stage	из кабинета out of the office	вечером in the evening	девушка girl	сборник collection	на связь to get in touch
на крыльцо on the porch	из машины out of the car	рано early	старик old man	роман novel	прогуляться for a walk
в свет into society	из подъезда out of the entrance	через полчаса in half an hour	жена wife	книжка book	встречать to meet
на балкон to the balcony	из квартиры out of the apartment	как раз just	отец father	фильм film	на поклоны for a bow
на дорогу to the road	оттуда from there	ночью at night	мама mother		подышать for a breath

Figure 3: the semantic sketch for the verb «выходить:идти:TO_WALK» ‘to go out’). Here the elements of the sketch are given with their rough translations.

ID	dev.sent.rus.116
target	наполнились
start	46
end	57
context	Когда доктор вошел, она вспыхнула, и глаза ее наполнились слезами

Table 1: The example of the context. The position of the target word *наполнились* ‘filled’ in the context ‘When the doctor came in, she flushed, and her eyes filled with tears’ is defined by the offsets.

Datasets

The task was meant to be solved in a few-shot or unsupervised learning manner. During the Shared Task, we provided the participants with two sets of data. In the first phase, the Trial data was published. It comprises three parts: a set of sketches, a set of contexts, and mapping between these two sets. The participants could use the data to get familiar with the formats, to test their hypotheses and to fine-tune their systems. During the second phase, we provided the participants with the main set of the sketches and corresponding contexts, which will be referred to as Dev data.

In contrast to trial data, where the mapping had been given, for Dev data the participants were asked to find the relations between the sketches and the contexts themselves.

For the third phase, we manually selected 100 sketches and evaluated the corresponding contexts. This data formed the gold standard set for the task, which we will refer to as Manual Dev data. Table 2 shows the size of the obtained datasets.

During the second phase, the participants were able to commit their answers to the CodaLab[3] to know the results on the Dev data and to choose the best decision. During the third phase, the performance of the best variants was finally evaluated on the Manual Dev data.

After the announcement of the results, we published the answers (the mapping between the sketches and the contexts) on the SemSketches github.

[3] https://competitions.codalab.org/competitions/29992

split	number of sketches	number of contexts
Trial	20	2000
Dev	895	44750
Manual Dev	100	4347

Table 2: The size of the SemSketches datasets, *Manual Dev* data forms a subset of *Dev* data

3.2 Evaluation metric

The submitted systems were evaluated using the **accuracy** metric. For the Shared Task, accuracy was calculated as the number of matched pairs between the participants' answers and test markup divided by the total number of contexts.

The evaluation script is publicly available on the SemSketches github.

3.3 Baseline

The participants were provided with a weak baseline solution. The solution was based on the masked language modeling (MLM) mechanism of the RuBERT [14] model.

For a given context *cont*, *sketch* was chosen according to the computed sketch scores based on MLM candidates. MLM candidates (MLM^N_{cont}) were calculated as follows:

1. syntactic analysis using the UDPipe ([23]) has been performed to find the direct dependents of the target predicate;
2. for each of the direct dependents, top-N mask replacements Rep^N_{dep} were stored;
3. stored replacements were intersected, i.e. $MLM^N_{cont} = \bigcap\{Rep^N_{dep}\ \forall dep \in cont\}$;
4. sketch *Score* was computed as the number of tokens present in the intersection of the sketch representation and the stored MLM candidates.

$$Score\,(sketch, cont) = \left|MLM^{1000}_{cont} \cap Tokens_{sketch}\right|$$

The intersection was performed over lemmas thus treating *на заре* and *заря* as intersecting entries.

The weak baseline system has shown 0,0094 accuracy on the Dev data set thus overperforming the random baseline.

3.4 Submitted systems

Three teams participated in the Shared Task: `paleksandrova`, `good501`, `smpl`. All teams suggested the solutions based on different approaches, and each solution managed to overcome the baseline. However, the final scores of each team turned out to be rather modest. To compare the results achieved, see Table 3 where the score of each team and the baseline score are presented.

Team	Dev Score	Manual Dev Score
`paleksandrova`	0.309	0.277
`good501`	0.104	0.127
`smpl`	0.182	0.121
`baseline`	0.0094	0.0035

Table 3: SemSketches Shared Task: the results of the submitted and baseline systems.

Let us now shortly characterize each decision and analyze what core problems they faced.

The team `smpl` used the brute-force approach. LM score has been used to rank sketches and choose the best one for each context. To estimate how well the predicate *pred* fits into the given sketch *sketch*,

the *LM score* was used. *LM score* is the average probability of $pred$ to replace *[MASK]* token in template sentence «*[MASK] cell*». Template sentences were generated for each *cell* present in *sketch*.

The team `Good501` used the approach based on the sentences' similarity objective, which is a popular objective when training language models. Target predicate was highlighted in the sentence using special tags. Sketch tables were flattened into pseudo-sentences. For the given sentence, the most similar sketch was chosen by using the Sentence-BERT[21] siamese similarity mechanism.

The team `paleksandrova` [1] used the MLM approach, which consisted of first restoring the covered predicate for each of the given sketches, and then picking the relevant sketch for the target sentence.

The covered predicates were restored by generating templates (e.g. «*[MASK] в школу*» — '[MASK] to school') using the sketch content cells. The most frequent predicate of all the MLM hypotheses for the sketch's templates was treated as the re-covered predicate. The first sketch whose predicate matched the sentence predicate was used as the system answer. When no sketch was found by exact matching, the sketch whose restored predicate was word2vec-closest [8] to the sentence predicate was used as the answer.

3.5 The analysis of the submitted systems

During the Shared Task, we formulated the experimental problem leaving enough room for different approaches. Although the performance of the submitted systems may be improved significantly, the proposed ideas were encouragingly diverse and thought provoking. The common feature of all three systems is using the pretrained Language Models.

The team `501good` which adopted the approach from Sentence Transformers introduced the only system that included training. The model was trained on the *Trial* data (20 sketches).

The systems of `smpl` and `paleksandrova` defined their unsupervised strategies for mapping the sketches and the contexts. While the `smpl` team estimated how well each target predicate fits to each sketch using the score from the Masked Language Model, the `paleksandrova` team suggested an original approach imitating the way humans guess the core of the anonymous sketch.

It is worth mentioning that the approaches of `paleksandrova` and `smpl` by design cannot disambiguate the polysemous predicate, as they take only the target verb into account but not its context.

The team `smpl` approach could be thought of as scoring how well the sketch could account as the sentence predicate core. LM is trained on sentence-level objective, therefore, the successful application of the similarity approach demands more sophisticated preprocessing of the input sequence, for example, taking the predicate contexts into account. Such modification could improve the results.

The team `paleksandrova` approach seems to be the most promising one. But the accuracy turned out to be rather low for the following reason. The sketch accumulates several verb forms, namely, it includes all tense, aspect and voice forms. For instance, the verbs *строить* 'build' <Imperfective, NonReflexive>, *построить* 'build' <Perfective, NonReflexive>, *строиться* 'build' <Imperfective, Reflexive>, *построиться* 'build' <Perfective, Reflexive> refer to one sketch. As far as `paleksandrova` approach is concerned, the team regarded such verbs as different candidates for a sketch. At the same time, they chose only one top candidate for each sketch. Therefore, only one grammatical form of the necessary set could be referred to the right sketch.

4 Discussion

In the current paper, we demonstrated the pilot corpus of the semantic sketches, gave a brief analysis of the problems we faced during the corpus creation, and described the results of the SemSketches Shared Task aimed at applying the machine processing tools to the corpus.

The sketches are based on the parser with full semantic mark-up, which defines their value and uniqueness: first, the sketches allow one to analyze not only the actant dependencies, but a full semantic model of a word; second, they differentiate between the various meanings of the verbs.

As far as the opportunities for theoretical investigations are concerned, the sketches can help in dealing with all problems bound with the semantic compatibility of words. Especially, the SRL and the WSD problems must be mentioned here.

As noted above, most researchers focus mainly only on the actant roles, while other dependencies do not usually get much attention. The semantic sketches suggest interesting data in this respect. The sketches include most frequent collocations, that is, the most natural, most typical contexts of a word. Among the dependencies the sketches include, modifiers and adjuncts are quite frequent. For some verbs, they seem to be even more specific than the actants and give more help in identifying the predicate.

For instance, the Locative is a typical circumstantial adjunct, but it is an obligatory slot for the verbs with the position meaning such as *быть* 'be', *находиться* 'be situated'. The Locative slot helps to differentiate between the 'be' with the position meaning and other be-homonyms, while the semantic role corresponding syntactically to the Subject of 'be' does not really contribute to differentiating between the be-homonyms.

It seems that the meaning of the adjuncts and the modifiers is sometimes underevaluated, therefore, an interesting task is to evaluate the correlation between the actant and circumstantial dependencies in the sketches.

As for the applied tasks, one of the promising directions in using the semantic sketches is their implementation for probing tasks for the pretrained language models. The interpretation of the linguistic knowledge encoded in the pretrained models has attracted much attention recently ([26], [19], [20]). We believe that the semantic sketches can serve as a basis for both probing tasks and linguistically-motivated fine-tune tasks for such models.

To summarize, the ideas from the proposed approaches can be used to embed effectively semantic sketches, making them not only a tool for manual lexicographical work but a semantic representation valid for automatic methods of Natural Language Processing.

5 Further plans

Our next plan is to add the sketches into the GICR, which brings two problems to consider.

The first one deals with the errors evaluation: in the current work, we did not check all the sketches in the pilot corpus manually — only the manual Dev data. Therefore, we did not evaluate the total number of the mistakes in the whole corpus. This task is still to be done, including work on both, that is, sketches that seem to be unsuitable (checking the manual Dev data shows that such cases are rare) and sketches containing single mistakes in either the semantic dependencies or their fillers.

The second question is about the processing tools the sketches should be provided with. The SemSketches Shared Task demonstrated that machine tools can be successfully applied to the sketches processing (in spite of the fact that the precision of the solutions suggested by the applicants was not really high). What the tools should look like, depends significantly on the tasks the sketches will be used to solve.

At the same time, we have recently started work on the English sketches, so our further plans include adding other languages to the sketch model, starting with the English sketches.

References

[1] Aleksandrova Polina, Mokhova Anna, Nikolaenkova Maria. Matching semantic sketches to predicates in context using the BERT model // Proc Dialogue, Russian International Conference on Computational Linguistics. — Moscow, 2021.

[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // 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). — Minneapolis, Minnesota : Association for Computational Linguistics, 2019. — Jun. — P. 4171–4186. — Access mode: https://www.aclweb.org/anthology/N19-1423.

[3] Big and diverse is beautiful: A large corpus of Russian to study linguistic variation / Alexander Piperski, Vladimir Belikov, Nikolay Kopylov et al. // Proc 8th Web as Corpus Workshop (WAC-8). — 2013.

[4] Corpus as language: from scalability to register variation / Vladimir Belikov, Nikolay Kopylov, Alexander Piperski et al. // Proc Dialogue, Russian International Conference on Computational Linguistics. — Bekasovo, 2013.

[5] Deep Contextualized Word Representations / Matthew Peters, Mark Neumann, Mohit Iyyer et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). — New Orleans, Louisiana : Association for Computational Linguistics, 2018. — Jun. — P. 2227–2237. — Access mode: https://www.aclweb.org/anthology/N18-1202.

[6] Deep Semantic Role Labeling With Self-Attention / Zhixing Tan, Mingxuan Wang, Jun Xie et al. // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018 / Ed. by Sheila A. McIlraith, Kilian Q. Weinberger. — AAAI Press, 2018. — P. 4929–4936. — Access mode: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16725.

[7] Differential Semantic Sketches For Russian Internet-Corpora / Julia Detkova, Valeriy Novitskiy, Maria Petrova, Vladimir Selegey // Proc Dialogue, Russian International Conference on Computational Linguistics. — Moscow, 2020.

[8] Distributed Representations of Words and Phrases and their Compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Neural and Information Processing System (NIPS). — 2013. — Access mode: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[9] Enriching Word Vectors with Subword Information / Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — P. 135–146.

[10] Firth J. A Synopsis of Linguistic Theory 1930-1955 // Studies in Linguistic Analysis. — Philological Society, Oxford, 1957. — reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

[11] Generalized Inference with Multiple Semantic Role Labeling Systems / Peter Koomen, Vasin Punyakanok, Dan Roth, Wen-tau Yih // Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). — Ann Arbor, Michigan : Association for Computational Linguistics, 2005. — Jun. — P. 181–184. — Access mode: https://www.aclweb.org/anthology/W05-0625.

[12] Gildea Daniel, Jurafsky Daniel. Automatic labeling of semantic roles // Computational Linguistics. — 2002. — Vol. 28, no. 3. — P. 245–288.

[13] Howard Jeremy, Ruder Sebastian. Fine-tuned Language Models for Text Classification // CoRR. — 2018. — Vol. abs/1801.06146. — 1801.06146.

[14] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. — 2019. — 1905.07213.

[15] Lang Joel, Lapata Mirella. Unsupervised Semantic Role Induction with Graph Partitioning // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. — Edinburgh, Scotland, UK. : Association for Computational Linguistics, 2011. — Jul. — P. 1320–1331. — Access mode: https://www.aclweb.org/anthology/D11-1122.

[16] Learning Structured Natural Language Representations for Semantic Parsing / Jianpeng Cheng, Siva Reddy, Vijay Saraswat, Mirella Lapata // Proceedings of the 55th Annual Meeting of the Association for

Computational Linguistics (Volume 1: Long Papers). — Vancouver, Canada : Association for Computational Linguistics, 2017. — Jul. — P. 44–55. — Access mode: https://www.aclweb.org/anthology/P17-1005.

[17] Palmer Martha Stone. Semantic role labeling. Synthesis lectures on human language technologies ; #6. — San Rafael, Calif.] : Morgan & Claypool Publishers, 2010. — ISBN: 9781598298314.

[18] Petrova M.A. The Compreno Semantic Model: The Universality Problem // International Journal of Lexicography. — 2013. — 12. — Vol. 27, no. 2. — P. 105–129. — https://academic.oup.com/ijl/article-pdf/27/2/105/2731792/ect038.pdf.

[19] Probing Pretrained Language Models for Lexical Semantics / Ivan Vulić, E. Ponti, Robert Litschko et al. // ArXiv. — 2020. — Vol. abs/2010.05731.

[20] Ravichander Abhilasha, Belinkov Yonatan, Hovy Eduard. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? — 2021. — 2005.00719.

[21] Reimers Nils, Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 3982–3992. — Access mode: https://www.aclweb.org/anthology/D19-1410.

[22] The Sketch Engine: ten years on / Adam Kilgarriff, Vít Baisa, Jan Bušta et al. // Lexicography. — 2014. — P. 7–36.

[23] Straka Milan, Straková Jana. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe // Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. — Vancouver, Canada : Association for Computational Linguistics, 2017. — August. — P. 88–99. — Access mode: http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

[24] Syntactic and semantic parser based on ABBYY Compreno linguistic technologies / K. V. Anisimovich, K. Ju. Druzhkin, F. R. Minlos et al. // Proc Dialogue, Russian International Conference on Computational Linguistics. — Bekasovo, 2012.

[25] Syntax for Semantic Role Labeling, To Be, Or Not To Be / Shexia He, Zuchao Li, Hai Zhao, Hongxiao Bai // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 2018. — Jul. — P. 2061–2071. — Access mode: https://www.aclweb.org/anthology/P18-1192.

[26] What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties / Alexis Conneau, German Kruszewski, Guillaume Lample et al. // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 2018. — Jul. — P. 2126–2136. — Access mode: https://www.aclweb.org/anthology/P18-1198.

Short Text Clustering with Transformers

Leonid Pugachev
Moscow Institute of
Physics and Technology
9 Institutskiy per., Dolgoprudny,
Moscow Region, 141701,
Russian Federation
leonid.pugachev@phystech.edu

Mikhail Burtsev
Moscow Institute of
Physics and Technology
9 Institutskiy per., Dolgoprudny,
Moscow Region, 141701,
Russian Federation
burtsev.m@gmail.com

Abstract

Recent techniques for the task of short text clustering often rely on word embeddings as a transfer learning component. This paper shows that sentence vector representations from Transformers in conjunction with different clustering methods can be successfully applied to address the task. Furthermore, we demonstrate that the algorithm of enhancement of clustering via iterative classification can further improve initial clustering performance with classifiers based on pre-trained Transformer language models.

Keywords: short text clustering, language models, transformers
DOI: 10.28995/2075-7182-2021-20-571-577

Кластеризация коротких текстов с помощью трансформеров

Пугачёв Леонид
Московский
физико-технический
институт
141707, Московская область,
г. Долгопрудный,
Институтский пер., д. 9
leonid.pugachev@phystech.edu

Бурцев Михаил
Московский
физико-технический
институт
141707, Московская область,
г. Долгопрудный,
Институтский пер., д. 9
burtsev.m@gmail.com

Аннотация

Методы для решения задачи кластеризации коротких текстов часто используют векторные представления слов для переноса обучения. В этой статье показано, что для решения задачи вместе с различными методами кластеризации могут успешно применяться векторные представления предложений из трансформеров. Более того, показано что алгоритм улучшения кластеризации с помощью итеративной классификации может дополнительно улучшить качество исходной кластеризации с помощью классификаторов, которые основываются на предобученных трансформенных языковых моделях.

Ключевые слова: кластеризация коротких текстов, языковые модели, трансформеры

1 Introduction

There are currently a lot of techniques developed for short text clustering (STC), including topic models and neural networks. The most recent and successful approaches leverage transfer learning through the use of pre-trained word embeddings. In this work, we show that high quality for STC on the range of datasets can be achieved with modern sentence level transfer learning techniques as well. We use deep sentence representations obtained using the Universal Sentence Encoder (USE) [16, 9].

Training of deep architectures can be effective for particular clustering tasks as well. However, application of deep models to clustering directly is difficult since we do not have labels a priori. We show that

fine-tuning of classifiers such as BERT [2] and RoBERTa [11] for clustering can be done with the Enhancement of Clustering by Iterative Classification (ECIC) algorithm [3]. Thus, we develop a combined approach to STC, which benefits from the usage of deep sentence representations obtained using USE and fine-tuning of Transformer models.

The main contributions of the work are as follows. First, we demonstrate that sentence level Transformer transfer learning for clustering gives good results on the range of datasets for STC. Second, fine-tuning of deep models for clustering is hindered because of the lack of labeled data and we propose to use the ECIC algorithm with deep Transformer models which has not been done before to tackle this problem. We called our method Transformer-based Enhancement of Clustering by Iterative Classification (TECIC). Third, we analyzed different combinations of components as constitutional parts of the algorithm, tested different schemes to handle weights during fine-tuning over iterations and developed a new stopping criterion for the algorithm.

2 Related work

One major direction in STC is based on Dirichlet multinomial mixture topic models [17, 15] including GSDPMM [18]. Some variants of these models incorporate word embeddings [6, 4, 15]. These models assume that each document contains only one or a few topics. The models have several advantages over conventional topic modeling such as latent Dirichlet allocation, when used for short texts. First, they better cope with the sparseness of short texts, which carry limited information about word co-occurrences. Second, these models can automatically infer the number of topics. Since only one topic is presented for each document, it is straightforward to use these topic models for clustering, assuming all documents with the same topic as belonging to the same cluster.

Recent works have considered a neural approach for STC. In [14, 12], authors propose to encode texts by pre-trained binary codes. Embeddings of words are then fed in the convolutional neural network which is trained to fit the binary codes. Finally, the obtained representations are used as features with k-means clustering algorithm. The work of [20] uses a somewhat similar strategy called Self-Taught Approach (STA). An autoencoder is pre-trained to obtain low-dimensional features and then learn it together with clustering algorithm by iteratively updating the weights of the autoencoder and centroids of clusters. Finally, they use the resulting features with k-means clustering algorithm. Another idea is to use attentive representation learning with adversarial training for STC [1]. The work of [3] sets the state-of-the-art results on the range of short text datasets using the ECIC algorithm which is simpler than in [20]. They use averaged word embeddings as features for short texts and clustering algorithms such as k-means, to get the initial label assignment. The clustering performance is then improved with iterative outlier detection and classification.

3 Model

In our work, we made several important modifications to the ECIC algorithm [3] to improve their results. Namely, we included modern deep learning components such as USE, BERT and RoBERTa in the algorithm as well tested various methods to handle weights during fine-tuning over iterations such as resumption and re-initialization and developed a new stopping criterion for the algorithm. The main steps of the algorithm are the following:

- Take a dataset D with N texts and K clusters.
- Apply initial clustering and labeling L.
- Set the number of iterations T.
- While $j \leq T$ and the stopping criterion δ is not reached do:
  - Sample P uniformly from $[P_1, P_2]$.
  - Apply outlier detection for each cluster from L to remove outliers from D.
  - If the number of texts in any cluster $n \geq P * N/K$ remove texts randomly from that cluster until $n \geq P * N/K$.
  - Add the rest of D to the train set and add all removed samples to the test set.

- Train a classifier on the train set and update L based on predictions of the classifier on the test set.
- Calculate the criterion δ and update j.

At the initial stage, clustering is carried out using one of the widely used clustering methods (see below). An algorithm for outlier detection is then used to split the dataset into train and test parts. Additional samples can be moved from the train to the test set based on the P number sampled randomly in the range from P_1 to P_2. The train part is used to train the classifier. Outliers and some number of the additional samples are used as a test set and predictions for the test set are used to relabel the dataset. Steps with outlier detection, classification, and relabeling are then repeated until the stopping criterion is reached or the maximum number of iterations is exceeded. As will be shown below, this iterative procedure leads to improved clustering results in many cases.

Averaged word embeddings were used as features in [3, 12]. One of the differences of our study is that we used USE representations[1] [16, 9] for short texts to plug them into one of the clustering algorithms: k-means, Hierarchical Agglomerative Clustering (HAC) or Spectral Clustering (SC). We did not consider the DBSCAN family of algorithms in this work since they infer the number of clusters automatically, while we studied the case when the number of clusters in a dataset is fixed. For k-means we chose the number of initializations 1000, the maximum number of iterations 300, the relative tolerance 10^{-4}. We used a full similarity matrix as well as k-NN and similarity distribution-based sparsification of the similarity matrix [5] with HAC. In both methods of sparsification, we set the number of non-zero elements in each row of the similarity matrix equal to the ratio of the number of samples in the dataset to the number of clusters. In addition, we tested all available linkage criteria for HAC such as single, complete, average, weighted, centroid and Ward. We used the euclidian metric with these criteria. For SC we chose ARPACK eigensolver, the stopping criterion for eigendecomposition of the Laplacian matrix equal 0, and the k-means strategy to assign labels in the embedding space with the number of initializations 10. We tried the Isolation Forest (IF) [8] and Local Outlier Factor (LOF) [7] for outlier detection. For IF we chose the number of base estimators in the ensemble 100. All train samples were used to train each estimator. The proportion of outliers in the dataset was determined automatically as in the original paper [8]. For LOC we chose the number of neighbours 20 and the euclidian metric. We used clustering and outlier detection algorithms implemented in the scikit-learn[2] and scipy[3] python libraries.

In contrast with [3], we used Transformer models such as BERT [2] and RoBERTa [11] for iterative fine-tuning and classification. For these models we used Adam optimizer and tried learning rates values among 2×10^{-5}, 3×10^{-5}, 5×10^{-5}. The number of training epochs per each iteration of the TECIC was varied among 2, 3 and 5. Constant and linear decay learning rate schedules were tested in different runs. We tried different models weight handling such as re-initialization after each iteration of the TECIC or resumption i.e. training with weights obtained at the previous iteration. We used batch size 32 and maximum sequence length 64 for both Transformer models. In addition, we used Multinomial Logistic Regression (MLR) as in other works. For MLR we tried different values of the maximum number of iterations for the solver to converge among 100, 1000, 10000 and the tolerance for stopping criteria among 10^{-4}, 10^{-5}, 10^{-6}. The rest of the parameters for MLR were taken as default in the scikit-learn.

The number of iterations T was set to be 10 for neural classifiers and 50 for MLR. We tried values for P_1 in the range from 0.5 to 0.8 and for P_2 in the range from 0.8 to 0.99. We consider two different stopping criteria. The first stopping criterion [3] is defined as follows $\delta = \frac{1}{N}\sum_i |c_i - c_i'| < \epsilon$ where c_i and c_i' are sizes of clusters determined by the current labeling L and previous labeling L', respectively, and i is a cluster number. For the first stopping criterion we tried ϵ equal to 0.03 and 0.05. The second criterion is reached immediately when δ has a minimum value.

[1] https://tfhub.dev/google/collections/universal-sentence-encoder/1

[2] https://scikit-learn.org/stable/index.html

[3] https://www.scipy.org/

Dataset	K	N	M
Stack Overflow	20	20000	8.2
AG News	4	8000	22.5
Biomedical corpus	20	20000	12.9
Search Snippets	8	12340	17.0
Tweet	89	2472	8.4
Google News TS	152	11109	28.0
Google News T	152	11109	6.2
Google News S	152	11109	21.8

Table 1: Statistics on the datasets used in the study. K is the number of clusters, N is the number of samples, M is the average number of words in a document.

4 Datasets

Our study uses the same datasets as those in a number of previous studies [10, 12, 20, 3] on STC. The statistics on the datasets are presented in Table 1. The Search Snippets dataset is composed of Google search results of 8 different domains [10]. The texts in the Search Snippets dataset represent sets of key words, rather than being coherent texts. The Biomedical corpus is a subset of one of the BioAsQ[4] challenge datasets, where 20000 paper titles were randomly selected from 20 groups [12]. The texts in this dataset contain special terms from biology and medicine. The Stack Overflow is a subset of the challenge data published on Kaggle[5], where question titles 20000 from 20 categories were randomly selected [12]. AG News is a subset of a news titles dataset that was used in [19], where 2000 samples of news titles with descriptions from each of the four categories were taken randomly. The Tweet dataset consists of 2472 tweets which are highly relevant to 89 queries [18]. The Google News TS consists of 11109 news articles titles and snippets about 152 events, while T version of the dataset contains only titles and S contains only snippets of these articles [18].

Note that the former and the latter four datasets can be grouped by the number of clusters. The first group contains relatively low numbers of clusters, while the second has greater numbers of clusters.

5 Results

To measure the performance of our algorithm, we used such metrics as accuracy and Normalized Mutual Information (NMI). The same metrics were used in the number of the previous studies [12, 20, 3, 15]. The value of NMI does not depend on the absolute values of labels. The accuracy is calculated using the Hungarian algorithm [12]. It allows one to rearrange absolute label values to maximize accuracy.

Our experiments on initial clustering tested which of the USE versions and which clustering algorithm should be used to obtain the best quality in terms of both aforementioned metrics. As a result, the old version of USE [16] proved to be better (by a few percent) than the newer one [9] in terms of both metrics on all 8 datasets. We tested k-means, HAC, and Spectral Clustering algorithms with these sentence embeddings. Interestingly, we found that the best clustering method was k-means for the whole group of datasets with the smaller number of clusters (see Table 2). Since k-means is not a deterministic algorithm and its result depends on a particular initialization, we averaged the results over 5 runs. On the contrary, HAC proved to be the best clustering method for datasets with the greater number of clusters (see Table 3). Note we does not provide variance for HAC since this algorithm is determenistic. Overall, k-NN sparsification with the average linkage criterion gave the best results for the four datasets with the greater number of clusters. This differs from the results of [3], where a sparsification based on similarity distribution and the Ward linkage criterion are described as the most effective ones.

[4] http://bioasq.org

[5] https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow/download/train.zip

Method	Metric	Stack Over.	AG News	Biomedical c.	Search Snip.
k-means	Acc.	**81.84±0.01**	**83.87±0.02**	**43.84±0.20**	**74.76±0.13**
	NMI	**80.80±0.01**	**61.88±0.04**	**37.85±0.13**	**54.25±0.16**
HAC Ward	Acc.	74.90	56.89	36.74	61.64
full	NMI	75.73	51.45	32.60	49.79
HAC Ward	Acc.	80.23	57.21	41.27	55.93
d.-b. spar.	NMI	80.44	58.48	37.00	50.23
HAC Ward	Acc.	80.53	58.10	41.79	64.98
k-NN spar.	NMI	80.69	55.74	36.75	52.59

Table 2: Comparison of accuracy and NMI scores for various clustering algorithms for datasets with the smaller number of clusters. Four best performing algorithms are presented.

Method	Metric	Tweet	G. News TS	G. News T	G. News S
HAC weight.	Acc.	**81.59**	64.81	62.95	70.71
full	NMI	**91.97**	86.49	85.95	91.03
HAC Ward	Acc.	74.11	74.71	**80.09**	**85.72**
full	NMI	90.49	90.21	90.67	94.47
HAC aver.	Acc.	78.20	**84.64**	77.56	80.34
k-NN spar.	NMI	91.28	**94.77**	91.14	91.96
HAC weight.	Acc.	74.96	79.56	79.15	83.95
k-NN spar.	NMI	90.42	91.28	**91.23**	**94.39**

Table 3: Comparison of accuracy and NMI scores for various clustering algorithms for datasets with the larger number of clusters. Four best performing algorithms are presented.

We obtained highly competitive results for two (Stack Overflow and AG News) of the four datasets from the first group of datasets (see Table 4). However, we did not get comparable results on the other two datasets (Search Snippets and Biomedical corpus), which can be easily explained. The Search Snippets dataset texts are sets of key words, rather than being coherent texts. Since USE was trained on coherent texts, it cannot produce a good result. The Biomedical dataset almost completely consists of special terms. USE probably did not see many of these terms during training, which explains its poor performance on this dataset. We got the best results for all four datasets from the second group in terms of NMI but not in terms of accuracy (see Table 5).

To improve the results of initial clustering, we tested the iterative classification algorithm with MLR and with neural pre-trained classifiers, such as BERT and RoBERTa. For the neural classifier, the best value for the learning rate was found to be 3×10^{-5} and the number of epochs to train during each iteration was found to be 2. The use of the warm start i.e. training resumption after each iteration instead of re-initialization, and learning rate linear decaying schedule instead of the constant learning rate, did not show any considerable improvement. RoBERTa gave approximately one half percent improvement over the BERT performance. We set T to be 50 for MLR, since the algorithm worked more stable and had potential to improve for the more iterations than for neural classifiers. We found that the use of the second stopping criterion with neural classifiers gives better results than the first one. We did not use any criterion for MLR and collected the metrics at the end of 50 iterations, since both considered metrics grew monotonically for this classifier. We found that P_1 equal 0.7 and P_2 equal 0.95 were the best values for both types of classifiers. We averaged our results over 3 runs in both cases. We did not find any difference in the use of IF or LOF for outlier detection with all classifiers.

The iterative classification achieved the state-of-the-art results on the Stack Overflow and AG News datasets with both types of classifiers and improved the good initial clustering result further (see Table 4).

Method	Metric	Stack Over.	AG News	Biomedical c.	Search Snip.
ECIC [3]	Acc.	78.73±0.17	84.52±0.50	47.78±0.51	**87.67±0.63**
	NMI	73.44±0.35	59.07±0.84	41.27±0.36	**71.93±1.04**
STA [20]	Acc.	59.8±1.9	-	**54.8±2.3**	77.1±1.1
	NMI	54.8±1.0	-	**47.1±0.8**	56.7±1.0
Init. clust.	Acc.	**81.84±0.01**	83.87±0.02	43.84±0.20	74.76±0.13
k-means	NMI	**80.80±0.01**	**61.88±0.04**	37.85±0.13	54.25±0.16
Iter. class.	Acc.	**84.72±0.20**	**84.64±0.08**	44.85±0.20	74.97±0.15
RoBERTa	NMI	**80.63±0.97**	**62.69±0.20**	38.40±0.13	55.17±0.26
Iter. class.	Acc.	**83.31±0.05**	**86.53±0.1**	44.96±0.17	75.87±0.15
Log. Reg.	NMI	**80.68±0.01**	**65.99±0.28**	39.18±0.04	57.36±0.08

Table 4: Comparison with published results of accuracy and NMI scores for datasets with the smaller number of clusters.

Method	Metric	Tweet	G. News TS	G. News T	G. News S
PYPM [13]	NMI	89.8±0.5	**94.9±0.1**	89.0±0.3	91.6±0.2
ECIC [3]	Acc.	**91.52±0.99**	93.56±0.27	**87.18±0.21**	**89.02±0.12**
	NMI	86.87±0.13	94.40±0.11	87.87±1.00	89.96±0.11
GSDPMM [18]	NMI	87.5±0.5	91.2±0.3	87.3±0.2	89.1±0.4
Init. clust.	Acc.	78.20	84.64	77.56	80.34
HAC	NMI	**91.28**	94.77	**91.14**	**91.96**

Table 5: Comparison with published results of accuracy and NMI scores for datasets with the larger number of clusters.

The neural classifier showed a one percent better performance for the Stack Overflow in terms of accuracy than MLR. We did not get comparable results for the Biomedical and Search Snippets datasets, since the iterative classification algorithm can improve the initial clustering result by a limited number of percent and it was low efficient for these two datasets. We did not observe any improvement for the second group of datasets, since it is more difficult for the algorithm to converge to the correct solution during iterations in the case of greater number of clusters.

6 Conclusions

The sentence embeddings based algorithm for enhanced clustering by iterative classification was applied to 8 datasets with short texts. The algorithm demonstrates state-of-the-art results for the 5 out of 8 datasets in terms of NMI and for 2 of 8 in terms of accuracy. We argue that the lack of coherent and common texts causes an inferior performance of the algorithm for the remaining datasets.

The quality of the whole algorithm strongly depends on the initial clustering quality. Initial clustering with USE representations has already allowed us to achieve a competitive performance for a number of datasets. Therefore, due to transfer learning these representations can be readily applied to other datasets even without iterative classification.

References

[1] Attentive Representation Learning with Adversarial Training for Short Text Clustering / Wei Zhang, Chao Dong, Jianhua Yin, Jianyong Wang // arXiv preprint arXiv:1912.03720. — 2019.

[2] Bert: Pre-training of deep bidirectional transformers for language understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.

[3] Enhancement of Short Text Clustering by Iterative Classification / Md Rashadul Hasan Rakib, Norbert Zeh, Magdalena Jankowska, Evangelos Milios // Natural Language Processing and Information Systems / Ed. by Elisabeth Métais, Farid Meziane, Helmut Horacek, Philipp Cimiano. — Cham : Springer International Publishing, 2020. — P. 105–117.

[4] Enhancing topic modeling for short texts with auxiliary word embeddings / Chenliang Li, Yu Duan, Haoran Wang et al. // ACM Transactions on Information Systems (TOIS). — 2017. — Vol. 36, no. 2. — P. 1–30.

[5] Improving Short Text Clustering by Similarity Matrix Sparsification / Md Rashadul Hasan Rakib, Magdalena Jankowska, Norbert Zeh, Evangelos Milios // Proceedings of the ACM Symposium on Document Engineering 2018. — 2018. — P. 1–4.

[6] Improving topic models with latent feature word representations / Dat Quoc Nguyen, Richard Billingsley, Lan Du, Mark Johnson // Transactions of the Association for Computational Linguistics. — 2015. — Vol. 3. — P. 299–313.

[7] LOF: identifying density-based local outliers / Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, Jörg Sander // Proceedings of the 2000 ACM SIGMOD international conference on Management of data. — 2000. — P. 93–104.

[8] Liu Fei Tony, Ting Kai Ming, Zhou Zhi-Hua. Isolation forest // 2008 Eighth IEEE International Conference on Data Mining / IEEE. — 2008. — P. 413–422.

[9] Multilingual Universal Sentence Encoder for Semantic Retrieval / Yinfei Yang, Daniel Cer, Amin Ahmad et al. // CoRR. — 2019. — Vol. abs/1907.04307. — 1907.04307.

[10] Phan Xuan-Hieu, Nguyen Le-Minh, Horiguchi Susumu. Learning to classify short and sparse text & web with hidden topics from large-scale data collections // Proceedings of the 17th international conference on World Wide Web. — 2008. — P. 91–100.

[11] Roberta: A robustly optimized bert pretraining approach / Yinhan Liu, Myle Ott, Naman Goyal et al. // arXiv preprint arXiv:1907.11692. — 2019.

[12] Self-taught convolutional neural networks for short text clustering / Jiaming Xu, Bo Xu, Peng Wang et al. // Neural Networks. — 2017. — Vol. 88. — P. 22–31.

[13] Short text clustering based on Pitman-Yor process mixture model / Jipeng Qiang, Yun Li, Yunhao Yuan, Xindong Wu // Applied Intelligence. — 2018. — Vol. 48, no. 7. — P. 1802–1812.

[14] Short text clustering via convolutional neural networks / Jiaming Xu, Peng Wang, Guanhua Tian et al. // Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. — 2015. — P. 62–69.

[15] Short text topic modeling techniques, applications, and performance: a survey / Qiang Jipeng, Qian Zhenyu, Li Yun et al. // arXiv preprint arXiv:1904.07695. — 2019.

[16] Universal Sentence Encoder / Daniel Cer, Yinfei Yang, Sheng-yi Kong et al. // CoRR. — 2018. — Vol. abs/1803.11175. — 1803.11175.

[17] Yin Jianhua, Wang Jianyong. A dirichlet multinomial mixture model-based approach for short text clustering // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2014. — P. 233–242.

[18] Yin Jianhua, Wang Jianyong. A model-based approach for text clustering with outlier detection // 2016 IEEE 32nd International Conference on Data Engineering (ICDE) / IEEE. — 2016. — P. 625–636.

[19] Zhang Xiang, LeCun Yann. Text understanding from scratch // arXiv preprint arXiv:1502.01710. — 2015.

[20] A self-training approach for short text clustering / Amir Hadifar, Lucas Sterckx, Thomas Demeester, Chris Develder // Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). — 2019. — P. 194–199.

Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection

Maxim Rachinskiy
HSE University
Moscow, Russia
myurachinskiy@edu.hse.ru

Nikolay Arefyev
Samsung Research Center Russia
Lomonosov Moscow State University
HSE University
Moscow, Russia
narefjev@cs.msu.ru

Abstract

Consulting word definitions from a dictionary is a familiar way for a human to find out which senses a particular word has. We hypothesize that a system that can select a proper definition for a particular word occurrence can also naturally solve Semantic Change Detection (SCD) task. To verify our hypothesis, we followed an approach previously proposed for Word Sense Disambiguation (WSD) and trained a system that embeds word definitions and word occurrences into the same vector space. In this space, the embedding of the most appropriate definition has the largest dot product with a contextualized word embedding.

The system is trained on an English WSD corpus. To make it work for the Russian language, we replaced BERT with the multilingual XLM-R language model and exploited its zero-shot cross-lingual transferability. Despite not finetuning the encoder model on any Russian data, this system achieves the second place in the competition, and likely works for any of one hundred other languages XLM-R was pre-trained on, though the performance may vary. We then measure the impact of such WSD pre-training and show that this procedure is crucial for our results. Since our model was trained to choose a proper definition for a word, we propose an algorithm for the interpretation and visualization of the semantic changes through time.

By employing additional labeled data in Russian and training a simple regression model, that converts the distances between output contextualized embeddings into more human-like scores of sense similarity between word occurrences, we further improve our results and achieve the first place in the competition.

Keywords: Semantic change detection, SCD, gloss-informed models, GLM
DOI: 10.28995/2075-7182-2021-20-578-586

Межъязыковой перенос без дообучения толковой языковой модели для обнаружения семантических сдвигов

Рачинский Максим[◇]
Москва, Россия
myurachinskiy@edu.hse.ru

Арефьев Николай[△▽◇]
Москва, Россия
narefjev@cs.msu.ru

[◇]Национальный исследовательский университет «Высшая школа экономики»
[△]Московский Исследовательский Центр Самсунг
[▽]Московский Государственный Университет им. М. В. Ломоносова

Аннотация

Обращение к определениям из словаря — это привычный для человека способ выяснить, какие значения имеет то или иное слово. Мы предполагаем, что система, которая может выбрать из толкового словаря или глоссария правильное определение для конкретного вхождения слова, также может естественным образом решить задачу обнаружения изменений значений слов с течением времени (семантических сдвигов). Чтобы проверить нашу гипотезу, мы использовали подход, ранее предложенный для разрешения лексической многозначности (WSD), и обучили систему, которая проецирует определения слов и их вхождения в тексты в одно и то же векторное пространство. В этом пространстве вектор наиболее подходящего определения имеет самое большое скалярное произведение с контекстуализированным вектором вхождения слова.

Система обучается разрешать лексическую многозначность (выбирать самое подходящее определение) на англоязычном корпусе. Для того чтобы работать с текстами на русском языке, мы заменили англоязычный BERT на многоязычную языковую модель XLM-R и использовали ее способность к межъязыковому переносу. Несмотря на отсутствие дообучения модели на каких-либо данных на русском языке, такая система заняла второе место в соревновании и, вероятно, работает на любом из ста других языков, на которых

XLM-R был предварительно обучен, хотя в зависимости от языка качество может варьироваться. Мы оцениваем влияние обучения модели выбору наиболее подходящего определения и показываем, что эта процедура имеет решающее значение для наших результатов. Поскольку наша модель была обучена подбору правильного определения слова, мы используем это свойство и предлагаем метод интерпретации и визуализации семантических сдвигов во времени.

Используя дополнительные размеченные данные на русском языке и обучая простую регрессионную модель, которая преобразует расстояния между контекстуализированными векторами вхождений слов в оценки смыслового сходства, близкие к человеческим, мы улучшили наши результаты. Дообученная на русскоязычных данных система заняла первое место в соревновании.

Ключевые слова: семантические сдвиги, модели на основе определений

1 Introduction

RuShiftEval [6] is a semantic change detection task for the Russian language.[1] Each test sample in the competition consisted of a single Russian word. The participants were asked to predict how much test words have changed their meanings between three epochs: pre-Soviet, Soviet and post-Soviet. The mean of the three Spearman correlation coefficients of the predicted and gold scores was utilized as the main performance metric.

Through the evaluation period, our model which did not use any Russian data for finetuning achieved the second place in the competition. As there was no domain adaptation, this system likely works for any of one hundred other languages XLM-R was pre-trained on, though the performance may vary. After adding labeled data in Russian and training a simple regression model, that converts the distances between output contextualized embeddings we achieved the first place.[2]

Our main interest was whether the semantic change detection systems can benefit from using gloss information and how these systems can be interpreted.

2 Background

Here we summarize the prior work linking word occurrences and word definitions. One of the first approaches in this field [7] calculated the lexical overlap between the context of a particular word occurrence and all possible definitions of this word. This approach did not take into account word synonymy or other lexical relations. The recent works tried to combine state-of-the-art language models with glosses from some dictionaries.

One of such methods has been proposed in [15]. Their EWISE system used a pre-training procedure for a gloss encoder, that learned knowledge graph embeddings from WordNet [8]. After this pre-training, the authors froze the gloss encoder and started to train a context encoder with labeled WSD data. The ablation study of this work has shown the importance of such gloss encoder pre-training.

While the previous method requires relational information from a knowledge graph, the method proposed in [5] relies fully on gloss information. The developed system jointly encodes the context with all possible glosses of the target word. The authors used a pre-trained BERT [1] model as initialization for their encoder. The results demonstrated a big gap in the performance between the developed method and a simple context encoder without any gloss information.

A similar approach has been proposed in [2], where authors trained two separate Transformer-based encoders for word occurrences (Context encoder) and word definitions (Gloss encoder), both initialized with BERT weights [1]. To represent a word occurrence, the outputs of the Context encoder for all of its subwords were averaged. To represent a definition, the output of the Gloss encoder from [CLS] token was taken. Finally, for a word occurrence and all of its definitions, the dot products between those outputs were calculated and the softmax function was applied to them, resulting in a probability distribution over possible word senses. The whole model was trained using cross-entropy loss to select the correct word sense on WSD data.

While BEM [2] and GlossBERT [5] are based on BERT [1] encoder, our system exploits XLM-RoBERTa (XLM-R) [13] architecture. XLM-R is based on RoBERTa [10] and is pre-trained on unlabeled

[1] `https://competitions.codalab.org/competitions/28340`

[2] In order to make our results reproducible, we publish the code of our experiments: `https://github.com/myrachins/RuShiftEval`.

ID	Model	P1	P2	P3	Aver.
GLM, zero-shot cross-lingual transfer to Russian					
1	Manhattan+norm GLM xlmr.large	74.1	77.9	79.8	77.3
GLM + regression to human scores trained on RuSemShift					
2	Linear regression on GLM xlmr.large distances	77.0	80.1	81.8	79.6
3	Linear regression on GLM xlmr.large+base distances	78.1	**80.3**	**82.2**	**80.2**
4	Knn regression on GLM xlmr.large+base distances	71.8	76.2	80.9	76.3
5	Random.forest regr. (1K est.) on GLM xlmr.large+base dist.	75.2	78.7	81.6	78.5
6	Random.forest regr. (2K est.) on GLM xlmr.large+base dist.	75.0	78.7	81.7	78.5
7	Random.forest regr. (5K est.) on GLM xlmr.large+base dist.	75.8	78.4	81.3	78.5
The top3 best results of other teams					
-	DeepMistake	**79.8**	77.3	80.3	79.1
-	vanyatko	67.8	74.6	73.7	72.0
-	aryzhova	46.9	45.0	45.3	45.7

Table 1: Test Spearman score for each of our submissions. P1, P2, P3 columns stand for the pre-Soviet - Soviet, Soviet - post-Soviet and pre-Soviet - post-Soviet pairs respectively. The features for 2-7 submissions were taken from 7 different distance measures: L1, L1+norm, L2, L2+norm, Dot product, Dot product+norm(Manh), Cosine. Top-head models from 2-7 submissions were trained with RuSemShift data [11]. GLM pre-trained XLM-RoBERTa encoders were used in a feature-based setting and were not fine-tuned.

data with MLM (Masked Language Modeling) objective. But in contrast to BERT and RoBERTa, it is pre-trained not on monolingual data but on 2.5 TB of texts from CommonCrawl in 100 languages.

3 System overview

Our approach employs contextualized embeddings obtained from a gloss-based Word Sense Disambiguation (WSD) model to measure an average sense similarity between occurrences of a particular word in two corpora. This model was pre-trained with Gloss Language Modeling (GLM) procedure, which we will discuss further in detail and compare with pure Masked Language Modeling (MLM) pre-training.

Based on the previous observations, that the strongest signal in the contextualized embeddings of a language model pre-trained with MLM corresponds to the word form, not the word meaning [4], we try to fix it by fine-tuning the model to select a gloss (a definition) from WordNet, that is most appropriate for a particular word occurrence. We call this model a Gloss Language Model (GLM) and show that this training procedure results in much more appropriate contextualized embeddings for the SCD task. To initialize the GLM before training, we use the weights of the XLM-R model, which was pre-trained with MLM objective on 2.5TB of texts from 100 languages [13]. Despite using only English WSD data for training, our model still produces sensible contextualized embeddings for Russian, which alone gives a strong performance in the SCD task. Since we do not use any Russian data or resources for the pure GLM-based SCD model, this model will likely work for other languages too, though the performance may vary.

To solve the SCD task, we sample sentences containing a particular target word from each of the given three epochs. Finally, for each pair of epochs, we calculate the average distance between contextualized word embeddings from GLM.

3.1 Gloss-informed embeddings

In order to learn sense-dependent representations of words, we pre-train our system on the Word Sense Disambiguation task. Following the BEM model [2], our system consists of two separate encoders: Context Encoder and Gloss Encoder.

Context encoder (T_c) takes a sentence $c = c_0, \ldots, c_{i-1}, w_c, c_{i+1}, \ldots, c_n$ containing a target word w_c to be disambiguated, where w_c is the i^{th} word in the sentence. The encoder then produces the target word

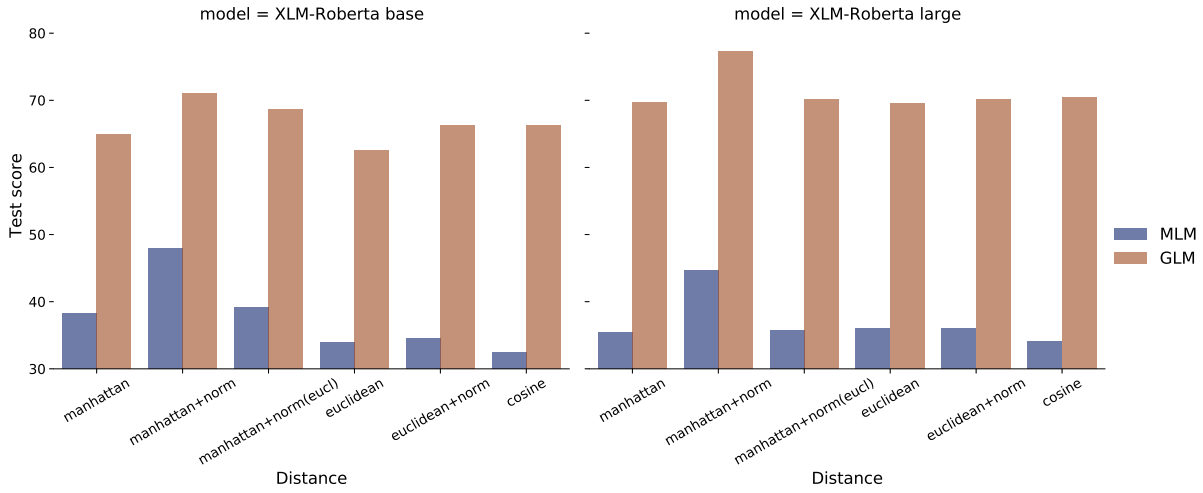


Figure 1: Mean test Spearman score for the GLM and MLM models for each distance measure.

representation:

$$r_{w_c} = T_c(c)[i]$$

For target words that are tokenized into multiple subword units, we average representations of these subwords.

Gloss encoder (T_g) takes as input a gloss $g_s = g_0, g_1, ..., g_m$ that defines a word sense s and encodes it as:

$$r_s = T_g(g_s)[0]$$

Taking the output from the first input token, which should be [CLS] or <s> token.

We can score each of the possible senses $s \in S_w$, for a target word w_c by taking the dot product of r_{w_c} against every r_s for $s \in S_w$:

$$\phi(w_c, s) = r_{w_c}^T r_s$$

As there is no such big WSD dataset as SemCor [14] for non-English languages, we extend BEM [2] system to the multilingual setting by replacing BERT with XLM-RoBERTa model [13] and exploiting its zero-shot cross-lingual transferability.

Both encoders were initialized with XLM-R base or large weights. Then the whole system was pre-trained on WSD data with cross-entropy loss. We denote this pre-training procedure as Gloss Language Modeling (GLM). In our experiments, these encoder models were not fine-tuned on any Russian data.

3.2 Sentence sampling

Following the competition's recommendations, we exploited Russian National Corpus (RNC) to sample sentences from the given epochs. We used rulemma[3] lemmatizer to find all occurrences of the target words in all forms in the corpus. For each target word, we sampled no more than 100 sentences per epoch.

3.3 Inter-epoch difference

In order to calculate desired $compare_{e_1,e_2}(w)$ value which denotes predicted compare metric for the word w in e_1, e_2 epochs pair, we build contextualized word representations from the sampled sentences and calculate an inter-corpus average distance between them. More precisely, we compute $d(w_{c_{e_1}}, w_{c_{e_2}})$ which is a distance between contextualized embeddings obtained from the context encoder of our WSD

[3] `https://github.com/Koziev/rulemma`

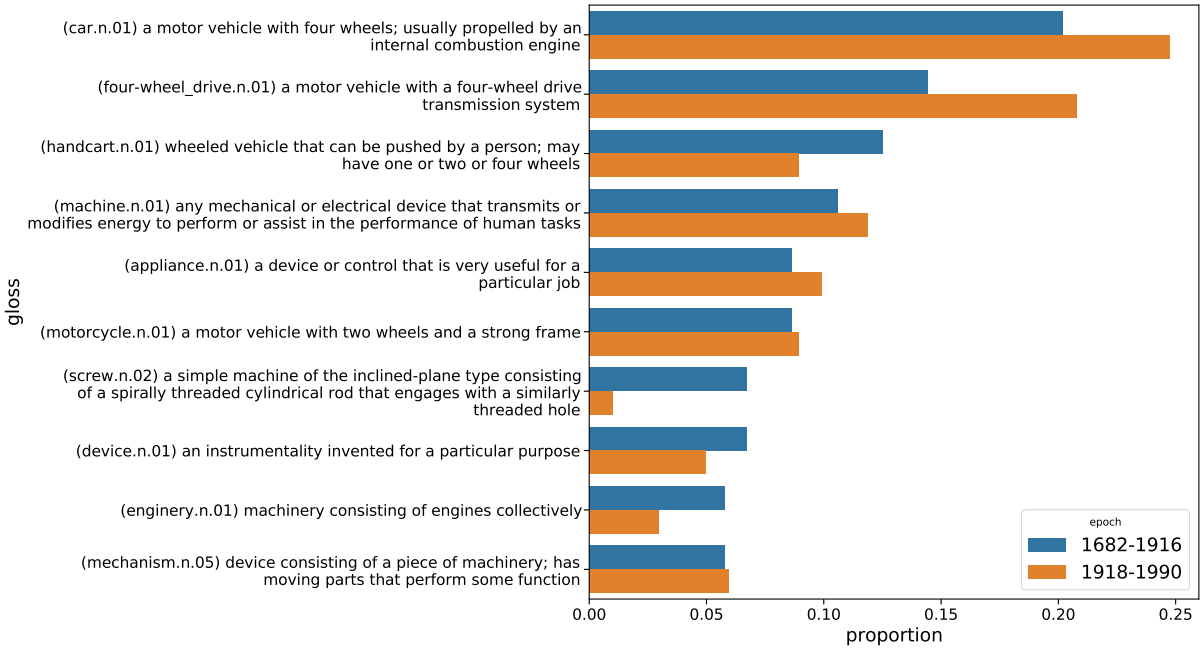


Figure 2: A proportion of examples with the word *машина* (car, vehicle, engine, computer) from the RuSemShift [11] samples where a particular gloss was selected (in top3) by GLM model. As this word did not occur in the post-Soviet part of RuSemShift [11], we show only the first two epochs.

pre-trained system for the word entries $w_{c_{e_1}}$ and $w_{c_{e_2}}$ of the word w from epochs e_1 and e_2 respectively. $compare_{e_1,e_2}(w)$ is calculated as average $d(\cdot)$ for all sampled sentence pairs $S_w(e_1, e_2)$ with the word w from the considered epochs pair e_1, e_2.

During the competition, we experimented on the $d(\cdot)$ definition based on our gloss-informed models. Here we propose methods that fully rely on the Context encoder and thus do not require any additional vocabulary or glosses. We achieve such generalization by using only outputs from the trained Context encoder.

1. **Euclidian (L2)**: Euclidian distance between outputs of the encoder.
2. **Euclidian+norm**: Euclidian distance between L2 normalized outputs of the encoder.
3. **Manhattan (L1)**: Manhattan distance between outputs of the encoder.
4. **Manhattan+norm**: Manhattan distance between L1 normalized outputs of the encoder.
5. **Manhattan+norm(Eucl)**: Manhattan distance between L2 normalized outputs of the encoder.
6. **Dot product**: Dot product similarity between outputs of the encoder.
7. **Dot product+norm(Manh)**: Dot product similarity between L1 normalized outputs of the encoder.
8. **Cosine**: Cosine similarity between outputs of the encoder.

As the bigger L1 or L2 distance means the bigger semantic change, to get the positive Spearman correlations with the gold scores we inverted our final average distances.

Instead of a single distance function, we can use a trainable combination of them. During the competition, we trained several regression models with features taken from 7 different distance measures: L1, L1+norm, L2, L2+norm, Dot product, Dot product+norm(Manh), Cosine. The models were trained to approximate human scores for the RuSemShift [11] sentence pairs.

Besides comparing distance measures, we also experimented on the encoders' initialization. In the result section, we compare the performance of the models, initialized with XLM-R base and XLM-R large [13].

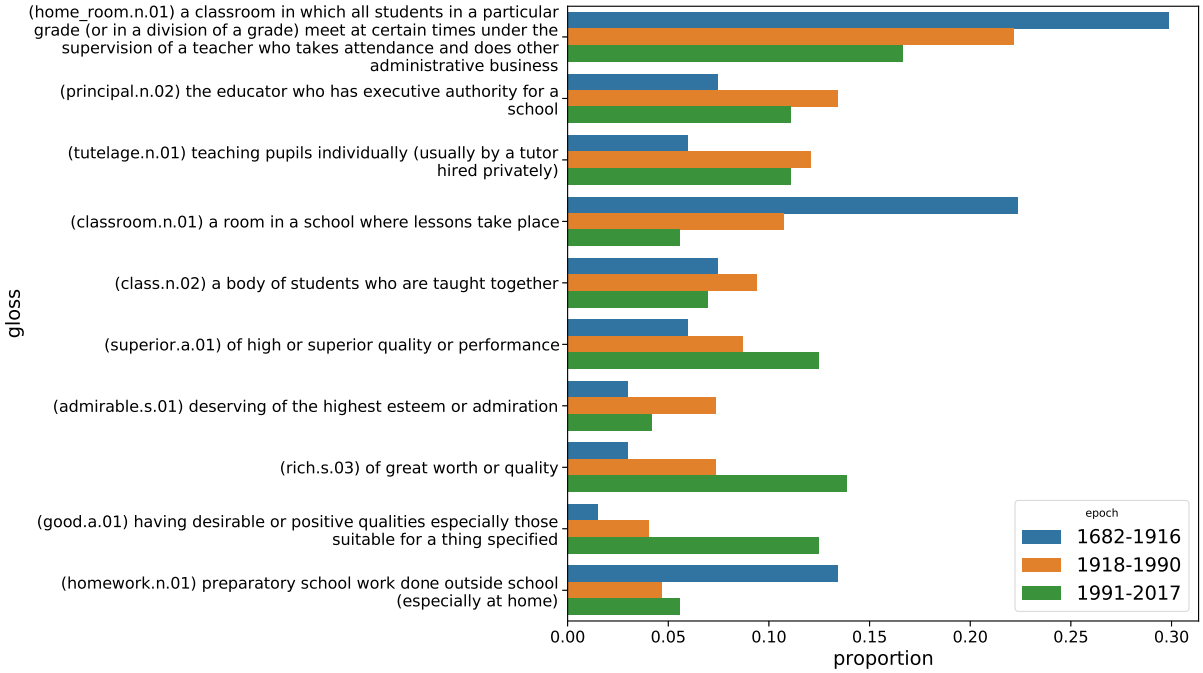


Figure 3: A proportion of examples with the word *классный* (classroom, cool, classy) from the RuSemShift [11] samples where a particular gloss was selected (in top3) by GLM model.

4 Experiments and Results

We trained our models on English SemCor [14] with glosses from WordNet 3.0 [8]. Systems based on XLM-RoBERTa base and XLM-RoBERTa large [13] were trained 20 and 10 epochs respectively. The Context and Gloss encoders were optimized on separate V100 GPUs for about 3 days for each backbone. Following standard practices, we used SemEval-2007 [12] as our development set to choose the final checkpoints. We evaluated our systems with the WSD framework proposed in [9].

4.1 GLM vs MLM

Figure 1 shows the gap between MLM and GLM pre-training, where we use distances between contextualized word embeddings obtained from the Context encoder pre-trained to solve the WSD task. Experiments show that models trained with GLM procedure strongly outperform their MLM counterparts regardless of distance measure or backbone. We also see that for both backbones manhattan+norm distance performs the best. In addition, the figure shows that GLM large model outperforms the base counter-part, but with MLM pre-training the base model performs slightly better.

4.2 Submissions

Table 1 shows overall results for the competition's test set for each of our submissions.

5 Qualitative analysis

As our GLM models were pre-trained to choose a proper definition for a target word, it is natural to try to interpret some of the models predictions. With the gloss knowledge from WordNet [8] we tried to find out how the meanings of the words were changing through time. We took the set of considered words from the book [3], where authors described the history of 20 Russian words. For each considered word w we run the following algorithm:

1. We sample all sentences with the target word w from RuSemShift [11].

Figure 4: Meanings' shift charts for the words *машина* (car, vehicle, engine, computer) and *классный* (classroom, cool, classy) from the book [3]. We also provide brief English translations of the columns and the words' meanings.

2. For each of these sentences we retrieve 3 glosses from WordNet having the maximum dot product with the contextualized embedding of the target word.
3. For each gloss we can calculate a proportion of examples with the target word w where this particular gloss was selected (in top3) by GLM model.

Figures 2, 3 show the dynamic of the meanings' changes for the words *машина* (car, vehicle, engine, computer) and *классный* (classroom, cool, classy) respectively. We took a pre-trained model with XLM-R large backbone for this purpose.

As we can see from these figures our model can choose sensible English glosses even from all WordNet [8] synsets, although the target words and their contexts are in Russian. Moreover, for these particular words, we can see consistency with the charts from the book [3] (Figure 4).

However, this approach with decoding Russian word senses with English WordNet [8] has several limitations. We have seen one of them during experiments with the word *пионер* (pioneer, scout), which of course drastically changed its meaning in the Soviet epoch. The Soviet meaning of this word is strongly connected to the Communist ideology, but in English, the nearest concept for this Soviet meaning is *scout*, which of course doesn't mean exactly the same, and consequently interpretation model can not find the proper English gloss and this leads to poor performance on such words.

6 Conclusion

In this paper, we proposed training a Gloss Language Model (GLM) to obtain better contextualized embeddings for the Russian Semantic Change Detection task. We have shown that this training procedure greatly boosts the performance compared to the traditional embeddings from a Masked Language Model (MLM) regardless of the distance measure employed.

Apart from that, we proposed a technique for the interpretation and visualization of the semantic

changes through time by linking Russian word occurrences to the reasonable definitions from the English WordNet and comparing distributions over those definitions for each epoch. Also, we discussed the limitations of this algorithm due to the difficult-to-translate concepts.

Acknowledgments

This research was supported in part through computational resources of HPC facilities at NRU HSE.

References

[1] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // 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). — Minneapolis, Minnesota : Association for Computational Linguistics, 2019. — Jun. — P. 4171–4186. — Access mode: https://www.aclweb.org/anthology/N19-1423.

[2] Blevins Terra, Zettlemoyer Luke. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders // Proceedings of the 58th Association for Computational Linguistics. — 2020. — Access mode: https://blvns.github.io/papers/acl2020.pdf.

[3] Daniehl' M. A., Dobrushina N. R. Dva veka v dvadtsati slovakh. — NRU Higher School of Economics Publ. House, 2016. — ISBN: 9785759811480.

[4] Laicher Severin, Kurtyigit Sinan, Schlechtweg Dominik et al. Explaining and Improving BERT Performance on Lexical Semantic Change Detection. — 2021. — 2103.07259.

[5] GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge / Luyao Huang, Chi Sun, Xipeng Qiu, Xuanjing Huang // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 3507–3512. — Access mode: https://www.aclweb.org/anthology/D19-1355.

[6] Kutuzov Andrey, Pivovarova Lidia. RuShiftEval: a shared task on semantic shift detection for Russian // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2021.

[7] Lesk Michael. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone // Proceedings of the 5th Annual International Conference on Systems Documentation. — SIGDOC '86. — New York, NY, USA : Association for Computing Machinery, 1986. — P. 24–26. — Access mode: https://doi.org/10.1145/318723.318728.

[8] Miller George A. WordNet: A Lexical Database for English // Commun. ACM. — 1995. — Nov. — Vol. 38, no. 11. — P. 39–41. — Access mode: https://doi.org/10.1145/219717.219748.

[9] Raganato Alessandro, Camacho-Collados Jose, Navigli Roberto. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. — Valencia, Spain : Association for Computational Linguistics, 2017. — Apr. — P. 99–110. — Access mode: https://www.aclweb.org/anthology/E17-1010.

[10] Liu Yinhan, Ott Myle, Goyal Naman et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. — 2019. — 1907.11692.

[11] Rodina Julia, Kutuzov Andrey. RuSemShift: a dataset of historical lexical semantic change in Russian // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 1037–1047. — Access mode: https://www.aclweb.org/anthology/2020.coling-main.90.

[12] SemEval-2007 Task-17: English Lexical Sample, SRL and All Words / Sameer Pradhan, Edward Loper, Dmitriy Dligach, Martha Palmer // Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). — Prague, Czech Republic : Association for Computational Linguistics, 2007. — Jun. — P. 87–92. — Access mode: https://www.aclweb.org/anthology/S07-1016.

[13] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // arXiv preprint arXiv:1911.02116. — 2019.

[14] Using a Semantic Concordance for Sense Identification / George A. Miller, Martin Chodorow, Shari Landes et al. // Proceedings of the Workshop on Human Language Technology. — HLT '94. — USA : Association for Computational Linguistics, 1994. — P. 240–243. — Access mode: https://doi.org/10.3115/1075812.1075866.

[15] Zero-shot Word Sense Disambiguation using Sense Definition Embeddings / Sawan Kumar, Sharmistha Jat, Karan Saxena, Partha Talukdar // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 5670–5681. — Access mode: https://www.aclweb.org/anthology/P19-1568.

Switching to Work in an Inclusivity Workshop: Multimodal Analysis of Interaction

Rudneva E. A.
Institute for Linguistic Studies,
St. Petersburg, Russia
katja1985mt@yandex.ru

Abstract

The study focuses on switching from talk to work in an "inclusivity workshop" for people with mental disabilities. Work activities and conversation about general topics can be approached from the perspective of *multiactivity* and considered courses of actions intertwined in social interaction. The order of activities is negotiated among participants using both linguistic and non-linguistic means. The data are extracts of video recordings containing a participant getting others to do things. The paper provides multimodal analysis of 6 cases of an instructor getting an autistic participant to switch to work, which occurred within a 17-minute conversation about animals. In the data, the autistic participant never provides a second-pair response to a directive. In 5 out of 6 cases analysed in the paper he fulfils the action to different extents, demonstrating various degrees of involvement. Getting the autistic person to switch to work is more effective when suggesting actions one by one, through concrete embodied actions, and when orienting to phases of the ongoing talk. The study highlights differences between autistic and non-autistic participants switching from one course of actions to another. Considering goals of an inclusivity workshop, success of switching to work can be also determined by the opportunities for the smooth conversation.

Key words: multimodality; multimodal analysis; multiactivity; autism
DOI: 10.28995/2075-7182-2021-20-587-596

Переключение к рабочей деятельности в инклюзивной мастерской: мультимодальный анализ взаимодействия

Руднева Е. А.
Институт лингвистических исследований РАН,
Санкт-Петербург, Россия
katja1985mt@yandex.ru

Аннотация

Исследование посвящено переключению от разговора к работе в «инклюзивной мастерской для людей с особенностями развития психики и интеллекта». Рабочая активность и общение с коллегами на отвлеченные темы могут рассматриваться как виды деятельности, которые пересекаются в социальном взаимодействии. Согласование действий между собеседниками происходит при помощи языковых и неязыковых средств. Материал составили фрагменты видеозаписей, где происходит побуждение к действию. В статье анализируются 6 попыток инструктора подвигнуть сотрудника с аутизмом возобновить рабочую активность, которые совершаются в течение 17-минутного разговора о животных. В 5 из 6 случаев сотрудник выполняет действие в той или иной степени, переключаясь к работе скорее поэтапно и демонстрируя разный уровень включенности. Он ни разу не отвечает на побудительный речевой акт словами. Два курса деятельности соотносятся таким образом, что он говорит только о животных, при этом выполняет действия, связанные с работой. Эффективнее оказывается побуждение, которое сопровождается движениями, требующими ответных, к одному действию (а не к двум сразу); кроме того, важен ход разговора. Мультимодальный анализ выявляет различия в том, как происходит переключение между видами деятельности у аутичных и нейротипичных участников. На переключение к рабочей деятельности можно взглянуть с учетом задач инклюзивного пространства и посчитать успешным, если оно позволяет поддерживать текущую беседу.

Ключевые слова: мультимодальность; мультимодальный анализ; многозадачность; аутизм

1 Введение

Инклюзивные мастерские представляют собой пространство, «где взрослые люди с особенностями развития психики и интеллекта работают и занимаются творчеством на равных с мастерами и волонтерами»[1]. Основная задача – создание условий для «осмысленного труда и общения». Часть сотрудников с ментальными особенностями официально трудоустроена и получает зарплату за работу в мастерских: столярной, швейной, графической, керамической. Присоединиться в качестве волонтера могут желающие, посетив установочную встречу. Кроме того, двери мастерских открыты для посетителей.

Я осуществляла включенное наблюдение, посещая мастерские в качестве волонтера в феврале – марте 2020 г., и не скрывала своего исследовательского интереса, при необходимости, объясняя, что снимаю видео «для науки»[2]. Цель исследования – проанализировать, каким образом осуществляется взаимодействие на рабочем месте, с участием людей с ментальными особенностями. Статья посвящена одному из аспектов изучаемого взаимодействия – переключению между рабочей деятельностью и разговором.

2 Согласование действий между собеседниками

Рабочая активность и общение с коллегами на отвлеченные темы могут рассматриваться с точки зрения понятия *мультиактивности* (multiactivity), т.е. как виды деятельности, которые пересекаются в социальном взаимодействии различными способами [Haddington et al. 2014: 3]. Многозадачность на рабочем месте подразумевает согласование порядка действий между участниками. При этом возможны разные способы согласования и переключения к работе: в частности, с помощью движений телом или побудительных реплик [Kamunen, Haddington 2020].

Тому, как осуществляется побуждение в спонтанном взаимодействии, в особенности просьбам и ответным реакциях на них, посвящено множество работ в рамках прагматики и конверсационного анализа [см., напр., Drew, Couper-Kuhlen 2014, Ogiermann 2015, Rauniomaa, Keisanen 2012, Rossi 2015]. Пристальное изучение видео- и аудиозаписей позволяет расширять список реакций на побуждение, добавляя к стандартным (отказу, согласию, выполнению просьбы) другие, например, делегирование действия третьему лицу и ответное подшучивание [Rudneva 2019]. Отмечается, что на успешность просьбы влияет распределение внимания участников в момент ее совершения [Rauniomaa, Keisanen 2012]. При этом в работах по конверсационному анализу рассматриваются не только побудительные речевые акты, но шире – процесс вовлечения кого-то в деятельность (recruiting) – «использование широкого спектра семиотических средств для достижения цели с помощью другого участника» [Drew, Couper-Kuhlen 2014: 29].

В рамках теории вежливости П. Браун и С. Левинсона, побудительные речевые акты несут угрозу социальным лицам собеседников, которая смягчается при помощи стратегий лингвистической вежливости. *Позитивная* вежливость направлена на демонстрацию солидарности и общего между собеседниками, а *негативная* – на проявление уважения, уменьшение неудобства [Brown, Levinson 1987].

В центре внимания данной статьи находится переключение к рабочей деятельности сотрудника, у которого диагностирован аутизм. В следующем разделе вкратце представлены основные подходы к изучению взаимодействия с аутичными участниками.

3 Основные подходы к изучению коммуникации с аутичными участниками

Особенности коммуникации людей с аутизмом традиционно описываются в терминах *нарушений* – прежде всего, нарушений «модели психического» (theory of mind) [Baron-Cohen, Leslie, Frith 1985] и *прагматических нарушений* [Cummings 2012: 295]. Данный подход распространен в психологии [Медведовская, Лебедева 2011] и в области *клинической прагматики* [Cummings 2012].

[1] Информация с официального сайта организации «Простые вещи»: https://prostieveschi.ru/about-us.

[2] Все сотрудники при поступлении на работу подписывают согласие на видеосъемку. Кроме того, я каждый раз спрашивала у всех присутствующих разрешение на осуществление записи.

Работы *по конверсационному анализу* ставят своей целью описать функции действий, напр., *эхолалии* (автоматических повторений) и формульных выражений, характерных для людей с аутизмом [Local, Wootton 1995; Dobbinson, Perkins, Boucher 2003]. Анализируется взаимодействие в институциональном контексте [Maynard, Turowetz 2020], а также внутри семьи (см. обзор в [Rae, Ramey 2020: 66−69]). К анализу семейного общения применяется, кроме того, *лингвоантропологический* подход [Ochs et al. 2004].

Сложности коммуникации аутичных и неаутичных участников могут рассматриваться как интеракционная проблема [Milton 2012: 884]. В рамках подхода *«обоюдной эмпатии»* Д. Милтона, непонимание – не результат нарушений человека с аутизмом, а общая проблема участников [Milton 2012, Williams 2020].

Настоящее исследование опирается на последние три подхода, а именно в нем 1) применяется методология конверсационного анализа; 2) анализируется взаимодействие, за которым велось длительное включенное наблюдение и в максимальной степени учитывается контекст; 3) понимание рассматривается как достижение всех участников, демонстрируются различия в точках зрения коммуникантов, при этом ни одна из них (ни аутичная, ни нейротипичная) не считается единственной нормой.

4 Методология исследования

Из видеозаписей, сделанных в инклюзивной мастерской (20 часов), были выбраны 20 фрагментов, где происходит побуждение к действию. В статье подробно анализируется один из эпизодов, записанный в керамической мастерской. На рабочем месте завязался разговор о содержании различных животных дома (17 минут), в течение которого инструктор совершает 6 попыток подвигнуть аутичного сотрудника возобновить рабочую деятельность. При помощи мультимодального анализа видеозаписи выясняется: 1) как организовано переключение к рабочей деятельности; 2) чем обусловлено достижение цели побуждения к возобновлению рабочей активности. В фокусе внимания оказываются побудительные действия инструктора, реакции аутичного сотрудника, роль побуждения в общем ходе взаимодействия.

Транскрипция видеозаписи выполнена по системе, разработанной лингвистом Л. Мондадой [Mondada 2018]. Мультимодальное транскрибирование представляет собой этап анализа данных и опирается на релевантность телесных действий для участников. В отличие от кодирования (когда используется конечное число заданных кодов для обозначения разных действий), при создании транскрипции есть возможность характеризовать движения с разной степенью детальности. По системе Л. Мондады, фиксируется соотношение речевых и телесных действий: для движения или жеста указывается точный момент начала и конца относительно речевой цепочки. Для каждого участника устанавливаются свои значки одновременности телесного действия с речевым, которыми обозначаются начало и конец (в речевой цепочке и под ней). Ниже представлены обозначения, используемые в настоящей статье[3]:

Δ знак обозначает начало и конец телесного действия Марии
+ знак обозначает начало и конец телесного действия Василия
÷ знак обозначает начало и конец телесного действия Павла
>> действие началось до начала фрагмента
--> действие продолжается далее до момента, обозначенного аналогичным знаком и знаком участника, напр.: -->Δ
..... подготовка к действию
,,,,, завершение действия

Речь участников выделяется жирным шрифтом, а инициал говорящего обозначается заглавной буквой, в то время как в случае телесных действий – строчной.

[3] В остальном транскрипция в настоящей статье следует принципам конверсационного анализа (список используемых обозначений представлен в приложении). Нумеруются все строки (в отличие от системы Л. Мондады), что удобнее при отсылке к транскрипции. Если неречевое действие следует за речевым, то неречевое действие называется и не обрамляется значками – это тоже удобнее для демонстрации разворачивания взаимодействия.

5 Анализ материала

Фрагмент 1

В керамической мастерской завязался разговор о животных. Его участники: сотрудники с ментальными особенностями (Василий и Павел, у которого диагностирован аутизм), инструктор (называемый в сообществе «мастером») (Мария), волонтеры (Екатерина, Светлана). Отметим, что еще до начала диалога Павел прекратил лепку; он стоял и чистил фартук, беззвучно шевеля губами. Через 5 минут после начала разговора о животных Мария совершает первую, а затем вторую попытку подвигнуть Павла возобновить рабочую активность:

01 **М: Δтак давай, Δ**
02 Δсмотрит на столΔ
03 **Δделаем еще Δ**
04 Δсмотрит на настенные часыΔ
05 **÷делаем еще тарелку÷, ÷еще успеваем ÷**
06 п: ÷смотрит на М.------÷ ÷смотрит на часы÷
07 смотрит на стол (1.0)
08 **М: [так Паш]**
09 **П: [ну и что]тоже**
10 **М: счас при[думаем]**
11 **П: [трудно]**
12 **÷ли держать игуану или чере черепаху?**
13 ÷трогает глину-->
14 (0.3)
15 **М: Δта:к эти все готовыΔ**
16 Δпроверяя формы-----Δ
17 **Е: ну черепаху мне кажется не трудно**
18 (0.2)
19 **÷может [÷ей не очень нравится]**
20 **М: [÷у меня есть предложение]**
21 ÷,,,,-->÷
22 **М: Δдавай сделаем #уборку,**
23 Δпоказывает П. руками, смотря на П-->
#рис.1

24 **М:и отомнем еще одну тарелочку, ты отомнешь**
25 смотрит на П
26 п:смотрит в стол
27 **М: потому что время ÷еще Δ есть а формы ÷уже все Δзалиты,**
28 Δ,,,,,,,,,,,,,,,,,,,,,>Δ
29 п: ÷.....÷собирает крошки-->
30 **М: все почищены и все сделано**
31 **В: черепахи они очень ÷медленные,÷**
32 ÷,,,,,,,-->÷
33 **÷их проще сам- очень легко содержать**
34 п: ÷вытирает стол>>

В анализируемом отрывке мастер по керамике Мария совершает попытки переключить Пашу к рабочей деятельности (1–5, 8–10, 20–33). Она начинает делать это в возникшую паузу, когда, казалось бы, разговор о животных завершен. Сначала инструктор предлагает сделать «еще одну тарелочку», используя форму совместного действия (1 л. мн ч.) с частицей *давай*, при этом обращая внимание на время (сначала смотрит на настенные часы, потом отсылает фразой «еще успеваем») (1, 3, 5). Обращение по имени и употребление формы совместного действия относятся, по Браун и Левинсону, к стратегиям позитивной вежливости, сближающей дистанцию между собеседниками. Далее следует аргументация (27, 30).

Павел, провожая взгляд Марии, также смотрит на часы, но, кроме этого, никак не реагирует на побуждение, продолжая разговор о животных. В данном фрагменте наблюдается некоторая несогласованность действий участников: Мария разговаривает о работе, а Павел совсем не переключается с беседы (8–12). Здесь наблюдается нетипичная ситуация, когда участники не реагируют на фразы друг друга: в частности, ожидается, что за побуждением последует ответное действие. Мастер, со своей стороны, игнорирует реплики о животных, даже вопрос (12), который представляет собой инициирующую реплику пары. Она пытается привлечь внимание сначала обращением (8), а затем телесными действиями (23, рис.1). Во втором случае Мария начинает фразу так, как будто совершает побудительное действие впервые: «У меня есть предложение» (20), и уже предлагает сначала заняться уборкой (22–23), а затем – тарелкой (24). Интересно, что в формулировке второго компонента этого «предложения», как его назвала Мария, содержится самопоправка: «и отомнем еще одну тарелочку, ты отомнешь» (24). После этого Павел начинает собирать крошки со стола.

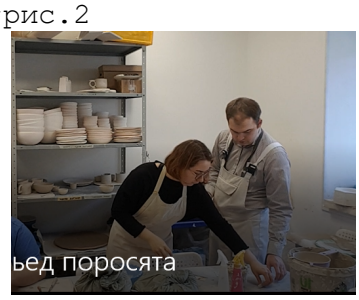
Отметим, что в данном фрагменте происходит побуждение к двум действиям – лепке новой тарелки и уборке рабочего места (сперва именно в таком порядке). Во-первых, эти два действия могут различаться по степени необходимой включенности, в частности для Павла. Во-вторых, уборка отличается еще и тем, что нет четких границ между действиями, которые относятся и не относятся к ней: в частности, если человек собирает мелкие крошки со стола, то это уборка, но если он их кладет обратно, то нет.

В следующем фрагменте Мария 2 раза предлагает заняться уборкой:

Фрагмент 2

```
01  п:>>трогает крошки на стол-->
02  В: +типа кого?              +
03    >>очищает изделие-->
04     +поворачивает голову на М+
05  М: ну вт, любые дикие +животные+
06  в:                    +,,,,,-->+
07  М: [не(???) человека ]
08  В: [+не почему кстати]
09      +повернул голову к М и П -->
10     про енотов говорят то что,
11     [если на  ]
12  П: [ну а что?]÷
13            -->÷
14  м: смотрит на В
15  П: [Δнельзя держать дома        ]
16  В: [Δна природе же, он же восемь]
17                 лет проживает, а если дома, Δ
18  м:  Δсмотрит на стол,чешет голову----------Δ
19  В: [то двадцать лет                ]
20  М: [Паш, давай Δ убираемся?       Δ]
21                 Δпоказывает на столΔ
22  П: смотрит, куда показывает МАР
```

Е: Δ÷но:+ вопрос еще, каких еще÷ Δ
м: Δнаклоняется к столу, #собирает палочки--Δ
#рис.2



П: ÷протягивает руку к палочкам÷
-->+
смотрит на Е, улыбаясь
Е: [Δ*двадцать лет*]
В: [Δи вообще, Δ двадцать лет]
П: [Δи еще Δ,нельзя ли де]ржать дома (.)
м: Δдает палочки ПΔ
п: берет палочки и держит в руках
[÷Δлемура или обезьяну?] ÷
В: [÷ну эти профессионалы ко]торые÷
п: ÷перекладывает палочки в руках÷
м: Δсмотрит на П. -->
В: они те грят двадцать лет
[Δ проживут дома]
М: [Δ я думаю],Δчто ей тоже лучше гораздо Δ в природеΔ
п: Δ.....Δставит палочки в стакан---Δ,,,,,,,,,,Δ
В: [не]
П: [÷а по]росят?
÷смотрит на М-->
в: смотрит на П
М: поросята они ÷домашние÷ особенно карликовые
п: ÷,,,,,-->÷
м: ухмыляется
в: ухмыляется
п: смотрит и улыбается
П: минипигиΔ
-->Δ
М: Δминипиги, да Δ
Δулыбается Δ
Δмотает головойΔ
говорят они правда не очень добрые, нравом
убирает крошки
п: [стирает крошки с сита]
м: [изображает злого поросенка]#рис.3
#рис.3

Е: ÷что вроде уже их на поводке
выгуливают в питере÷ и москве÷
п: ÷стряхивает сито---÷,,,,,,,,,÷

62 м: улыбается
63 С: о:
64 М: Δ#так, паш Δ
65 Δжест рукойΔ
#рис.4

66 **÷давай берем ÷**
67 ÷берет глиняные крошки÷
68 **[вот это вот]**
69 **П: [вообще-то**(???)**]**
70 М: Δ÷#счищаем это все на край Δ÷
71 Δсобирает на столе крошки---Δ
72 п: ÷наблюдает за М----------÷
#рис.5

73 П: вообще-то ÷свиньи поросята Δвсеядные Δ ÷
74 ÷качается, держа в руке кусочек глины÷
75 м: Δотряхивает рукиΔ
76 М: да:, можно все что угодно с кухни им отдавать
77 п: убирает крошки

Здесь участники разбиваются на пары; анализ, представленный ниже, сосредотачивается на одной из них – Павла и Марии. Побудительная реплика снова оформляется как предложение с глагольной формой совместного действия (1 л. мн. ч.) и частицей *давай* (20) (позитивная вежливость, по Браун и Левинсону). Здесь происходит наложение реплик (19–20). После этого Мария наклоняется и собирает рукой палочки-стеки, Павел смотрит в направлении ее взгляда и наклоняется туда же (24–25). Мария, практически выхватывая палочки из рук Павла (рис. 2), собирает их и отдает ему. Павел таким образом возобновляет рабочую активность, но в пассивном формате, реагируя на телесные действия Марии (которые сложнее игнорировать, чем словесные). Параллельно с этими действиями, не требующими полного внимания, Павел продолжает разговор о животных, и формулировка вопроса повторяет структуру, которая использовалась им неоднократно ранее: «нельзя ли держать дома» (30). Затем он перекладывает палочки из рук в руки (35) – это действие находится на грани между рабочей деятельностью и нерабочей: совершаются манипуляции с предметами, но бесцельно, механически. Он ставит их в стакан (40), тем самым завершая этап уборки, который несколько растянулся. После этого, сначала переключившись на поросят/ «минипигов» (42 и далее), Павел самостоятельно переходит к следующему этапу уборки – вытирает крошки с сита (57) и стряхивает его (довольно резкими и громкими движениями) (61).

Следующее побудительное действие Марии (64–71) частично накладывается на начало еще одной реплики Павла о поросятах (68–69). Здесь мастер снова сопровождает инструкцию движениями тела и жестами, за которыми следит Павел (65, 67, 71, рис.5). Только завершив свою фразу (73), Павел приступает к уборке крошек (77). Отметим, что Мария и Павел практически не смотрят друг на друга, но в связи с данной совместной работой, это не является нарушением норм общения.

Таким образом, здесь Павлу и Марии удается сочетать два курса деятельности, которые перемежаются мелкими фазами. Если сравнить этот фрагмент с первым, то разница прослеживается, прежде всего, в телесных действиях. В попытках, которые оказались успешными в том смысле, что за ними следуют необходимые действия, Мария активно использует движения телом, на которые сложнее не реагировать Паше. Кроме того, роль играет общий ход диалога. Так или иначе, получается, что Павел разговаривает только о животных, в то время как действия телом включают и рабочие.
Т.к. Павел продолжает уборку, но не переходит к изготовлению новой тарелки, Мария еще раз совершает побудительное речевое действие:

Фрагмент 3

```
1   п: >>вытирает со стола-->
2   М: так Паш, давай быстренько тарелочку
3        у нас полчаса осталось÷
4   п:                      -->÷
5   м:берет банку с краской
6     уносит банку
7   п:останавливается, ждет
```

В терминах модели Браун и Левинсона, здесь побудительный речевой акт (2) также реализован по стратегии позитивной вежливости, к которой относятся форма совместного действия и диминутивные суффиксы (в данном случае делающие речь менее формальной).
Павел опять никак не реагирует словами на побуждение. Потом он, убравшись, уходит (можно предположить, что выливать воду), а Мария ищет его глазами. Затем, через 16 минут после первой попытки Марии побудить Павла заняться новой тарелкой, он приступает к этому действию:

Фрагмент 4

```
01  м: >>чистит кружку>>
02     смотрит в сторону двери
03  М: Паш у нас же еще не конец занятия?
04     а ты уже поменял воду, мы ее еще запачкать успеем
05  п: проходит сзади М
06     останавливается, смотря в пол
07  в: лепит
08  М: давай тарелочку слепим ÷ быстренько   ÷
09  п:                        ÷поворачивается÷
10  М: ÷у нас еще полчаса есть÷
11  п: ÷берет форму----------÷
12     смотрит на часы
13  в: смотрит на часы
14  В: му меньше полчаса
15  *минут пять у нас есть*
16  п: ставит форму, обтирает
```

В данном фрагменте побудительный речевой акт (8) оформлен аналогично варианту из предыдущего фрагмента (с позитивной вежливостью, по Браун и Левинсону). За побудительной репликой сразу следует действие: Павел поворачивается и берет форму для лепки тарелки. Однако к такому результату привела не одна фраза Марии, а в совокупности все попытки, но главное, общий ход взаимодействия: разговор о животных закончен, уборка завершена, и поэтапно, в своем режиме, Павел уже готовился к лепке.
Таким образом, переключение к рабочей деятельности может стать проблемой для участников. Описанные попытки можно рассматривать как составляющие процесса побуждения, который значительно растянут по времени. Переключение к рабочей деятельности оказывается для Павла многоэтапным процессом.

6 Выводы

Переключение между двумя видами деятельности (разговором и ручным трудом) осуществляется по-разному у разных участников: у нейротипичных – более резко, а у аутичного – поэтапно. Два курса деятельности соотносятся для аутичного сотрудника таким образом, что он говорит только о животных, но выполняет действия, связанные с работой. Ни в одном из случаев он не отвечает на реплики, касающиеся рабочей деятельности, словами. В целом, разговор о животных занимает его в большей степени.

На основании анализа фрагментов можно выделить следующие закономерности относительно эффективности побуждения аутичного участника к действию. Успешнее оказывается побуждение, если оно сопровождается невербальными действиями, требующими ответных: напр., дать палочки. Кроме того, важен общий ход взаимодействия: побуждение результативнее, если происходит в момент, когда разговор на нерабочие темы подошел к завершению или в случае естественной паузы в нем. Анализ, пусть и такого ограниченного материала, с точки зрения модели вежливости, показывает, что реакция аутичного участника на побудительное действие не зависит от языковой формы речевого акта (в том числе от средств вежливости).

Отсутствие какого бы то ни было ответа на побуждение – инициирующие реплики первой пары – даже в тех случаях, когда желаемое действие не выполняется, нарушает связность взаимодействия и отличается от «типичного» сценария (по которому в случае невыполнения действия, формулируется отказ или предлагаются другие варианты). Можно предложить не трактовать отсутствие ответа на побудительный речевой акт как его игнорирование и связать с особенностями переключения между разными видами деятельности.

На успешность переключения к рабочей деятельности можно взглянуть с учетом более широкого контекста и задач инклюзивного пространства, одна из которых – создание условий для непринужденного общения. В этом смысле переключение к рабочей деятельности можно назвать успешным, если оно позволяет поддерживать текущую беседу, не нарушая общего хода взаимодействия и гармонично встраиваясь в него.

Итак, мультимодальный анализ взаимодействия демонстрирует различия в том, как происходит переключение между видами деятельности у аутичных и нейротипичных участников. Подход, при котором принципы нейротипичного поведения не принимаются за единственно возможную норму, вносит коррективы в моделирование человеческой коммуникации. Результаты подобных исследований могут быть использованы для компьютерного моделирования нейроразнообразного взаимодействия.

References

[1] Baron-Cohen Simon, Leslie Alan. M., Frith Uta. Does the autistic child have a “theory of mind”? // Cognition. 21. 1985. — P. 37–46.

[2] Brown Penelope, Levinson Stephen. Politeness: Some Universals in Language Usage. — Cambridge: Cambridge University Press, 1987. 358 p.

[3] Cummings Louise. Pragmatic disorders // Cognitive pragmatics [Handbook of Pragmatics, Vol. 4], ed. by H.-J. Schmid. 2012. — P. 291–315.

[4] Dobbinson Sushie, Perkins Michael R., Boucher Jill. The interactional significance of formulas in autistic language // Clinical Linguistics & Phonetics. 17(4–5). 2003. — P. 299–307.

[5] Drew Paul, Couper-Kuhlen Elizabeth. Requesting – from speech act to recruitment // Requesting in Social Interaction, Studies in Language and Social Interaction / ed. by P. Drew, E. Couper-Kuhlen. Amsterdam / Philadelphia: John Benjamins, 2014. — P. 1–34.

[6] Haddington Pentti, Keisanen Tiina, Mondada Lorenza, Nevile Maurice. Towards multiactivity as a social and interactional phenomenon // Multiactivity in social interaction: Beyond multitasking, ed. by Haddington P., Keisanen T., Mondada L., Nevile M. John Benjamins Publishing Company. 2014. — P. 3–32.

[7] Kamunen Antti, Haddington Pentti. From monitoring to co-monitoring: Projecting and prompting activity transitions at the workplace // Gespächsforschung 21. 2020. — P. 82–122.

[8] Local John, Wootton Anthony J. Interactional and Phonetic Aspects of Immediate Echolalia in Autism: A Case Study // Clinical Linguistics and Phonetics. 9. 1995. — P. 155–194.

[9] Maynard Douglas. W., Turowetz Jason. Sequence and Consequence: Transposing Responsive Actions into Provocations in Forensic and Clinical Encounters Involving Youths with Autism // Atypical Interaction. Ed. by Wilkinson R., Rae J., Rasmussen G. 2020. — P. 39−63.

[10] Medvedovskaja T.A., Lebedeva E.I. (2011) How we understand behaviour of other people, or “theory of mind” [Kak my ponimaem povedenie drugih ljudej, ili «model' psihicheskogo»] // Autism and developmental disorders [Autizm i narushenija razvitija], 2 (33), pp. 22–34.
[11] Milton Damian. On the ontological status of autism: the double empathy problem // Disability and Society. 27 (6). 2012. — P. 883–887.
[12] Mondada Lorenza. Conventions for multimodal transcription. 2018. https://www.lorenzamondada.net/multimodal-transcription.
[13] Ochs Elinor, Kremer-Sadlik Tamar, Sirota Karen G., Solomon Olga. Autism and the social world: an anthropological perspective // Discourse Studies. 6. 2004. — P. 147–183.
[14] Ogiermann Eva. Object requests: Rights and obligations surrounding object possession and object transfer // Journal of Pragmatics. 82. 2015. — P. 1–4.
[15] Rae John P., Ramey Monica. Making and Taking Opportunities for Co-participation in an Interaction Between a Boy with Autism Spectrum Disorder and His Father // Atypical Interaction. Ed. by Wilkinson R., Rae J., Rasmussen G. 2020. — P. 65−92.
[16] Rossi Giovanni. The Request System in Italian Interaction. Ph.D Thesis. Radboud University, Nijmegen, 2015. xvii, 287 p.
[17] Rudneva Ekaterina. How Russians pre-request and seek assistance: a study of interaction in two communities of practice // Russian Linguistics. 43 (2). 2019. — P. 127–142.
[18] Williams Gemma. Talking together at the edge of meaning: Mutual (mis)understanding between autistic and non-autistic speakers. PhD thesis. University of Brighton. 2020.

Приложение А. Список обозначений, используемых в транскрипции

,	Завершенность синтагмы и небольшая пауза.
(0.3)	Пауза более 0,1 секунды с указанием длины паузы.
та:к	Растягивание звука.
ВОду	Увеличение громкости.
?	Интонация вопроса.
М: [так Паш] П: [ну и что]тоже	Одновременное говорение.
минут пять у нас есть	Слова произносятся тише по сравнению с предшествующими.
(???)	Неразборчивая речь.

Detection of Semantic Changes in Russian Nouns with Distributional Models and Grammatical Features

Anastasiia Ryzhova
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Lomonosov Moscow State University
Moscow, Russia
ryzhova@tesyan.ru

Daria Ryzhova
HSE University
Moscow, Russia
dryzhova@hse.ru

Ilya Sochenkov
HSE University
Moscow, Russia
isochenkov@hse.ru

Abstract

The paper presents the models detecting the degree of semantic change in Russian nouns developed by the team *aryzhova* within the RuShiftEval competition of the Dialogue 2021 conference. We base our algorithms mostly on unsupervised distributional models and additionally test a model that uses vectors representing morphological preferences of the words in question. The best results are obtained by the model built on the ELMo architecture with a small window, while the quality of performance of the "grammatical" model is comparable to that of the models based on much more sophisticated algorithms.

Keywords: semantic change, COMPARE metric, distributional models, word2vec, ELMo, RuBERT, grammatical profile

DOI: 10.28995/2075-7182-2021-20-597-606

Оценка степени семантических изменений у русских существительных с помощью дистрибутивных моделей и грамматических профилей

Анастасия Александровна Рыжова
Федеральный исследовательский центр «Информатика и управление» Российской академии наук, МГУ имени М.В. Ломоносова
Москва, Россия
ryzhova@tesyan.ru

Дарья Александровна Рыжова
Национальный исследовательский университет «Высшая школа экономики»
Москва, Россия
dryzhova@hse.ru

Илья Владимирович Соченков
Национальный исследовательский университет «Высшая школа экономики»
Москва, Россия
isochenkov@hse.ru

Аннотация

В статье представлены методы оценки степени семантических изменений русскоязычных существительных, разработанные в рамках соревнования RuShiftEval (Диалог-21) командой *aryzhova*. В качестве основы используются дистрибутивные модели в различных архитектурах (word2vec, ELMo, RuBERT), а также грамматические профили – вектора частотностей морфологических форм, в которых встречаются в корпусах разных временных периодов анализируемые существительные. Лучшие результаты показывает модель на базе архитектуры ELMo, учитывающей также ближайший контекст (окно = 1), а модель на основе одних только грамматических профилей дает результаты, сопоставимые с показателями значительно более сложных алгоритмов.

Ключевые слова: семантический сдвиг, метрика COMPARE, дистрибутивные модели, word2vec, ELMo, RuBERT, грамматический профиль

1 Introduction

The paper presents a series of experiments conducted by the team *aryzhova* within the Dialogue'21 RuShiftEval competition aimed at automatic detection of a degree of semantic change in Russian nouns throughout three time periods: a pre-Soviet (1700-1916), a Soviet (1918-1990), and a post-Soviet (1991-2016) ones [20].

Semantic change is a shift in the meaning of a lexeme occurring due to socio-cultural or purely linguistic reasons. Such changes can be substantial and clearly seen. A common example is a new technical device labeled with an already existing lexical item, cf. a computer *mouse*, or Russian *mama* 'mother' acquiring the meaning 'motherboard'. At the same time, some changes affect just a particular aspect of a word's meaning, such as its general connotations and associations, cf. *Kosovo* which is strongly associated with war thematics after the conflicts in 1998-1999 [21]. Between these extremities, there are plenty of intermediate cases.

The RuShiftEval shared task consisted in ranging a set of Russian nouns according to the degree of semantic change which they underwent in a given time span. We test a set of distributional models (word2vec [22], ELMo [25], RuBERT [16]) with different parameter settings in application to this task.

The aim of our study is twofold. On the one hand, we intend to test the applicability of the most prominent semantic change detection methods that have proven their efficiency in languages other than Russian (mostly in English). Since Russian, in contrast to English, has a rich morphology, we make several attempts to take it into consideration. On the other hand, we expect the results of the experiments to shed new light on the linguistic nature of semantic change. Thus some of our experiments test certain linguistic properties of the given lexical items.

2 Related work

Semantic change is a linguistic phenomenon intriguing scholars from a long time ago (see, e.g. [3]). However, as a specific domain of studies, it shaped relatively recently. The main achievements in the domain are still mostly limited to the detailed descriptions of individual words' trajectories of semantic change (cf. [1]; [14], among others) or the research of individual mechanisms of semantic shift, such as metaphor, metonymy, and others ([15]; [9], to mention just a few).

The emergence of large-scale corpora gave rise to computational studies of semantic change ([17]; [6]) and allowed to formulate and test some general laws underlying these semantic processes. Among these general laws are a correlation between a word's frequency, its level of polysemy, and its aptitude to meaning change [11], or dependency of the degree of semantic change on the level of the word's prototypicality in the previous time period [7]. However, the lion's share of the studies is still based solely on the English data.

In Russian linguistic tradition, there is a fundamental study by V. V. Vinogradov [31] describing an impressive number of words and expressions through the prism of their diachronic change, followed by a recent volume [5] representing an in-depth corpus analysis of twenty words across two centuries. Due to the lack of diachronic corpus data in open access, computational analysis of semantic change in Russian was limited until very recently. Now, the release of the diachronic subcorpora of the RNC opens up new perspectives in the field.

The most prominent methodologies of semantic change detection in English (and some other languages) make use of various distributional models [17]. We use these algorithms as a basis of our experiments as well.

3 Dataset

The organizers of the competition provided three datasets: train, development, and test.

The train set consists of two parts, called RuSemShift1 and RuSemShift2, the first semantic change datasets in Russian [27]. The RuSemShift1 covers pre-Soviet and Soviet times and includes 48 words; the RuSemShift2 covers Soviet and post-Soviet times with 51 words. The datasets contain both nouns and adjectives. However, we exclude the adjectives from consideration since the development and test sets

Dataset	Number of tokens
pre-Soviet	73542513
Soviet	95043479
post-Soviet	83269542

Table 1: The number of tokens in RNC texts.

contain only nouns. The resulting datasets include 44 and 43 words in the RuSemShift1 and RuSemShift2, respectively.

We used recently released diachronic sub-corpora of the Russian National Corpus for our experiments, which correspond to the three time periods. Table 1 presents the volumes of these corpora in tokens.

The semantic change value is measured by COMPARE metric [29]. In brief, for each word, the annotators get random pairs of sentences. The sentences belong to two different periods. The annotators evaluate them, placing scores from 1 to 4, where the lowest score corresponds to the strongest change. The resulting score for each pair of sentences is the average of the scores of the annotators. The score for each word is the average of the scores for each pair of sentences.

The development dataset includes 12 words, the test dataset – 99 words. For each word the participants of the competition have to predict semantic change score (COMPARE) in three pairs of time periods: pre-Soviet - Soviet (RuSemShift1), Soviet - post-Soviet (RuSemShift2), pre-Soviet - post-Soviet (RuSemShift3). The words are ranged according to these scores, and the results are evaluated with the Spearman correlation between the produced ranking and the ranking obtained from human annotation.

4 Experiments

Our main experiments are based on the model architectures suggested in [17]. We use static word embeddings, word2vec [22], and contextualized word embeddings, ELMo [25]. The word2vec model considers only corpus statistics, but it can capture some semantic properties of words, assigning to each word exactly one vector representation. The ELMo model has the BiLSTM architecture and is believed to catch deeper semantic properties. The embedding of each word depends on the given context, so for each word mention the model provides a separate embedding vector. In our research, we consider models trained both on lemmas and on tokens. In addition, we conduct a simplistic experiment testing whether a change in a morphological profile correlates with a change in meaning.

Below we describe each experimental setup in more detail.

4.1 Experiment 1: Word2vec

In the first experiment series, we use the word2vec models provided by the organizers of the competition. These models are trained separately on the Russian diachronic corpora using CBOW algorithm, context window size equals 5, vector size is 300.

As a baseline, we range the words by the cosine similarity of their word2vec representations in the models corresponding to the different time periods. The models were trained on lemmas and aligned with Procrustes alignment (cf. [11]).

On the word2vec models trained on tokens we run three different experiments. We compute final scores in the same way as in the baseline, but we test three types of vector representations:

1. Vector of the word is the average vector of all its word forms.
2. Vector of the word is the vector of its most frequent word form.
3. Vector of the word is the vector of its word form which displays the highest rate of semantic change in the given period. We computed the cosine similarity of word embeddings for each word form, and then took the one with the lowest final score.

4.2 Experiment 2: ELMo

The following experiment series is based on ELMo contextualized embeddings.

For each word we find all the sentences where these words occur and construct 100 random pairs. The first sentence in a pair belongs to one time period, the second to the other period. For example, in the RuSemShift1 task, the first sentence is picked from the pre-Soviet time corpus, the second – from the Soviet one. For each sentence we compute the ELMo embedding of the word for which we want to measure the semantic shift value. We also build additional models where each sentence is presented by the average vector of the target word embedding and the embeddings of the words from its closest context (one, two, or three items away from the target word), which proved efficient in similar experiments from [18]. The final score is the average cosine similarity for all pairs. To achieve a certain level of robustness, we take the average value of the cosine similarity from 5 experiments on different 100 random pairs in each iteration.

Another parameter that we vary in this experiment series is sentence preprocessing: the target and the context words are either treated as tokens or substituted by their lemmas. For the experiment on tokens, we use the tokens ELMo model *ruwikiruscorpora_tokens_elmo_1024_2019*, trained on RNC and Wikipedia, from RusVectores [19]. For the experiment on lemmas, we lemmatize texts with the *morpho_ru_syntagrus_pymorphy* model from DeepPavlov library, which also allows excluding personal names from consideration. It is a neural morphological tagger; the algorithm is described in [13]. The ELMo model *ruwikiruscorpora_lemmas_elmo_1024_2019* is also taken from RusVectores.

4.3 Experiment 3: RuBERT model

This experiment repeats the previous algorithm with another model, RuBERT, presented in [16]. It is a standard BERT model with transformer architecture, trained on the Russian Wikipedia and news, based on the multilingual BERT model. In this case we take into account only the target word embedding.

4.4 Experiment 4: Grammatical features

Previous theoretical studies of various semantic phenomena show that a meaning change can affect not only the distributional properties of a word but its grammatical profile as well (see, for example, [4]). Trying to test this hypothesis on large-scale data, we represent each target noun with a vector of its grammatical preferences. These grammatical vectors consist of 12 dimensions that correspond to different morphological forms appropriate for Russian nouns, i.e., the combinations of six cases and two grammatical numbers. The values of these dimensions are computed as raw counts of tokens of the target word in the given morphological form within each time period. The final scores, which are expected to signal the level of semantic change, are computed as in the previous experiment series – as the cosine similarity of two vectors representing the same word in different time spans.

4.5 Experiment 5: Regression

In the final experiment, we try to benefit from distributional and grammatical properties of words taken together. To achieve this, we train a linear regression with regularization given by the l2-norm (Ridge Regression). The regularization strength alpha is selected with the gridsearch on 5 fold cross-validation and equals to 0.1. We use the cosine similarities from the experiment in Section 4.2 with an ELMo model (tokens + context of window size 1), and the cosine similarities of the corresponding grammatical vectors as features.

5 Results

We conducted all our experiments on the train RuSemShift1 dataset, and then evaluated the best methods on other datasets. Tables 2-6 present the results we obtained. In Tables 5 and 6, *Spearman correlation 1*, *Spearman correlation 2*, and *Spearman correlation 3* state for correlation coefficients corresponding to the three time period pairs: pre-Soviet and Soviet, Soviet and post-Soviet, pre-Soviet and post-Soviet, respectively; an asterisk in the same tables indicates the results where the p-value is lower than 0.05.

Model	Spearman correlation
word2vec similarity	0.485
ELMo, layers='average'	0.508
ELMo, layers='top'	0.526
ELMo with context, 'average', window=2	0.579
ELMo with context, 'top', window=2	0.555

Table 2: Results on the training set RuSemShift1 (44 nouns), lemmatized texts

Model	Spearman correlation
word2vec, average of all word forms	0.282
word2vec, the most frequent word form	0.361
word2vec, the most changed word form	0.28
ELMo, layers = 'average'	0.54
ELMo with context, 'average', window=3	0.593
ELMo with context, 'average', window=2	0.593
ELMo with context, 'average', window=1	0.621
RuBERT	0.411
grammatical vectors	0.30
linear regression	0.602

Table 3: Results on the training set RuSemShift1 (44 nouns), tokens

Tables 2-6 show that the results differ depending on the dataset. However, it is clearly seen that the ELMo model with a small window size outperforms both word2vec and ELMo with a bigger window size models. RuBERT performs worse than ELMo, but it should be taken into account that the ELMo models were trained on the RNC collection, while we do not fine-tune RuBERT on the same corpora.

Interestingly, pretty simple grammatical vectors get rather high scores, even outperforming word2vec and RuBERT models on some datasets. It means that change in meaning is indeed correlated with grammatical re-profiling. Figure 1a, representing the distribution of morphological forms of the noun *svalka* in the pre-Soviet and the Soviet subcorpora, gives an illustration of a clear grammatical shift which can be easily explained from the semantic point of view. The most frequent morphological form of this word in the pre-Soviet period is the nominative singular, while in the Soviet period, the accusative singular form becomes almost as frequent as the nominative singular one. The prevalent meaning of this word in the pre-Soviet subcorpus was that of a fight [24]. Since the word denoted an event, it was frequently used in existential contexts, declaring that a fight was taking place (see Example 1) – hence the preference for the nominative case. In contrary, in the Soviet period this meaning is rarely attested, giving way to the meaning 'dumping ground'. This new semantics triggers the usages of the word *svalka* in the accusative case, because, denoting a specific place, it often plays the role of the goal of a motion (Example 2) typically marked with the accusative case in Russian.

(1) *Totčas že na zemle zakipela* ***svalka.NomSg****, i desyatki tel smešalis' v odnu obš'uyu kričaš'uyu massu.*

'A **scuffle** ensued, with dozens of women in a bawling, struggling mass on the ground.' [Aleksandr Kuprin. Olesya (Stepan Apresyan, 1982)]

(2) *Govoryat šefu: stanok slomalsya. On verit, volokut stanok na* ***svalku.AccSg****.*

'They would tell the boss that a lathe was broken. He would believe them and they would drag the lathe out on to the *rubbish dump*.' [Anatoly Kuznetsov. Babi Yar (David Floyd, 1970)]

Model	Spearman correlation
word2vec on lemmas	0.545
ELMo tokens with context, layers='average', window =1	0.715
RuBERT	0.211
grammatical vectors	0.465
linear regression	0.733

Table 4: Results on the training set RuSemShift2 (43 nouns)

Model	Spearman correlation 1	Spearman correlation 2	Spearman correlation 3
word2vec on lemmas	0.49	0.538	0.622*
ELMo lemmas, 'average'	0.559	0.343	0.664*
ELMo tokens + context, 'average', window =1	0.636*	0.769*	0.818*
RuBERT	0.322	0.531*	0.517*
grammatical vectors	0.392	0.259	0.252
linear regression	0.741*	0.727*	0.832*

Table 5: Results on the development dataset (12 nouns)

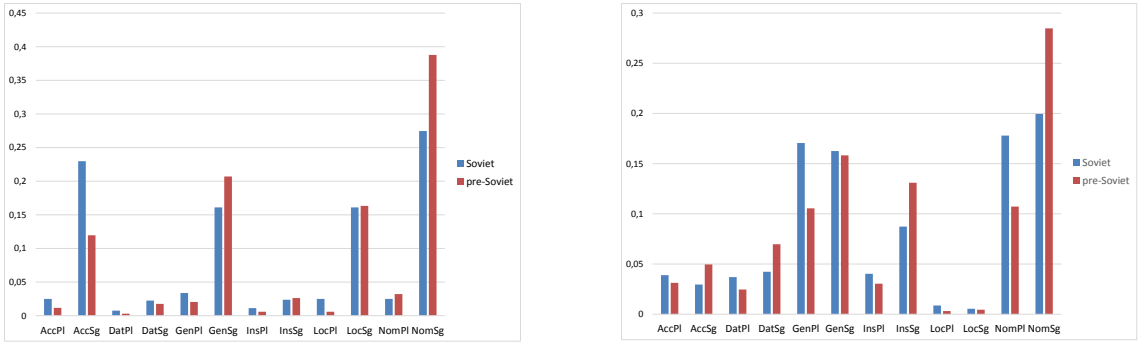
Model	Spearman correlation 1	Spearman correlation 2	Spearman correlation 3
word2vec on lemmas	0.141	0.246*	0.330*
ELMo lemmas, 'average'	0.469*	0.450*	0.453*
ELMo tokens + context, 'average', window =1	0.430*	0.451*	0.469*
RuBERT	0.380*	0.429*	0.448*
grammatical vectors	0.157	0.199*	0.343
linear regression	0.480*	0.487*	0.560*

Table 6: Results on the test dataset (99 nouns)

It is also clear that change in morphological preferences is more crucial in some cases, while in others they seem insignificant. For example, the noun *element* 'element' exhibits substantial semantic differences between the pre-Soviet and the Soviet periods, according to the annotators (its COMPARE equals 1.91), nevertheless it remains grammatically stable.

The opposite situation, where grammatical re-profiling is present, while almost no semantic change is detected (i.e., COMPARE value is higher than 3) is also attested. This is usually the case for rare words, such as *rožděstvo* 'Christmas' or *agenstvo* 'agency'. However, it seems that, for some words, grammatical changes reveal some interesting tendencies remained unnoticed by the annotators. For example, the word *pravitel'*, which does not change in meaning from the pre-Soviet to the Soviet times, according to human annotation (COMPARE equals 3.38), shows a clear shift towards plural forms with almost no change in the case forms ratio. This phenomenon could reflect a cultural (political) change in the corresponding linguistic society, i.e. the shift from monarchy to socialism. However, such effects require further investigation.

As for the regression model, it only slightly outperforms other models, being in general compatible with



(a) Svalka (b) Pravitel'

Figure 1: Changes in grammatical profiles of the words *svalka* and *pravitel'* between the pre-Soviet and the Soviet periods.

the best ELMo model. However, it seems to add some robustness to the algorithm and shows substantial improvement on the datasets with words that have undergone more or less significant grammatical re-profiling.

6 Discussion

Most evidently, the lack of textual data leads to a worse model performance, and this factor is expectedly more crucial for models based on the word2vec CBOW architecture than for those based on the ELMo algorithm. The pre-Soviet subcorpus is the most problematic in this respect: while covering the longest time span, it is the most modest in size, and it is less representative in terms of the text genres. In addition to it, this subcorpus is quite heterogeneous and seems to split into two parts: the texts written app. before the 1830s versus the texts written app. after 1830. This borderline is mentioned in a range of case studies of semantic changes occurred in individual lexical items (cf. [30] on *znatnyj* 'famous', [23] on *mama* 'mother', [26] on *slavnyj* 'famous', among others). It is grounded on observations that it is only after the 1830s that the traces of colloquial speech - the register which is mostly prone to semantic changes - start to penetrate into written texts reflected in the corpus. Thus, many words are used rather differently in these parts of the pre-Soviet subcorpus. This piece of information is lost when a word representation is built on this subcorpus as a whole. It could be better to divide this time period (and, consequently, the subcorpus) into two parts, or even totally exclude the texts written before the 1830s from consideration. However, the amount of textual data representing this period, being already insufficient, would decrease even more.

Another problematic area is the evaluation metrics. The 'golden standard' dataset that was used in the competition is compiled from human judgements on the level of semantic differences within pairs of usages of one and the same lexical item randomly picked from the corresponding text corpora (see [29] and Section 2). This benchmark has many advantages. First, annotators deal with a word in context, which allows them to estimate semantic similarity more accurately than it is done in traditional datasets on semantic similarity, such as WordSim353 [8], where words are given in isolation. Second, the annotation procedure is not very difficult and allows for crowdsourcing. Third, this approach to construction of an evaluation dataset is fully data-driven, it is not based on the previous knowledge acquired from dictionaries or other resources. Finally, because of the random sampling of the context pairs, the dataset roughly represents the respective frequencies of different word senses in the corpus.

At the same time, this benchmark has some drawbacks, the most important of which, to our mind, concerns the vague nature of semantic change. Semantic change is an umbrella term for very differ-

ent processes, including metaphoric and metonymic shifts, grammaticalization and pragmaticalization, specification and generalization of a word meaning, cf. [2], [9]. Heterogeneity of these processes complicates the task of annotating and ranking words according to the level of semantic change they have undergone, since word meanings change in very different ways and aspects. The complexity of the task for humans is clearly seen from the level of inter-annotator agreement, which sometimes does not exceed 0.2.

It might be fruitful to develop a classification of linguistic phenomena related to semantic change and to annotate evaluation datasets according to it, assigning different weights to different classes (e.g., 1 for metonymy, 2 for metaphor, 3 for grammaticalization, etc.). For every lexical item within a pair of time periods, such an annotation would include the list of the attested types of semantic changes together with the total scores counted as the sum of these types weights.

It should be admitted that construction of such a dataset would be pretty costly: it would require a certain level of annotators' linguistic proficiency and a more rigorous analysis of various data sources. However, it could be useful for theoretical studies of semantic change, since it would allow for testing different methodologies for different kinds of processes (cf. different metrics for culturally versus linguistically driven shifts in [10]). The results of such experiments could help to reveal new regularities appropriate for various linguistic phenomena.

For example, it is already well known that a change in meaning usually correlates with a change of the usage context. This assumption underlies the so-called distributional hypothesis [12, 28], which, in its turn, gives rise to distributional semantics, i.e. to the most prominent methodology used to complete the task of semantic change estimation. However, our experiments show that the correlation between the level of semantic change and the level of change of the word's morphological profile is sometimes also rather high. This effect is expected for grammaticalization and pragmaticalization processes, but the examples of such changes were not so numerous in the datasets. It seems that some grammatical effects are appropriate for semantic shifts of other types as well – this topic deserves further investigation.

7 Conclusion

The task of an automated semantic shift detection in Russian is a promising field for future experiments, interesting for both computational and theoretical linguistics. The methods that we had implemented did not receive the highest scores in the RuShiftEval competition. Our model based on the ELMo algorithm with the smallest window took the 6th place among 14 participating systems, while the regression model was not submitted. There is definitely much room for further improvement: additional corpus data, fine-tuning of the RuBERT model, or usage of the train dataset for a supervised model generation could result in better performance.

One of the most interesting theoretical outcomes that we got is a rather strong correlation between grammatical re-profiling and semantic change. We find it an interesting topic for further research which could shed additional light on the nature of semantic shifting.

We would also like to highlight that the competition was devoted to nouns, and it would be an interesting challenge to look for the best ways of semantic change detection in adjectives and verbs. We hope that the next competition within the Dialog conference will include this task.

Acknowledgements

We deeply thank Lidia Pivovarova and Andrey Kutuzov for the organization of the competition and for their constant support of its participants, as well as the anonymous Reviewers for their insightful comments which helped us to improve this text in many respects. We are also grateful to Ekaterina Rakhilina for the idea that grammar matters. Anastasiia Ryzhova and Ilya Sochenkov are supported by RFBR, research project no. 18-29-03187, and partially by the Ministry of Science and Higher Education of the Russian Federation according to the agreement between the Lomonosov Moscow State University and the Foundation of project support of the National Technology Initiative No 7/1251/2019 dated 15.08.2019 within the Research Program "Center of Big Data Storage and Analysis" of the National Technology Initiative Competence Center (project "Text mining tools for big data"). Daria Ryzhova is supported by RFBR, research project no. 20-012-00240.

References

[1] Blank Andreas. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. — De Gruyter Mouton, 2013.

[2] Bloomfield L. Language. — London : Allen & Unwin, 1933.

[3] Bréal Michel. Essai de sémantique (2nd ed.). — Paris, France : Hachette, 1899.

[4] Divjak Dagmar, Gries Stefan Th. Ways of trying in Russian: Clustering behavioral profiles // Corpus Linguistics and Linguistic Theory. — 2006. — Vol. 2, no. 1. — P. 23–60.

[5] Dobrushina Nina R. Daniel Mikhail A. Danova Margarita K. Opachanova Anastasiia S. Pechurina Varvara S. Skorinkin Daniil A. …Sheshenina Aleksandra V. Two centuries in twenty words [Dva veka v dvadcati slovah]. — the National Research University Higher School of Economics, 2016.

[6] Dubossarsky Haim. Semantic change at large: A computational approach for semantic change research : Ph. D. thesis / Haim Dubossarsky ; Ph. D. thesis, Hebrew University of Jerusalem, Edmond and Lily Safra Center for Brain Sciences. — 2018.

[7] Dubossarsky Haim, Weinshall Daphna, Grossman Eitan. Outta control: Laws of semantic change and inherent biases in word representation models // Proceedings of the 2017 conference on empirical methods in natural language processing. — 2017. — P. 1136–1145.

[8] Finkelstein Lev G., Matias Evgeniy et al. Placing Search in Context: The Concept Revisited // ACM Transactions on Information Systems. — 2002. — Vol. 20, no. 1. — P. 116–131.

[9] Geeraerts Dirk. Diachronic prototype semantics: A contribution to historical lexicology. — Oxford University Press, 1997.

[10] Hamilton William L, Leskovec Jure, Jurafsky Dan. Cultural shift or linguistic drift? comparing two computational measures of semantic change // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing / NIH Public Access. — Vol. 2016. — 2016. — P. 2116.

[11] Hamilton William L, Leskovec Jure, Jurafsky Dan. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2016. — P. 1489–1501.

[12] Harris Zellig S. Co-occurrence and transformation in linguistic structure // Language. — 1957. — Vol. 33, no. 3. — P. 283–340.

[13] Heigold Georg, Neumann Guenter, van Genabith Josef. An extensive empirical evaluation of character-based morphological tagging for 14 languages // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. — 2017. — P. 505–513.

[14] Hopper Paul J, Traugott Elizabeth Closs. Grammaticalization. — Cambridge, UK : Cambridge University Press, 2003.

[15] Kövecses Zoltán. Metaphor: A practical introduction. — New York : Oxford University Press, 2010.

[16] Kuratov Yuri, Arkhipov Mikhail. Adaptation of deep bidirectional multilingual transformers for Russian language // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2019. — P. 333–339.

[17] Kutuzov Andrey. Distributional word embeddings in modeling diachronic semantic change : Ph. D. thesis / Andrey Kutuzov ; Ph. D. thesis, University of Oslo. — 2020.

[18] Kutuzov Andrey, Giulianelli Mario. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection // Proceedings of the Fourteenth Workshop on Semantic Evaluation. — 2020. — P. 126–134.

[19] Kutuzov Andrey, Kuzmenko Elizaveta. WebVectors: a toolkit for building web interfaces for vector semantic models // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2016. — P. 155–161.

[20] Kutuzov Andrey, Pivovarova Lidia. RuShiftEval: a shared task on semantic shift detection for Russian // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2021.

[21] Kutuzov Andrei, Velldal Erik, Øvrelid Lilja. Tracing armed conflicts with diachronic word embedding models. — 2017. — 01. — P. 31–36.

[22] Mikolov Tomas, Sutskever Ilya et al. Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems. — 2013. — 10. — Vol. 26.

[23] Opachanova Anastasiia S., Dobrushina Nina R. Mother [Mama] // Two centuries in twenty words [Dva veka v dvadcati slovah]. — the National Research University Higher School of Economics, 2016. — P. 72–93.

[24] Pechurina Varvara S., Dobrushina Nina R. Dumping ground [Svalka] // Two centuries in twenty words [Dva veka v dvadcati slovah]. — the National Research University Higher School of Economics, 2016. — P. 317–338.

[25] Peters Matthew, Neumann Mark et al. Deep Contextualized Word Representations // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). — 2018. — P. 2227–2237.

[26] Rakhilina Ekaterina V., A. Ryzhova Daria. Microhistory of semantic shifts: the case of the Russian adjective slavnyj [Slavnii korabl' - omulevaia bochka. K mikroistorii semanticheskih perehodov] // Proceedings of the Vinogradov Institute of the Russian Language/Trudi instituta russkogo iazika im. V.V. Vinogradova. — 2019. — Vol. 20. — P. 241–256.

[27] Rodina Julia, Kutuzov Andrey. RuSemShift: a dataset of historical lexical semantic change in Russian // Proceedings of the 28th International Conference on Computational Linguistics. — 2020. — P. 1037–1047.

[28] Sahlgren Magnus. The distributional hypothesis // Italian Journal of Disability Studies. — 2008. — Vol. 20. — P. 33–53.

[29] Schlechtweg Dominik, im Walde Sabine Schulte, Eckmann Stefanie. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change // Proceedings of NAACL-HLT. — 2018. — P. 169–174.

[30] Skorinkin Daniil. Famous [Znatnyj] // Two centuries in twenty words [Dva veka v dvadcati slovah]. — the National Research University Higher School of Economics, 2016. — P. 13–38.

[31] Vinogradov Victor V. The history of words: About 1500 words and expressions and more than 5000 words associated with them [Istoriia slov: Ok. 1500 slov i vyrazhenii i bolee 5000 slov, s nimi sviaz. — Russian language [Russkii yazyk], 1999.

RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian

Andrey Sakhovskiy[1,0]
andrey.sakhovskiy@gmail.com

Alexandra Izhevskaya[2,0]
alexandra.izhevskaya@gmail.com

Alena Pestova[2,0]
alpestova1818@gmail.com

Elena Tutubalina[1,2]
elvtutubalina@kpfu.ru

Valentin Malykh[1,3]
valentin.malykh@phystech.edu

Ivan Smurov[4,5]
ivan.smurov@abbyy.com

Ekaterina Artemova[2,3]
elartemova@hse.ru

[1]Kazan Federal University, Kazan, Russia
[2]National Research University Higher School of Economics, Moscow, Russia
[3]Huawei Noah's Ark lab, Moscow, Russia
[4]ABBYY, Moscow, Russia
[5]Moscow Institute of Physics and Technology, Moscow, Russia

Abstract

This report presents the results from the RuSimpleSentEval Shared Task conducted as a part of the Dialogue 2021 evaluation campaign. For the RSSE Shared Task, devoted to sentence simplification in Russian, a new middle-scale dataset is created from scratch. It enumerates more than 3000 sentences sampled from popular Wikipedia pages. Each sentence is aligned with 2.2 simplified modifications, on average. The Shared Task implies sequence-to-sequence approaches: given an input complex sentence, a system should provide with its simplified version. A popular sentence simplification measure, SARI, is used to evaluate the system's performance.

Fourteen teams participated in the Shared Task, submitting almost 350 runs involving different sentence simplification strategies. The Shared Task was conducted in two phases, with the public test phase allowing an unlimited number of submissions and the brief private test phase accepting one submission only. The post-evaluation phase remains open even after the end of private testing. The RSSE Shared Task has achieved its objective by providing a common ground for evaluating state-of-the-art models. We hope that the research community will benefit from the presented evaluation campaign.

https://github.com/dialogue-evaluation/RuSimpleSentEval/.

Keywords: sentence simplification, seq2seq models, cross-lingual models
DOI: 10.28995/2075-7182-2021-20-607-617

RuSimpleSentEval-2021: соревнование по симплификации предложений на русском

Сахновский А.[1,0]
andrey.sakhovskiy@gmail.com

Ижевская А.[2,0]
alexandra.izhevskaya@gmail.com

Пестова А. С.[2,0]
alpestova1818@gmail.com

Тутубалина Е. В.[1,2]
elvtutubalina@kpfu.ru

Малых В. А.[1,3]
valentin.malykh@phystech.edu

Смуров И. М.[4,5]
ivan.smurov@abbyy.com

Артемова Е. Л.[2,3]
elartemova@hse.ru

[0]AS, AI and AP contributed equally

[1]Казанский федеральный университет, Казань, РФ
[2]Национальный исследовательский университет Высшая школа экономики, Москва, РФ
[3]Huawei Noah's Ark lab, Москва, РФ
[4]ABBYY, Москва, РФ
[5]Московский Физико-технический Институт, Москва, РФ

Аннотация

В отчете представлены результаты соревнования RuSimpleSentEval, приуроченного к конференции Диалог 2021. Соревнование RuSimpleSentEval посвящено упрощению предложений на русском языке. Специально для этого соревания авторы подоготовили новый набор данных, насчитывающий 3 тысячи сложных предложений. Каждое предложение снабжено несколькими вариантами упрощения. Среднее число упрощений на сложное представление составляет 2.2 упрощения. Сложные предложения собраны из веб-энциклопедии Википедия, а их упрощения подготовлены работниками краудсорсинговой платформы Яндекс.Толока. Постановка, рассматривая в рамках соревнования, предполагает решение задачи по аналогии с задачей машинного перевода: на вход системе подается сложное предложение, на выходе система выдает упрощенную версию входного предложения. В качестве показателя успешности системы используется широкораспространенная мера SARI, используемая для оценивания систем упрощения предложений.

В соревновании RuSimpleSentEval приняли участие 14 команд. Сумарно было получено 350 вариантов решений, использующих различные стратегии упрощения предложений. Соревнование проводилось в два этапа. На первом этапе – публичном тестировании – участники соревнования могли подавать столько вариантов решений, сколько им нужно. На втором этапе – скрытом тестировании – участники могли отправить только одно, лучшее на их взгляд решение. Данные соревнования, за исключением ответов на скрытом тестовом множестве, опубликованы в открытом доступе. Платформа, на которой проводилось соревнование будет открыта для всех, кто захочет принять участие в проекте уже после завершения соревнования. Авторы считают, что соревнование прошло успешно: подготовлен новый набор данных, уникальный для русского языка, и создана новая площадка для исследования моделей машинного обучения. Авторы надеются, что проведенная работа будет интересна и полезна для исследователей, занимающихся развитием методов машинного обучения и их применению к материалу русского языка.

https://github.com/dialogue-evaluation/RuSimpleSentEval/.

Ключевые слова: симплификация, упрощение предложений, модели последовательностей, межъязычные модели

1 Introduction

The objective of sentence simplification is to transform a source sentence to become easier to read and comprehend. Being able to simplify texts allows better access to information for non-native speakers, people with cognitive disabilities, and children. Although possessing a significant social impact, sentence simplification is not yet widespread in real-life applications due to the lack of parallel corpora, in which a source sentence is matched with its simplified form.

Sentence simplification can be seen as a sequence-to-sequence (seq2seq) problem, which neural language models can efficiently approach. Such a model inputs a source sentence and outputs its simplified version. A large body of research, conducted on seq2seq models evaluation, offers a few performance metrics, of which SARI is usually preferred for sentence simplification.

With this in mind, we organized the shared task on sentence simplification for the Russian language at the Dialogue 2021 conference. There is still no Russian dataset available for this task; we aimed to close this gap and created a general-purpose corpus for simplification in Russian. We adopted a broad definition of the task so that the task itself does not separate lexical simplification, sentence compression, and paraphrasing.

We hope that the corpus and the shared task setup will raise interest for the research and industrial communities, studying NLP for Russian. For example, the techniques developed for seq2seq model training can be adapted to other tasks, such as paraphrasing and question answering.

2 Related work

Recently, there have been multiple achievements in solving sentence simplification problems. A transformer-based model mBART initially developed for machine translation has proven effective for dealing with monolingual tasks of this kind [20] [16]. Adding special control tokens to the model helped

to achieve high quality. While current models rely primarily on the encoder-decoder approach, it is also often accompanied by additional tools. DRESS model [29], for instance, has an encoder-decoder architecture, which is also complemented with deep reinforcement learning to explore possible simplifications and find the best one.

However, success also depends on the quality of the data used for training. The most significant publicly available datasets belong to the English domain. Some of the most prominent ones are PWKP [30] and Wiki-large [29]. The latter is a large-scale parallel English corpus consisting of complex sentences extracted from Wikipedia and their aligned simplified versions.

Nevertheless, focusing mainly on Wikipedia may limit research and lead to inadequacy. Such consideration resulted in the creation of Newsela corpus [22]. It includes news articles edited by professionals, which promises a significant rise in quality. Another corpus of better quality, TurkCorpus [28], was created by asking workers to simplify original sentences on a crowdsourcing platform. One of the latest corpora is ASSET [1], in which each simplification contains several transformations.

There has also been progress in overcoming the lack of simplification corpora for languages other than English [16]. The possibilities of zero-shot learning were investigated to tackle this problem for a low-resource language [17]. In addition, translation data in the form of paraphrases proved to help improve the model's quality.

As for datasets, in Japanese, a 15k sentences corpus was created via a crowdsourcing platform [11]. In Italian, the PaCCSS-IT [23] contains approximately 63k sentences. However, the problem of scarce data resources for sentence simplification in other than English languages remains relevant. Though it is possible to find such corpora in some languages, there is still no Russian dataset available for this task.

3 Dataset

The dataset used for the shared task consists of two parts:

1. The English WikiLarge dataset [29] (EnWikiLarge), translated with the help of a commercial machine translation engine;
2. The RSSE dataset, created specifically for the shared task from scratch.

3.1 Translating English WikiLarge

We utilized theEnWikiLarge for two purposes: first, we used it as is, in English. Secondly, using a commercial machine translation API, we translated WikiLarge to Russian (further, we address this dataset as RuWikiLarge). In total in RuWikiLarge there are 246978 train sentence pairs, 768 dev sentence pairs and 365 test sentence pairs.

Table 1 presents an example of an original-simplified sentence pair in English and its translation into Russian.

3.2 The RSSE dataset

As there are no resources in Russian, which we could mine for simplifications, we collected the dataset via crowd-sourcing. We roughly followed the approach of [22], who showed that crowd workers provide simplifications of good quality and diversity. First, we utilized WikiMedia[1] rankings to pick up the most popular Wikipedia pages in Russian Wikipedia during the last year. Next, from these Wikipedia pages we extracted raw texts, which, in turn, we preprocessed in the following way. We removed lists, references, figure captions, and other parts of Wikipedia articles, that do not belong to the article's body. Next, we splited the remained raw texts into paragraphs and sentences. We sampled first sentences from the paragraphs to avoid undesired coreferent and anaphoric links, which may only complicate the task for crowd workers. Finally, we filtered sentences based on their length in tokens and average IPM (instance per million). To this end, we used the Razdel tool[2] both to split sentences and tokenize them and the frequency dictionary by [14]. Selected sentences, which comprise from 12 to 25 tokens and have an

[1] https://stats.wikimedia.org/

[2] https://github.com/natasha/razdel

Source	Sentence
Original (English)	Before Persephone was released to Hermes , who had been sent to retrieve her , Hades tricked her into eating pomegranate seeds , (six or three according to the telling) which forced her to return to the underworld for a period each year .
Simplified (English)	When Demeter went to the Underworld to rescue her Persephone , Hades forced Persephone to eat the pomegranate . After she ate this fruit it was supposed to keep her in the underworld with Hades so she would be forced to marry him .
Original (Translated)	Перед тем, как Персефона была отпущена Гермесу , который был отправлен за ней, Аид обманом заставил ее съесть семена граната (шесть или три , согласно рассказам) , что вынудило ее возвращаться в подземный мир на период каждый год .
Simplified (Translated)	Когда Деметра отправилась в Подземный мир , чтобы спасти свою Персефону , Аид заставил Персефону съесть гранат . После того , как она съела этот фрукт , предполагалось , что она останется в подземном мире с Аидом , чтобы она была вынуждена выйти за него замуж .

Table 1: A sample from RuWikiLarge dataset

average IPM not lower than 0.95, form the final pool, which we used further to create tasks for crowd workers.

We hired workers on Yandex.Toloka platform to simplify selected sentences. Workers were asked to rewrite a sentence in a simplier way, preserving its meaning, but removing some parts, they might consider unnecessary. Splitting the sentence into two parts and paraphrasing complex terms with simpler or even on colloquial synonyms was considered acceptable. To reject malicious workers, we asked them first to re-write five sentences free of payment and manually inspected such submissions. If we were satisfied with the quality of the trial task, we provided access for the worker to the final pool.

Lastly, before accepting the simplified sentences from the crowd workers, we applied a few more filters. We rejected those sentences, which were exact copy of the original ones or were too similar. We estimated the similarity based on the edit distance in tokens and on the length of the longest common string. The former should have been less than three, while the later should have been less than 90% of the original sentence's length.

In total, we collected more than 3000 sentences. A sample is presented in Table 2. The number of reference simplifications ranges from 3 to 5 with an average of 2.2.

3.3 Dataset statistics

We used an Text Evaluator of Russian texts (TAR) tool[3] [5] to compute descriptive and morphological statistics of original and simplified texts (see Table 3). In particular, text descriptive metrics include one of the most commonly used methods of assessing text readability Flesch-Kincaid Grade Level (FKGL) [8, 6]. It relies on average sentence length (ASL) and word length in syllables (AWL), so short sentences would get good scores even if they are ungrammatical, or do not preserve meaning [27]. Two adaptation of FKGL to Russian are available: Oborneva's formula (O) [21] and SIS formula [25]. We use word frequencies from [15] in TAR tool. Table 3 shows that elaboration of simplified sentences comprised reduction of its length (18.12 words → 12.36 words, 50.78 syllables → 32.98 syllables), which manifests

[3] http://tykau.pythonanywhere.com

Source	Sentence
Original	Климат Казани – умеренно континентальный , сильные морозы и палящая жара редки и не характерны для города .
Simplified 1	В Казани редко бывают и сильные морозы , и жаркая летняя погода .
Simplified 2	В Казани зимой не слишком холодно , а летом не слишком жарко .
Simplified 3	В Казани зимой не очень холодно , а летней жары почти не бывает .

Table 2: A sample from the RSSE dataset

itself in less nouns (7.71 → 5.3), verbs (2.21 → 1.75), and adjectives (3.09 → 1.76). The average reading level estimated by both FKGL measures is decreased, i.e. FKGL (SIS) from 10.68 ± 3.59 to 7.79 ± 3.16 for original and simplified texts, respectively. Although the simplified sentences are shorter, the average number of pronouns is almost the same, while the number of nouns, used in genitive case, decreased. This indicates indirectly that the sentences became less complex.

	Original	**Simplified**
#words	18.12±6.47	12.36±4.7
#syllables	50.78± 18.77	32.98±13.08
ASL (sent. l.)	17.75±6.49	11.65±4.36
AWL (word l.)	2.81±0.47	2.71±0.51
FKGL (SIS)	10.68±3.39	7.79±3.16
FKGL (O)	16.99± 4.85	12.94± 4.58
word frequency	190.47± 112.5	197.85±147.11
#adjectives	3.09±1.9	1.76±1.44
#adverbs	0.77±0.99	0.41±0.71
#pronouns	0.32±1.04	0.38±0.75
#nouns	7.71±2.71	5.3±2.16
#verbs	2.21±1.47	1.75±1.18
avg. #nouns in different cases per sentence		
nominative	1.74±1.26	1.41±1.01
genitive	2.92±1.98	1.78±1.55
dative	0.35±0.68	0.22±0.52
accusative	1.14±1.16	0.85±0.97
instrumental	0.65±0.87	0.38±0.67
prepositional	0.87±0.96	0.63±0.81
avg. #verbs in different tenses per sentence		
present	0.75± 0.95	0.54± 0.78
future	0.02± 0.16	0.02± 0.15
past	1.15± 1.18	0.93± 1.01

Table 3: Statistics of our annotated corpus computed by the TAR tool. All metrics are averaged across sets of original or simplified sentences.

4 Baseline

We utilized mBART [19] which is a multilingual version of previously introduced BART [3]. Russian is well-represented in mBART, i.e. being the second largest language in terms of training corpus size. We trained mBART models in course on two corpora: first, on EnWikiLarge, and then on RuWikiLarge. We used FairSeq [31] and trained our models for 15 and 5 epochs, respectively. For training, we used the learning rate of $3 * 10^{-5}$ and Adam optimizer [12] with warm-up steps at the beginning. Each epoch took about 1 hour on a single machine with 4 NVIDIA P40 GPUs with a per-device batch size of 16.

5 Shared Task set-up

The RSSE shared task was hosted on CodaLab platform[4]. We use EASSE library [7] to compute SARI [22], which was selected as the main performance measure.

The shared task had two phases. During the public testing phase, the participants were provided with RuWikiLarge and the annotated development set, which contained 1000 unique original sentence and 9977 unique sentence pairs in total. The development set could be used for any purpose. The public test dataset consisted of 1000 unique sentences. During the first stage, the participants were allowed to make as many submissions, as needed. There were no restrictions on using any kind of additional data. The participants received immediate feedback from the platform, which returned SARI values for any submission. During the private test set phases, the participants had to test their submissions on the new private test set, that consisted of 1126 sentences. Only one submission was allowed to the platform. The top solutions were determined based on the performance on the private test set. Table 4 presents with the number of unique sentences, utilized at all phases.

Part	Original	Sentence pairs
Dev	1000	3406
Public test	1000	3398
Private test	1126	n/a

Table 4: The number of sentences in the RSSE shared task dataset

6 Results and analysis

6.1 Official results and best models description

We have received submissions from 14 teams for the public test and from 8 teams for the private test. Official shared task results are available in Table 5 (we also provide results on public test for reference in Table 6). All but one team have outperformed the baseline. The winning team `qbic` did not participate in the public testing, so their scores on the public test remain unknown.

All top-placed models used some form of filtering the training dataset or conditioning on control tokens and fine-tuning of large-scale pretrained language models.

The winning solution (`qbic`) is heavily based on Multilingual Unsupervised Sentence Simplification [16]. The model consists of mBART[19] fine-tuned on ParaPhraserPlus[9] and RuWikiSimple conditioned on specific control tokens (Levenshtein similarity, fraction of coinciding characters between original and simplified sentences, word rank, lexeme similarity).

Several other top placed models (second-placed `orzhan`, third-placed `ashatilov` and fifth-placed `alenusch`) are generative (GPT-based models) fine-tuned on the filtered RuWikiSimple.

To be more specific, the second-placed `orzhan` model is ruGPT-3 fine-tuned on the RSSE dev set and filtered RuWikiSimple, where filtering was conducted with the help of 6 different metrics (sentence embedding cosine similarity, named entity preservation score, lexical complexity score, dependency tree depth score, length score and reading ease score). Selecting the best candidate from the variants

[4]The shared task page: `https://competitions.codalab.org/competitions/29037`

User	SARI
qbic	39.6898
orzhan	39.2791
ashatilov	38.491
smpl	38.2379
alenusch	37.82
OnSlaught	36.9367
king_menin	36.6836
komleva.1999	33.1954

Table 5: SARI scores for the private leaderboard of the competition

User	SARI
orzhan	40.2332
alenusch	38.8703
ashatilov	38.8439
bogdansalyp	38.0651
Aroksak	38.0171
smpl	37.9967
phoenix120	37.8921
OnSlaught	37.0807
komleva.1999	37.0175
latticetower	36.0483
memy_pro_kotow	35.8878
letsjusttry	33.6007
cointegrated	31.2018
BASELINE	30.1515
svart	11.5705

Table 6: SARI scores for the public leaderboard of the competition

generated by the model by optimizing a combination of these six metrics instead of SARI allowed for 0.6 SARI improvement on private test.

Third-placed ashatilov solution used GPT-2[13] based model. Both RuWikiSimple filtering and candidate selection was performed with the help of four metrics (cosine similarity, ROUGE-L, and input and candidate length in tokens). However, unlike orzhan model, instead of manually combining the four metrics into aggregate ashatilov selects the best candidate by training a random-forest classifier, where four metrics used as features.

Several other solutions mBART-based enriched with some additional features and techniques. These included pre-training on additional sources of data (with ParaPhraserPlus being the most popular), various handcrafted features, and back-translation[26].

Analyzing the results, one can speculate that the choice between seq2seq pre-training (i. e. mBART-based) and generative models (GPT-based) has a limited impact on the final result. The usage of additional metrics for dataset filtering, candidate selection, and/or as control tokens conversely seem crucial to improve performance further. It appears that the choice of metrics in the top two models is better than the ones used in the third-placed one. On the other hand, training a separate model to select the best candidate (as it is done by the third-placed model) seems to cause less overfitting than using an aggregate metric with fixed parameters (as is evidenced by 0.4 SARI reduction of ashatilov model compared

to 1 SARI of orzhan and alenusch). Additional research has to be conducted in order to validate these claims.

6.2 Lexical richness evaluation

We use the Python package LexicalRichness [24] to compute several measures of textual lexical richness for the original and simplified sentences as well as for the top-3 solutions received from teams qbic, orzhan and ashatilov (see Table 7 for the results). Before the computations, the sentences were lemmatized and are converted to lowercase. All measures were computed on the sentence level and than averaged.Words and terms stand for the average number of words (w) and unique terms (t) in a sentence, correspondingly. The other measures are calculated as follows:

1. type-token ratio (TTR): $\frac{t}{w}$;
2. root TTR (RTTR) [2]: $\frac{t}{\sqrt{(w)}}$;
3. corrected TTR (CTTR) [4]: $\frac{t}{\sqrt{(2w)}}$;
4. Herdan TTR [10]: $\frac{\log(t)}{\log(w)}$;
5. Summer TTR: $\frac{\log(\log(t))}{\log(\log(w))}$;
6. Maas TTR [18]: $(\log(w) - \log(t))/(\log(w)^2)$.

	Original	Simplified	Baseline	qbic	orzhan	ashatilov
words	18.1261	12.7592	14.2247	17.532	9.4654	12.9192
terms	17.0169	12.2299	13.5382	14.4973	9.1918	12.1874
TTR	0.9475	0.9668	0.9614	0.8496	0.976	0.9535
RTTR	3.9746	3.3805	3.5356	3.4748	2.9608	3.3617
CTTR	2.8105	2.3904	2.5	2.457	2.0936	2.3771
Herdan TTR	0.9808	0.9867	0.9847	0.9386	0.989	0.9811
Summer TTR	0.9815	0.9858	0.9834	0.9377	0.9861	0.9795
Maas TTR	0.0066	0.005	0.0058	0.0213	0.0048	0.0072

Table 7: Textual lexical richness measures computed for the original sentences and its simplifications. All metrics are averaged across sets of original or simplified sentences.

The number of words and terms decreased in all simplified sentences compared to the original complex ones. The longest on average sentences are from the team qbic. However, most metrics show that the team qbic's sentences are on average less lexically diverse, due to the repetition of words within one sentence. The team orzhan's simplifications are the shortest ones and the difference between the average number of words and terms is extremely small, which is reflected in the high rates of lexical diversity in the TTR, Herdan and Summer TTR metrics. However, other metrics indicate, on the contrary, less lexical diversity, which is explained by the small number of terms and words in these sentences. The sentences from the team ashatilov are the most similar to human simplifications (Simplified) in terms of lexical diversity and the number of words and terms in the sentence. Our baseline simplifications also have fairly high rates of diversity, but on average longer than human ones, while at the same time longer than the team orzhan's and the team ashatilov's simplifications but shorter then the team qbic's ones.

6.3 Human evaluation

We have run a human evaluation of the top-3 submitted solutions for the shared task. These solutions are named after the participants (the ordering is according private leaderboard): qbic, orzhan, ashatilov. We have compared the output of these solutions with human written simplifications on 125 randomly chosen sentences from the private test set.

For the human evaluation task we have used Yandex.Toloka service, where we asked the crowd workers to choose one of four presented simplifications for a sentence. Three of these four were the parti-

Aggregation	Human	qbic	orzhan	ashatilov	No Preference	Overall
per label	**106**	92	*100*	77	N/A	375
majority	*25*	*25*	*25*	20	**30**	125

Table 8: The aggregation results for human evaluation of the top-3 shared task solutions.

cipants' ones, while the fourth was a human written simplification. The ordering of the simplifications was chosen randomly for each sentence. Each sentence was shown to 3 workers independently.

The human labels were aggregated in two different ways. In the first way we count the human preference labels for all the sentences jointly ("per label" aggregation). For the second aggregation way we have analysed the preferences within one sentence, if some variant had the majority (two or more votes), then we counted this sentence for the specific variant ("majority" aggregation). The results for the aggregation are presented at Tab. 8.

The first aggregation, being more fine-grained, shows an interesting pattern: the human written variants are preferred almost as often as team orzhan's and team qbic's ones. Another essential feature is that team orzhan's solution being the second one by SARI is better than the first place. The second aggregation pictures this comparison from another angle. In 30 sentences out of 125 (i.e. 24%) there is no preference. Thus, none of the presented variants could be considered definitely better than the others. The team qbic's and the orzhan's solutions alongside human-written text shown showed the same result of 25 sentences (20%) of their preeminence each. We could conclude that the top-2 solutions' output and human simplifications have about the same overall quality.

7 Conclusion

We have described the RSSE shared task on sentence simplification in Russian. For this shared task, we have created a new corpus consisting of complex sentences extracted from Wikipedia and aligned with their simplified version. Overall, we received submissions from 14 participants, utilizing a wide range of technologies from ranking models to pre-trained auto-regressive generators. The proposed task does not appear extremely difficult, as the majority of participants have beat the baseline. The received average SARI scores are in line with the expectations and are close to the values established for corpora in other languages. The human evaluation confirms that the best solutions almost reach human-level fluency and diversity.

For more significant impact, we realize the dataset and the code used for computing baseline. The shared task platform remains open for post-evaluation. We hope that the community of NLP practitioners could benefit from the RSSE shared task and its materials. Out future work includes, but is not limited to, developing better measures for text complexity evaluation and tools, which account for lexical and syntactical changes carried out by models.

Acknowledgments

Alena Pestova, Elena Tutubalina and Ekaterina Artemova are supported by the framework of the HSE University Basic Research Program.

References

[1] ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations / Fernando Alva-Manchego, Louis Martin, Antoine Bordes et al. // arXiv preprint arXiv:2005.00481. — 2020.

[2] André J. GUIRAUD (P.).-" Problèmes et méthodes de la statistique linguistique"(Book Review) // Revue de Philologie, de Littérature et d'Histoire Anciennes. — 1962. — Vol. 36. — P. 180.

[3] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension / Mike Lewis, Yinhan Liu, Naman Goyal et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — P. 7871–7880.

[4] Carroll John B. Language And Thought. — Prentice-Hall, 1964.

[5] Computing Descriptive Metrics and Propositions in Reading Texts and Recalls / Mariia Andreeva, Marina Solnyshkina, Valery Solovyev et al. // CEUR Workshop Proceedings. — 2020.

[6] Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel : Rep. : 8-75 / Chief of Naval Technical Training: Naval Air Station Memphis. ; Executor: J.P. Kincaid, R.P. Fishburne, R.L. Rogers, B.S. Chissom : 1975. — February. — 49 p.

[7] EASSE: Easier Automatic Sentence Simplification Evaluation / Fernando Alva-Manchego, Louis Martin, Carolina Scarton, Lucia Specia // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 49–54. — Access mode: https://www.aclweb.org/anthology/D19-3009.

[8] Flesch Rudolph. A new readability yardstick. // Journal of applied psychology. — 1948. — Vol. 32, no. 3. — P. 221.

[9] Gudkov Vadim, Mitrofanova Olga, Filippskikh Elizaveta. Automatically Ranked Russian Paraphrase Corpus for Text Generation // Proceedings of the Fourth Workshop on Neural Generation and Translation. — ACL. — 2020. — P. 54–59.

[10] Herdan G. Quantitative Linguistics // Journal of the Royal Statistical Society. Series A (General). — 1966. — 01. — Vol. 129.

[11] Katsuta Akihiro, Yamamoto Kazuhide. Crowdsourced corpus of sentence simplification with core vocabulary // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018.

[12] Kingma Diederik P, Ba Jimmy. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. — 2014.

[13] Language Models are Unsupervised Multitask Learners / Alec Radford, Jeff Wu, Rewon Child et al. — 2019.

[14] Ljaševskaja Olga N, Šarov Sergej A. Častotnyj slovar sovremennogo russkogo jazyka na materialach Nacionalnogo korpusa russkogo jazyka. — Azbukovnik, 2009.

[15] Lyashevskaya ON, Sharov SA. Chastotnyy slovar'sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka)[The frequency dictionary of the modern Russian language (on the materials of the National Corpus of the Russian language)]. Moscow, Azbukovnik Publ., 2009. 1112 p // dict. ruslang. ru/freq. php. — 2009.

[16] MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases / Louis Martin, Angela Fan, Éric de la Clergerie et al. // arXiv preprint arXiv:2005.00352. — 2021.

[17] Mallinson Jonathan, Sennrich Rico, Lapata Mirella. Zero-Shot Crosslingual Sentence Simplification / Association for Computational Linguistics. — 2020.

[18] Mass Heinz-Dieter. Über den zusammenhang zwischen wortschatzumfang und länge eines textes // Zeitschrift für Literaturwissenschaft und Linguistik. — 1972. — Vol. 2, no. 8. — P. 73.

[19] Multilingual denoising pre-training for neural machine translation / Yinhan Liu, Jiatao Gu, Naman Goyal et al. // Transactions of the Association for Computational Linguistics. — 2020. — Vol. 8. — P. 726–742.

[20] Nishihara Daiki, Kajiwara Tomoyuki, Arase Yuki. Controllable text simplification with lexical constraint loss // Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop. — 2019. — P. 260–266.

[21] Oborneva IV. Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov: dis.... kand. ped. nauk [The Computerized Estimation of Academic Texts Complexity on the Basis of Statistical Parameters. Cand. ped. sci. diss.]. — 2006.

[22] Optimizing statistical machine translation for text simplification / Wei Xu, Courtney Napoles, Ellie Pavlick et al. // Transactions of the Association for Computational Linguistics. — 2016. — Vol. 4. — P. 401–415.

[23] Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification / Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. — 2016. — P. 351–361.

[24] Shen Yan Shun. LexicalRichness Python module. — 2019.

[25] Solovyev Valery, Ivanov Vladimir, Solnyshkina Marina. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics // Journal of intelligent & fuzzy systems. — 2018. — Vol. 34, no. 5. — P. 3049–3058.

[26] Sugiyama Amane, Yoshinaga Naoki. Data augmentation using back-translation for context-aware neural machine translation // Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019). — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 35–44. — Access mode: https://www.aclweb.org/anthology/D19-6504.

[27] Wubben Sander, van den Bosch Antal, Krahmer Emiel. Sentence Simplification by Monolingual Machine Translation // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Jeju Island, Korea : ACL, 2012. — Jul. — P. 1015–1024. — Access mode: https://www.aclweb.org/anthology/P12-1107.

[28] Xu Wei, Callison-Burch Chris, Napoles Courtney. Problems in current text simplification research: New data can help // Transactions of the Association for Computational Linguistics. — 2015. — Vol. 3. — P. 283–297.

[29] Zhang Xingxing, Lapata Mirella. Sentence Simplification with Deep Reinforcement Learning // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — Copenhagen, Denmark : Association for Computational Linguistics, 2017. — P. 595–605. — Access mode: http://aclweb.org/anthology/D17-1063.

[30] Zhu Zhemin, Bernhard Delphine, Gurevych Iryna. A monolingual tree-based translation model for sentence simplification // Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). — 2010. — P. 1353–1361.

[31] fairseq: A Fast, Extensible Toolkit for Sequence Modeling / Myle Ott, Sergey Edunov, Alexei Baevski et al. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). — 2019. — P. 48–53.

Sentence simplification with ruGPT3

Shatilov A. A.
RANEPA
Moscow, Russia
shatilov-aa@ranepa.ru

Rey A. I.
RANEPA
Moscow, Russia
rey-ai@ranepa.ru

Abstract

This paper describes our solution for the RuSimpleSentEval shared task on sentence simplification held together with Dialogue 2021 conference. Our approach was to filter the provided dataset, finetune the pretrained ruGPT3 model on it and select generated simple candidates based on cosine similarity and ROUGE-L with a complex sentence as an input. The system achieved SARI 38.49 and took third place in the competition. We have reviewed and analyzed examples of simplified sentences produced by the model. The analysis showed that the sentences produced by the system lose the original meaning of the input sentence in about half of the cases.

Keywords: sentence simplification, ruGPT3, fine-tuning, text generation
DOI: 10.28995/2075-7182-2021-20-618-625

Упрощение предложений с помощью ruGPT3

Шатилов А. А.
РАНХиГС
Москва, Россия
shatilov-aa@ranepa.ru

Рей А. И.
РАНХиГС
Москва, Россия
rey-ai@ranepa.ru

Аннотация

В данной статье описано наше решение для соревнования по упрощению предложений RuSimpleSentEval, проводящегося в рамках конференции Диалог 2021. Наш подход заключался в фильтрации предоставленного набора данных, дообучении претренированной ruGPT3 модели и отборе сгенерированных с ее помощью примеров простых предложений на основе их косинусной близости и ROUGE-L ко входному сложному предложению. Система получила значение метрики SARI 38.49 и заняла третье место в соревновании. Мы провели обзор и анализ примеров упрощенных предложений, получаемых с помощью модели. Анализ показал, что упрощенные предложения теряют смысл оригинального сложного предложения примерно в половине случаев.

Ключевые слова: упрощение предложений, ruGPT3, дообучение, генерация текста

1 Introduction

Text simplification consists of modifying the content and structure of a text in order to make it easier to read and understand, while preserving its main idea and approximating its original meaning.

In the RuSimpleSentEval task simplification was performed at the sentence level. In this formulation, the goal is to obtain a simplified sentence from a complex one. The criteria for sentence complexity include presence of complex grammatical constructions, subordinate clauses, the presence of rare and ambiguous words, etc.

Our goal was to evaluate how well the finetuned autoregressive ruGPT3 model would handle the task. We approach the problem in three steps. At first, we use only Russian sentences from provided translated WikiLarge corpus and filter it by cosine similarity and ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation for Longest Common Subsequence) [10] metric between complex and simplified sentences. We keep pairs with high cosine similarity and medium ROUGE-L values.

Further, we finetune pretrained ruGPT3 on the filtered dataset similarly to finetuning for paraphrasing [18].

And finally, we use the finetuned model to generate simplified candidates using the complex sentence as a prompt. To choose from several generated candidates a random forest model is used. It is trained to predict SARI (System output Against References and against the Input sentence) [13] metric on the validation dataset on 4 features: cosine similarity, ROUGE-L, complex sentence length in tokens and generated simplified sentence length in tokens. The candidate with the highest predicted SARI is selected as an output.

Our system achieved SARI 38.49 points on the private test set, with a third place finish.

In this paper, we describe our approach in more detail and analyze quality of generated simple sentences.

2 Related work

Most text simplification models treat sentence simplification as a monolingual machine translation task. Phrase-based and syntax-based translation models [21] were succesfully used for this. There are also such approaches as deletion [4] and candidate reranking [19] models.

Lately, the task of sentence simplification has mostly been handled with Seq2Seq models [16], for example, Nisioi et al. [6]. Zhang and Lapata [20] combined Seq2Seq with reinforcement learning to optimize a reward based on simplicity, fluency, and relevance. Martin et al. [3] enhanced the transformer architecture with conditioning parameters such as length, lexical and syntactic complexity. To improve text generation Lagutin et al. [8] used Implicit Unlikelihood Training - a method for regularizing output by finetuning a language model with policy gradient reinforcement learning.

One of the latest works [12] uses an unsupervised approach to automatically create training corpora for simplification in multiple languages from raw Common Crawl web data and train simplification systems in any language with a controllable generation mechanism.

In similar tasks of summarization and headlines generation in Russian, finetuning of BERT-based models (BertSumAbs, mBART) is usually used [1, 11, 2].

3 Task description

Most text simplification models are trained on parallel data: pairs of complex sentence - simple sentence. There was no such dataset for the Russian language, so the organizers of the shared task prepared it for this competition [14].

The training dataset is based on the English Wikipedia and Simple English Wikipedia materials translated into Russian. It was also allowed to use additional data, such as a corpus of paraphrases of news headlines ParaPhraserPlus [7]. This corpus consists of different variants of headlines (from 2 to 15) for one news item.

The validation and test datasets for the shared task are gathered on a crowdsourcing platform. These datasets consist of pairs of one complex sentence and from 1 to 5 simple sentence variants. All data are presented on the competition github.[1] Datasets sizes[2] are presented in Table 1.

Dataset type	Dataset size
Training Wikipedia (all)	248,111
Training ParaPhraserPlus	1,725,393
Validation	1,000
Public test	1,000
Private test	1,126

Table 1: Dataset sizes

[1] https://github.com/dialogue-evaluation/RuSimpleSentEval

[2] The number of different news items is shown for ParaPhraserPlus, the number of pairs of complex input - simple references is shown for other datasets

SARI (System output Against References and against the Input sentence) is used as an automatic quality metric for the task. It is a lexical simplicity metric that measures ”how good” the words added, deleted and kept by a simplification model are. The metric compares the model’s output to multiple simplification references and the original sentence.

A specific implementation of the metric is used from the EASSE library [5] with a modification to account for a variable number of reference sentences. [3]

4 System description

4.1 Data

For our experiments we used only Russian sentences from the provided dataset. It contains 248,111 pairs of sentences and is quite noisy, so additional filtering was applied. To select good examples we used these metrics calculated between complex and simple sentences:

- cosine similarity of embeddings, obtained with BERT large model (uncased) for Sentence Embeddings in Russian language from Sberbank [4]. It shows how similar the sentences are in terms of meaning.
- ROUGE-L F1-score - Longest Common Subsequence (LCS) based statistics. It identifies longest co-occurring in sequence n-grams automatically. It shows how similar the sentences are in terms of common words.

Joint distributions of these metrics for different datasets are shown in Figure 1. For each dataset, histogram of cosine similarities is shown at the top and histogram of ROUGE-L F1 scores is shown on the right. It can be seen, that training dataset distributions differ from validation dataset, because of alignment errors and because sentences in the Simple English Wikipedia are not always simplified versions of sentences from the English Wikipedia.

Filtered dataset was obtained by choosing sentence pairs that have:

- cosine similarity between 0.6 and 0.99
- ROUGE-L between 0.1 and 0.8
- token length of the simple sentence which is less than or equal to the token length of the complex sentence

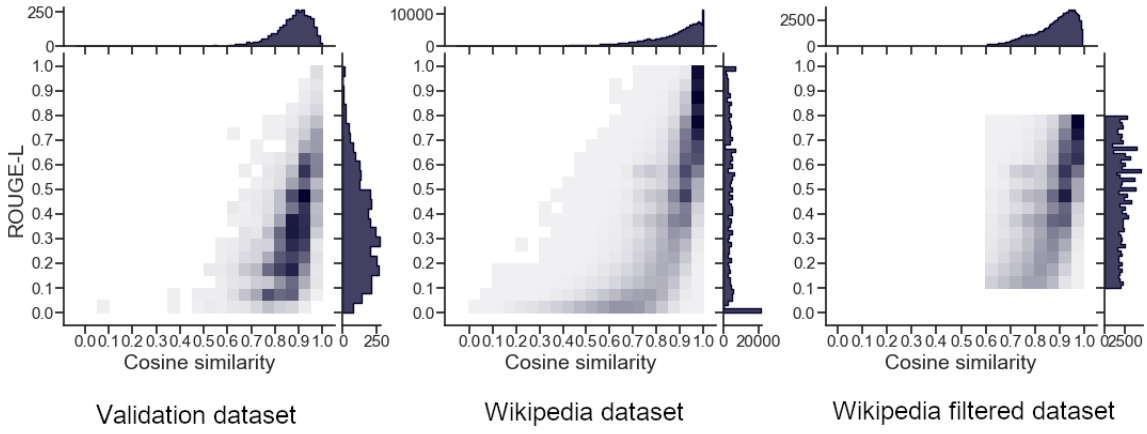


Figure 1: Joint distributions of cosine similarities and ROUGE-L

Final size of the filtered dataset is 120,765 sentence pairs - about half of the original.

[3] https://github.com/Andoree/sent_simplification
[4] https://huggingface.co/sberbank-ai/sbert_large_nlu_ru

4.2 Model

We used a pretrained autoregressive GPT2-like [9] model with 350M parameters from Sberbank [5] called rugpt3medium_based_on_gpt2 - it's the largest model that fit into one 11GB 2080Ti GPU.

Finetuning was done on the prepared examples from the filtered Wikipedia dataset using transformers library [17]. These examples were fed into the model with the addition of special tokens (`<|startoftext|>` - in the beginning, `<|sep|>` - between complex and simple sentences, `<|pad|>` - padding token):

```
<|startoftext|>Complex sentence.<|sep|>Simple sentence.
```

After finetuning it is possible to feed into the model a prepared example as follows:

`<|startoftext|>New complex sentence.<|sep|>` and have the model generate a simplified sentence.

4.3 Selection of generated candidates

Feeding the complex sentence as a prompt into the model can generate several simple sentences. Parameters that were used to generate candidate examples were chosen empirically and presented in Table 2:

Parameter	Value
top_k	50
top_p	0.95
temperature	0.9
max_length	200
length_penalty	0.7
num_return_sequences	5

Table 2: Generation parameters

To select one of the generated candidates as an output we do the following. At first, we generate 5 candidates for each input complex sentence in validation dataset and calculate SARI for each of them. Then, similarly to Wikipedia dataset filtration, we calculate cosine similarity and ROUGE-L between input complex sentence and each of the candidates. Additionally, two more features are calculated: token lengths of input sentence and candidate sentence.

After this we have 4 features:

- cosine similarity between input and candidate
- ROUGE-L between input and candidate
- input sentence token length
- candidate sentence token length

We use the calculated SARI metric as a target and fit a Random Forest model on these features of the provided validation dataset, so that we don't have to manually set tresholds. Setting a small `max_depth` value serves as a regularization. These parameters of the Random Forest model showed the best results: `n_estimators=1000`, `max_depth=5`.

Having trained a Random Forest model on the validation data, we can predict SARI for each generated candidate for the test dataset and choose one with the highest predicted value as an output.

5 Experiments and Results

Code is available on GitHub[6]. rugpt3medium_based_on_gpt2 model was finetuned on one 11GB 2080Ti GPU using transformers library with parameters[7] that are presented in Table 3

[5] https://github.com/sberbank-ai/ru-gpts

[6] https://github.com/InstituteForIndustrialEconomics/DialogueEvaluation21_RuSimpleSentEval

[7] Maximum batch size, that fitted on GPU was 4, for other batch sizes gradient accumulation was used

Parameter	Value
num_train_epochs	3
per_device_train_batch_size	4, 8, 32, 64
learning_rate	5e-5
lr_scheduler_type	linear
warmup_steps	500

Table 3: ruGPT3 finetuning parameters

Random Forest model for generated candidates selection was used from scikit-learn library [15] We used n_estimators=1000 and changed max_depth parameter of the model for the experiments: None, 3, 5, 10.

First, the quality of the models finetuned with the same parameters but on different datasets was compared. Model trained on the filtered Wikipedia dataset showed better results. After that we finetuned model with different batch sizes, and finally evaluated how max_depth parameter of Random Forest candidate selection model influences the result.

Results are presented in Table 4

dataset	Batch size	Max_depth of RF	SARI on the public leaderboard
Original	4	None	37.75
Filtered	4	None	38.30
Filtered	8	None	38.55
Filtered	32	None	38.22
Filtered	64	None	38.34
Filtered	8	3	38.58
Filtered	**8**	**5**	**38.84**
Filtered	8	10	38.65

Table 4: Experiments results

For final submission to private leaderboard the model finetuned on filtered dataset with batch_size=8 was used, max_depth of candidate selection Random Forest model was set to 5. Final SARI score achieved with this system is 38.49. Results of our final model, baseline and the best models are presented in Table 5

Team name	SARI on public	SARI on private	Place on private
Pestova (organizers' baseline)	30.15	-	-
ashatilov (our final model)	38.84	38.49	3
orzhan	40.23	39.28	2
qbic	-	39.69	1

Table 5: Competition results

6 Analysis and Examples

We annotated a sample of 100 simple sentences predicted for the private dataset for the loss of the original meaning of the input complex sentence. The annotation was made in three classes:

- loss of the original meaning - opposite meaning or nonsense
- partial loss of the original meaning - loss of some important detail, but overall meaning is retained
- no loss of the original meaning

Results are presented in the Table 6. As shown in the table, about half of the sentences completely or partially lose the original meaning of input complex sentence.

Result	Number of examples
No loss of meaning	47
Partial loss of meaning	18
Total loss of meaning	35

Table 6: Annotation results of prediction sample

Examples of annotations are presented in Tables 7, 8, 9:

Input:	Взрослых собак, не содержащихся во дворе или открытом вольере, рекомендуется выгуливать минимум два (а желательно три), раза в день.
Output:	Питомников, которые содержат маленьких детей, рекомендуется выпускать на прогулки минимум два или три раза в день.
Input:	Насаждение лучистой сосны в районе Kaingaroa Forest создали крупнейший в мире искусственно выращенный лес.
Output:	В районе Kaingaroa Forest выращивают самую большую в мире светлую сосну.
Input	Земля по отношению к Марсу является внутренней планетой, так же, как Венера для Земли.
Output:	Марс по отношению к Земле - внутренняя планета.

Table 7: Examples with loss of meaning

Input:	После дискуссии заместители получили право задавать вопросы в суде, голосовать на конференциях судей и получать всесторонние консультации.
Output:	Заместители имеют право задавать вопросы в суде и получать подробную информацию из судов.
Input:	Изменение крови характеризуется лейкоцитозом, при тяжёлых формах имеется резкий сдвиг влево с выходом в кровь миелоцитов и юных клеток.
Output:	Изменение крови характеризуется лейкоцитозом и множественными молодыми клетками.

Table 8: Examples with partial loss of meaning

Input:	Вооружённые силы в России не являются самостоятельным политическим субъектом, не принимают участия в борьбе за власть и формировании государственной политики.
Output:	Вооруженные силы России не являются отдельной политической организацией.
Input:	Фэнтези, в форме фольклора, легенд и фантастики, фигурировала в живописи ещё до оформления в отдельный жанр фэнтези-арта.
Output:	Фэнтези фигурировала в искусстве еще до создания отдельного жанра фэнтези-арта.

Table 9: Examples without the loss of meaning

7 Conclusion

In this paper, we evaluated a simple approach for sentence simplification by finetuning a pretrained ruGPT3 model and selecting generated samples based on similarity and difference between input complex and output simplified sentences. The system got SARI 38.49 and placed 3rd on the private leaderboard.

Yet despite high values of the automatic quality metric, sentences produced by the system completely or partially lose the original meaning of the input sentence in about half of the cases. As a result, the system can be used to generate several examples of simplified sentences for further manual selection.

References

[1] Bukhtiyarov Alexey, Gusev Ilya. Advances of Transformer-Based Models for News Headline Generation. — 2020. — 2007.05044.

[2] Chernyshev Daniil, Dobrov Boris. Abstractive Summarization of Russian News Learning on Quality Media // Analysis of Images, Social Networks and Texts / Ed. by Wil M. P. van der Aalst, Vladimir Batagelj, Dmitry I. Ignatov et al. — Cham : Springer International Publishing, 2021. — P. 96–104.

[3] Martin Louis, Sagot Benoît, Éric de la Clergerie, Bordes Antoine. Controllable Sentence Simplification. — 2020. — 1910.02677.

[4] Coster Will, Kauchak David. Learning to Simplify Sentences Using Wikipedia // Proceedings of the Workshop on Monolingual Text-To-Text Generation. — Portland, Oregon : Association for Computational Linguistics, 2011. — Jun. — P. 1–9. — Access mode: https://www.aclweb.org/anthology/W11-1601.

[5] EASSE: Easier Automatic Sentence Simplification Evaluation / Fernando Alva-Manchego, Louis Martin, Carolina Scarton, Lucia Specia // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 49–54. — Access mode: https://www.aclweb.org/anthology/D19-3009.

[6] Exploring Neural Text Simplification Models / Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, Liviu P. Dinu // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Vancouver, Canada : Association for Computational Linguistics, 2017. — Jul. — P. 85–91. — Access mode: https://www.aclweb.org/anthology/P17-2014.

[7] Gudkov Vadim, Mitrofanova Olga, Filippskikh Elizaveta. Automatically Ranked Russian Paraphrase Corpus for Text Generation // Proceedings of the Fourth Workshop on Neural Generation and Translation. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 54–59. — Access mode: https://www.aclweb.org/anthology/2020.ngt-1.6.

[8] Lagutin Evgeny, Gavrilov Daniil, Kalaidin Pavel. Implicit Unlikelihood Training: Improving Neural Text Generation with Reinforcement Learning. — 2021. — 2101.04229.

[9] Language Models are Unsupervised Multitask Learners / Alec Radford, Jeff Wu, Rewon Child et al. — 2019.

[10] Lin Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. — Barcelona, Spain : Association for Computational Linguistics, 2004. — Jul. — P. 74–81. — Access mode: https://www.aclweb.org/anthology/W04-1013.

[11] Malykh Valentin, Porplenko Denis, Tutubalina Elena. Generating Sport Summaries: A Case Study for Russian // Analysis of Images, Social Networks and Texts / Ed. by Wil M. P. van der Aalst, Vladimir Batagelj, Dmitry I. Ignatov et al. — Cham : Springer International Publishing, 2021. — P. 149–161.

[12] Multilingual Unsupervised Sentence Simplification / Louis Martin, Angela Fan, 'Eric de la Clergerie et al. // ArXiv. — 2020. — 05. — Vol. abs/2005.00352.

[13] Optimizing Statistical Machine Translation for Text Simplification / Wei Xu, Courtney Napoles, Ellie Pavlick et al. // Transactions of the Association for Computational Linguistics. — 2016. — Vol. 4. — P. 401–415. — Access mode: https://www.aclweb.org/anthology/Q16-1029.

[14] Sakhovskiy Andrey; Izhevskaya Alexandra; Pestova Alena; Tutubalina Elena; Malykh Valentin; Smurov Ivan; Artemova Ekaterina. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. XX. — 2021. — P. xx–xx.

[15] Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort et al. // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.

[16] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to Sequence Learning with Neural Networks // Advances in Neural Information Processing Systems / Ed. by Z. Ghahramani, M. Welling, C. Cortes et al. — Vol. 27. — Curran Associates, Inc., 2014. — Access mode: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

[17] Transformers: State-of-the-Art Natural Language Processing / Thomas Wolf, Lysandre Debut, Victor Sanh et al. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — Online : Association for Computational Linguistics, 2020. — Oct. — P. 38–45. — Access mode: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[18] Witteveen Sam, Andrews Martin. Paraphrasing with Large Language Models // Proceedings of the 3rd Workshop on Neural Generation and Translation. — Hong Kong : Association for Computational Linguistics, 2019. — Nov. — P. 215–220. — Access mode: https://www.aclweb.org/anthology/D19-5623.

[19] Wubben Sander, van den Bosch Antal, Krahmer Emiel. Sentence Simplification by Monolingual Machine Translation // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Jeju Island, Korea : Association for Computational Linguistics, 2012. — Jul. — P. 1015–1024. — Access mode: https://www.aclweb.org/anthology/P12-1107.

[20] Zhang Xingxing, Lapata Mirella. Sentence Simplification with Deep Reinforcement Learning // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — Copenhagen, Denmark : Association for Computational Linguistics, 2017. — Sep. — P. 584–594. — Access mode: https://www.aclweb.org/anthology/D17-1062.

[21] Zhu Zhemin, Bernhard Delphine, Gurevych Iryna. A Monolingual Tree-based Translation Model for Sentence Simplification // Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). — Beijing, China : Coling 2010 Organizing Committee, 2010. — Aug. — P. 1353–1361. — Access mode: https://www.aclweb.org/anthology/C10-1152.

The Russian particle *zhe* in the light of parallel corpora

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

Abstract

This paper deals with the Russian particle *zhe* and its use in the Russian translations from English and demonstrates the possibilities of "one-focus analysis" in contrastive studies based on the parallel corpora. It correlates the explications of *zhe* given in earlier studies (it makes special reference to the *Active Dictionary of Russian*) with the stimuli to translation, that is, fragments of the original English text that might cause the appearance of *zhe* in a Russian translation as a reaction to those stimuli. The study sought to validate, disprove or improve the semantic analysis of *zhe* made without recourse to electronic corpora.

The analysis of the stimuli that have led Russian translators to use the particle *zhe* reveals important characteristics of this word. It turns out that the Russian particle *zhe* is often pragmatically obligatory as its absence would violate the idiomatic nature of the utterance and change its illocutionary force. It is often the case that if a translator had given word-for-word translation, that is without a particle, they would convey the precise meaning, but the translation would be inadequate: the wrong implicature would appear. On the other hand, when they add the particle, they may impart new shades of meaning which the original text did not contain.

Keywords: semantics, language-specific particles, translation, parallel corpora, lexicography
DOI: 10.28995/2075-7182-2021-20-626-635

Русская частица *же* в зеркале параллельных корпусов

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

В статье рассматривается семантика русской частицы *же* на основе ее функционирования в англо-русском параллельном подкорпусе «Национального корпуса русского языка» (НКРЯ). Делается попытка подтвердить, опровергнуть или уточнить существующие описания, в частности в «Активном словаре русского языка» (Левонтина 2017; далее – АС).

Очевидно, что в английских оригиналах нет и не может быть единицы, для которой *же* было бы непосредственным переводным соответствием. Установив, каковы были причины, побудившие переводчика использовать эту частицу в русском тексте, мы имеем шанс проникнуть в ее семантические секреты, до сих пор ускользавшие от внимания или описанные неточно.

Итак, в центре внимания находится функционирование лингвоспецифичной частицы *же* в переводных текстах (в переводах с английского языка). Иными словами, исследование выполнено в русле «монофокусного» анализа [1]. Важно, что, в отличие от контрастивных корпусных

This paper has been written with the support of RFBR (project # 19-012-00505).

[1] В более ранних публикациях мы называли такой анализ «унидирекциональным» (Шмелев, Зализняк 2017: 370).

исследований, при таком анализе в центре внимания оказывается именно язык переводных текстов (в нашем случае русский), тогда как оригинальные тексты (в нашем случае – английские) используются как средство для выявления «семантического задания», стоящего перед переводчиком. Если сопоставительные исследования частиц на базе параллельных корпусов имеют уже относительно давнюю традицию и возможности использования параллельных корпусов для таких исследований демонстрировались в том числе в рамках конференции «Диалог»[2], то «монофокусные» исследования дискурсивных слов на базе параллельных корпусов стали осуществляться относительно недавно[3].

Вообще говоря, частица *же* неоднократно была предметом внимания со стороны лингвистов. Ее описания делались еще в докорпусную эпоху; самыми важными представляются наблюдения, сделанные в статье (Padučeva 1987), недавно напечатанной по-русски с некоторыми сокращениями и одновременно дополнениями (Падучева 2018: 333–352). В ней различные употребления *же* классифицируются с учетом ее положения в высказывании. *Же* в позиции, которую Е. В. Падучева называет «свободной», бывает двух типов: *же* тождества (*За столом сидела та же женщина*) и *же* незамедлительности (*Он пришел в первый же вечер*). В позиции Ваккернагеля (после первого ударного слова в клаузе), которую Е. В. Падучева называет «начальной», *же* может относиться к одному из четырех типов: *же* противительное (*Моею будет век Людмила, Руслан же гробу обречен*), *же* обоснования (*Он же гений*), *же* следствия (*Здорово же ты увлекся, если ничего не слышал*; *Что же мне в таком случае делать?*; *Так поди же попляши*) и *же* присоединительное, иллюстрируемое примером В. З. Санникова *Коля добр, доброта же зачастую беззащитна*[4]. Выделенные типы употребления слова *же* различаются своими просодическими характеристиками.

После появления электронных корпусов, в первую очередь НКРЯ, оказалось возможно использовать их данные. В частности, именно на основе этих данных было выполнено лексикографическое описание частицы *же* (которая после гласных в неформальной речи может выступать в виде *ж*) в АС. В соответствующей словарной статье вслед за традиционными описаниями разграничиваются употребления *же* в качестве частицы и в качестве союза. При этом отмечается, что употребления *же* в качестве союза не характерны для живой современной речи: первое из них (*же 4*) характеризуется как необиходное, а второе (*же 5*) – как книжное. Довольно редкое *же 5* толкуется в АС как ‘Говорящий во втором предложении поясняет то, о чем говорилось в первом предложении’; оно соответствует значению, которое Е. В. Падучева обозначила как *же* противительное. Заметим, что *же 4* соответствует *же* противительному по Е. В. Падучевой (примеры на *же 4* в АС – *Они уехали, мы же остались*; *Муж целыми днями работал, жена же ходила по магазинам*; *Раньше это слово было очень употребительно, в современном же языке встречается редко*); оно встречается в русских текстах несколько чаще, чем несколько искусственное для современного языка *же 5*, но тоже может считаться относительно редким.

Гораздо более характерны для современной речи употребления, рассматриваемые в АС как примеры не союза, а частицы *же* (именно они составляют подавляющее большинство употреблений *же* в русских переводах с английского языка, вошедших в параллельный подкорпус НКРЯ). Приведем синопсис этих употреблений по АС:

же 1 ‘ведь’: *Я же просил!*

же 2 усилительная: *Как же это могло случиться?*

же 3.1 отождествительная: *Адрес тот же?*

же 3.2 ‘с минимальным промежутком’: *В первый же день пребывания*

Помимо это в АС упоминается ряд идиом с частицей *же*.

Прежде чем рассматривать разные типы употребления частицы *же* в переводах, полезно обратиться к понятию прагматической обязательности, введенному в статье (Levontina, Shmelev 2005), тем более что, хотя статья в основном была посвящена русской частице *еще* (в одном из типов употреблений), в ней кратко рассматривалась и частица *же*. В этой статье упоминались некоторые контексты, в которых частица *же* оказывается прагматически обязательной в том смысле, что ее устранение делает высказывание неидиоматичным или меняет его иллокутивную

[2] Ср., напр., публикации некоторых из докладов (Кобозева, Орлова 2008; Добровольский, Левонтина 2012; 2014; 2015; 2017; Добровольский, Зализняк 2018).

[3] См., в частности (Шмелев 2015; Шмелев, Зализняк 2017).

[4] В книге (Санников 2008: 278) используется выражение «*же* присоединения или обобщения».

функцию. Так, фраза *Ну что же ты!* очень естественно понимается как упрек, в то время как фраза *Ну что ты!* может пониматься как выражение несогласия, или как поторапливание, или, возможно, как-то иначе, но не как упрек. Вопрос *Как так?* может пониматься только как выражение недоумения, в отличие от вопроса *Как же так?*, у которого гораздо больше возможных интерпретаций.

Там же отмечалась некоторая парадоксальность употребления лингвоспецифичных прагматически обязательных частиц в переводных текстах. Если в том или ином языке некоторая частица прагматически обязательна в данном типе контекстов и человек переводит подобный контекст на данный язык с языка, в котором аналогичного показателя нет, то он оказывается перед выбором. Если он переведет фразу буквально, без частицы, то, точно передав смысл, он тем не менее получит неадекватный перевод, содержащий ложную импликатуру. Если же он добавит во фразу частицу, а вместе с ней и смысл, которого, строго говоря, не было в исходном тексте, то, несмотря на добавление нового смысла, перевод будет более адекватным.

Это иллюстрировалось в том числе на примере частицы *же*, однако в качестве материала был использован всего один текст – *Winnie-the-Pooh* Алана Милна в переводе на русский язык (точнее, пересказе) Бориса Заходера (*Винни Пух и все-все-все*). Отмечалось, что частица *же* часто появляется в заходеровском переводе, когда в оригинале ей, казалось бы, ничего не соответствует. Впрочем, там же отмечалось, что *же* в переводе Заходера во многих случаях возникает тогда, когда в английском тексте в соответствующей фразе некоторые слова выделены курсивом (можно добавить – иногда посредством капитализации). Иными словами, частица *же* соответствует эмфатическому выделению в оригинале. По-видимому, именно использование *же* в качестве показателя эмфатического выделения дает основание говорить о значении «усиления», которое выделяется едва ли не во всех толковых словарях русского языка.

Впрочем, случаи появления частицы *же* (или любого другого дискурсивного слова) в переводе можно отнести к одному из двух типов: когда частица действительно оказывается прагматически обязательной (так что ее отсутствие привело бы либо к аномалии, либо к ложной импликатуре) и когда ее появление в тексте – сознательный или бессознательный выбор переводчика, хотя можно было бы обойтись и без нее (Zalizniak, Shmelev 2017). Вообще говоря, оба случая заслуживают внимания: семантика дискурсивного слова проясняется не только в тех контекстах, когда оно прагматически обязательно, но и в тех, когда, казалось бы, без него можно обойтись, но тем не менее носитель языка (в рассматриваемых случаях – переводчик) предпочитает его использовать.

Более того, случаи «прагматической обязательности» частицы тоже неоднородны. К ним относятся как случаи, когда действительно нельзя обойтись без данной частицы (иначе возникают нежелательные прагматические сдвиги), так и случаи, когда некоторый дискурсивный показатель необходим, но при этом возможен выбор между разными показателями: напр., при повторной просьбе или уговаривании могут использоваться разные частицы, но почти невозможно или прагматически неадекватно повторение просьбы без какого бы то ни было показателя: если на просьбу выключить телевизор адресат речи никак не реагирует, то в зависимости от ряда прагматически значимых факторов говорящий может повторить просьбу с разной степенью категоричности, сказав нечто вроде *Ну выключи телевизор*; *Да выключи телевизор*; *Выключи же телевизор*; но едва ли уместно будет просто сказать *Выключи, пожалуйста, телевизор*, не маркировав просьбу как повторную (Левонтина 1999).

Принимая во внимание все указанные разграничения, можно обратиться к случаям использования частицы *же* в русских переводах с английского языка. Сразу следует сказать, что ввиду несбалансированности англо-русского параллельного подкорпуса НКРЯ статистический анализ на нем не имеет особого смысла. Необходимо содержательное рассмотрение каждого отдельного случая. Ограниченный объем статьи не позволяет останавливаться на деталях – изложены будут лишь результаты анализа в самом общем виде.

Следует сразу сказать, что, по-видимому, сочетания *т*-местоимений с *же* можно рассматривать не как свободные сочетания слов, а как единые идиоматичные единицы, в которых *же* представляет собою нечто вроде постфикса, подобного таким постфиксам, как *-то* и *-нибудь*, присоединяемым к *к*-местоимениям. Эти местоименные единицы устойчиво используются для перевода английского *same* или выражений, содержащих *same*. К местоимениям *тот же*, *этот же* и

такой же часто присоединяется слово *самый*, а на базе местоимения *тот же* образуется фразеологический комплекс *один и тот же*.

Обратим внимание на различие выражений *тот же (самый)* и *один и тот же* (в английском оригинале им часто соответствует одно и то же выражение, включающее слово *same*). Местоимение *тот же* предполагает отождествление объекта A1 с некоторым A2, фигурирующим в ситуации S и предположительно известным адресату речи; при этом A2 или S может уже быть в поле внимания адресата речи, напр. упоминаться в предшествующем тексте, или же вводиться в подчиненной клаузе (*A1 то же, что и A2*; *A1 тот же, что и в S*). Напротив того, выражение *один и тот же* предполагает, что сравниваются A1 и A2 (а возможно, и A3, A4 и т. д.), коммуникативный статус которых в общем одинаков. Ср. примеры, в которых *тот же (самый)* и *один и тот же* очевидным образом не взаимозаменимы:

(1) *The answer is the same as for the genetic replicator. – Ответ будет тот же, что и для генетического репликатора. [Ричард Докинз. Расширенный фенотип: длинная рука гена (А. Гопко, 2010)]*

(2) *Just from me repeating the same words enough times, and pointing to objects, he had learned the Russian words for those objects. – Просто из-за того, что я повторял одни и те же слова достаточно времени и указывал на объекты, он узнал русские обозначения этих объектов. [Почему я учил своего сына говорить по-русски? (Inosmi.ru, 2018)]*

(3) *Thirty years ago Richard Collins wrote that pilots, by observing the uniform precision of autoland touchdowns, would improve their performance. Now they are saying the same thing about HUDs… – Тридцать лет назад Ричард Коллинз писал, что пилоты, наблюдая за однообразной точностью автоматических посадок, улучшат свои показатели. Теперь то же самое говорят об индикаторах на лобовом стекле… [Дэвид Минделл. Восстание машин отменяется! Мифы о роботизации. (В. Краснянская, 2017)]*

(4) *When everyone is claiming the same thing, it's almost impossible to make your product stand out. – Когда каждый твердит одно и то же, практически невозможно выделить ваш продукт из общей массы. [Народу виднее (ng.ru) (2015)]*

Используя данные параллельного англо-русского подкорпуса НКРЯ, попытаемся уточнить толкование значения *же 3.1*, сформулированное в АС. Приведем это толкование полностью: «*A1 же* ‘A1 совпадает или почти совпадает с тем, о чем шла речь раньше’ [часто после слов *тот*, *этот*, *такой*, *сей*, *столько*, *там*, *здесь*, *тут*, *туда*, *оттуда*, *тогда*, *так*, *потому*, *поэтому*, *затем*; *подобный*, *сходный*, *сам*]». Сразу можно сказать, что значение частицы *же* после *т*-местоимений (*тот*, *такой*, *столько*, *там*, *туда*, *оттуда*, *тогда*, *так*, *потому*, *затем*) не отвечает данному толкованию. По существу, она представляет собою постфикс, в соединении с которым местоимение выражает тождество, но не обязательно «с тем, о чем шла речь раньше»: очень часто это тождество с тем, о чем идет речь в последующей подчиненной клаузе:

(5) *South Ossetia is the same as Kosovo. – Южная Осетия — то же самое, что Косово. [Десять главных мифов о России, ее лидере и ее силе (inosmi.ru) (2008)]*

(6) *Elephants survived also, but the life of an elephant today is largely the same as it was millions of years ago. – Слоны, конечно, тоже выжили, но их жизнь такая же, как и миллионы лет назад. [Саймон Синек. Лидеры едят последними. Как создать команду мечты (Е. И. Животикова, 2015)]*

(7) *Indeed the two programs could swap physical computers every other game, each one running alternately in an IBM and an ICL computer, and the result at the end of the tournament will be the same as if one program consistently ran in the IBM and the other consistently ran in the ICL. – Программы вообще могут обмениваться носителями перед каждой новой партией — например, выполняясь по очереди в компьютерах IBM и ICL, — но результат в*

конце соревнования будет таким же, как если бы одна программа постоянно запускалась на компьютере IBM, а другая — на ICL. [Ричард Докинз. Расширенный фенотип: длинная рука гена (А. Гопко, 2010)]

При этом, когда тождество устанавливается с тем, о чем идет речь в последующей подчиненной клаузе, местоимение с *же* часто может быть заменено соответствующим *т*-местоимением без *же*. А когда тождество устанавливается с тем, о чем шла речь ранее, замена обычно невозможна, и это можно считать еще одним аргументом в пользу того, что *т*-местоимение с *же* представляет собою единый комплекс, который разумно считать особой единицей, «отождествительным» местоимением. Ср.:

(8) *Greece vehemently opposes its northern neighbor's use of "Macedonia" without a qualifier, because a region in Greece bears the same name. – Греция активно возражает против использования её северным соседом в своём названии слова «Македония» без какого-либо дополнительного определения, потому что в Греции есть регион, который называется точно так же. [Свежий импульс на Балканах (Inosmi.ru, 2018)]*

(9) *A few boys were making their way to the cricket-field; and two or three shopkeepers who were standing at their doors looked as if they should like to be making their way to the same spot… – Несколько молодых людей направлялись к крикетному полю, а два-три лавочника, стоя у дверей своих лавок, имели такой вид, словно им хотелось отправиться туда же… [Чарльз Диккенс. Посмертные записки Пиквикского клуба (А. В. Кривцова, Е. Л. Ланн, 1933)]*

Однако, как мы видели, в АС в ряду слов, после которых частица *же* часто используется в значении отождествления, упоминаются не только *т*-местоимения. Действительно, употребление *же* в значении отождествления не после *т*-местоимений также встречается, и, хотя частица *же* в них обычно оказывается факультативной (*Задал вопрос, и сам же на него ответил* ≈ *Задал вопрос, и сам на него ответил*), она тем не менее нередко появляется в переводных текстах, напр.:

(10) *And here he wrote 'a poor physician'. And it was he, without doubt, who scratched a calendar on this stone. – А вот тут он приписал: «несчастный доктор». И он же, конечно, нацарапал и этот календарь, вот здесь, на камне. [Чарльз Диккенс. Повесть о двух городах (М. П. Богословская, С. Я. Бобров, 1950-1960)]*

(11) A similar distancing happened with Einstein's wife, Mileva. – Подобное же отдаление произошло и между Эйнштейном и его женой, Милевой. [Дэвид Боданис. E=mc2. Биография самого знаменитого уравнения в мире (С. Б. Ильин, 2009)]

(12) *Here haunted of yore the fabulous Dragon of Wantley; here were fought many of the most desperate battles during the Civil Wars of the Roses; and here also flourished in ancient times those bands of gallant outlaws, whose deeds have been rendered so popular in English song. – По преданию, здесь некогда обитал сказочный уонтлейский дракон; здесь происходили ожесточенные битвы во время междоусобных войн Белой и Алой Розы; и здесь же в старину собирались ватаги тех отважных разбойников, подвиги и деяния которых прославлены в народных песнях. [Вальтер Скотт. Айвенго (Е. Бекетова, 1890-1902)]*

(13) *He gazed once again into the void of night, feeling dwarfed by the events he had put into motion. – И чтобы успокоиться, он снова выглянул во тьму ночи, чувствуя себя игрушкой в водовороте событий, которым сам же положил начало. [Дэн Браун. Код Да Винчи (Н. Рейн, 2004)]*

Именно такое употребление адекватно описывается толкованием, сформулированным в АС.

Впрочем, можно добавить, что стоило бы устранить из списка слов, после которых часто употребляется частица *же* в значении отождествления, слово *тут*. В принципе выражение *тут же*

может означать ‘приблизительно в этом же месте’ и примеры такого рода даже можно обнаружить в англо-русском подкорпусе НКРЯ. Ср.:

(14) *A young couple close by flirted a fan by turns, making an unpleasant draught. Francie and one of her lovers stood near. – Одна пара неподалеку от него кокетливо обмахивалась по очереди веером, и Сомсу был неприятен ветер, который они подняли. Тут же рядом остановилась Фрэнси с кем-то из своих поклонников. [Джон Голсуорси. Собственник (Н. Волжина, 1946)]*

Но несравненно чаще *тут же* используется в значении ‘в этот момент, немедленно’. Употребление *же* в этом случае непосредственно связано со значением, которое в АС обозначено как *же 3.2*; однако, поскольку рассматриваемое значение не складывается из значений *тут* и *же 3.2* и *же* нельзя устранить без коренного изменения смысла (в отличие от сочетаний, в которых фигурирует *же 3.2*, напр. *сразу же*, *сейчас же*, *завтра же*) естественно считать *тут же* идиомой; было бы целесообразно включить это выражение в список идиом с частицей *же*.

Это побуждает нас еще раз обратиться к упомянутому выше списку идиом из словарной статьи *же* в АС. Полностью он выглядит следующим образом: «***всё же*** *см*. **ВЕСЬ**; ***одно и то же*** *см*. **ОДИ́Н**; *к тому же см*. **ТО**, *надо же см*. **НА́ДО**, *или же см*. **И́ЛИ**, *туда же см*. **ТУДА́**, *Как же! см*. **КАК**». Сразу же возникают вопросы, уже не связанные с использованием частицы *же* в переводах с английского языка. Так идиома *всё же* содержит отсылку к словарной статье слова *весь* (автор – Е. В. Урысон), но в ней идиома *всё же* вовсе не упоминается. Зато в АС есть отдельная словарная статья *всё же* (автор – В. Ю. Апресян), и на нее и следовало бы сослаться. Остается непонятным, почему идиома *один и тот же* приводится в форме среднего рода. А в отношении идиомы *как же* следовало бы добавить, что, хотя ироническое употребление выражения *Как же!* вполне конвенционализовалось, так что в подавляющем числе случаев оно означает отрицание (≈ ‘Нет’), но любопытно, что если оно предваряется союзом *а*, то значение подтверждения, выраженное риторическим вопросом, сохраняется: *А как же!* (≈ ‘Да, конечно; а как же иначе? ’)[5]. Собственно, материалы НКРЯ это только подтверждают. Ср., с одной стороны, пример с подразумеваемым иронически-экспрессивным отрицанием:

(15) *“... I know something that's better.” “I bet you don't. ...” – Я знаю средство получше. – Знаешь ты, как же! [Марк Твен. Приключения Тома Сойера (Н. Дарузес, 1950)]*

И с другой – всего лишь два из многочисленных примеров, в которых очевидно подтверждение, выраженное риторическим вопросом:

(16) *Is your husband going over there to-night? Oh, yes. – А ваш муж поедет туда нынче ночью? - Ну а как же! [Марк Твен. Приключения Гекльберри Финна (Н. Дарузес, 1950)]*

(17) *Will we see you at the camp? – Yes, man. – Мы увидимся в лагере? – А как же! [Эрнест Хемингуэй. По ком звонит колокол (Н. Волжина, Е. Калашникова, 1968)*

Таким образом, упоминание идиом с частицей *же* в АС тоже нуждается в некотором уточнении.

Перейдем к употреблению частицы *же* в вопросительных предложениях. Соответствующее значение (*же 2*) истолковано в АС следующим образом: «*Какой же А?* ‘Задавая вопрос, говорящий подчеркивает, что он не может знать ответа на него’ [после вопросительных слов *кто*, *что*, *какой*, *чей*, *который*, *сколько*, *где*, *куда*, *откуда*, *когда*, *как*, *почему*, *зачем*]». Не требуется обращения к англо-русскому подкорпусу НКРЯ, чтобы увидеть неточность в данном толковании. В самом деле, если учитель говорит ученику: «Вы говорите, что группа “Освобождение труда” была создана в 1883. Хорошо. А *кто же* в нее входил?», – то совершенно не предполагается, что учитель не знает ответа на заданный вопрос: напротив того, скорее всего учитель знает ответ и хочет проверить, знает ли его ученик. Можно добавить, что в некоторых «риторических» употреблениях вопросительное предложение с *же* выражает возражение (напр., *Что же в этом плохого?*).

[5] Ср. [Шмелев 2009: 186].

По-видимому, АС толковал эту частицу лишь в контексте прототипического (прямого, информационного) вопроса, не учитывая иллокутивного разнообразия вопросительных предложений с частицей *же* после вопросительного местоимения.

Перечисление в квадратных скобках вопросительных слов, после которых *же* может употребляться в рассматриваемом значении, наводит на мысль, что *же* может употребляться только в частных, но не в общих вопросах. Ср. в связи с этим беглое замечание (Падучева 2018: 344; сноска), согласно которому *же* «невозможно в общем вопросе (**Так пошел же он в магазин?*)». Впрочем, в этой же сноске, со ссылкой на работу Д. Пайара, Е. В. Падучева отмечает, что *же* все же возможно в вопросах с частицей *неужели* (*Неужели же ты не узнал свою работу?*), но не в вопросах с частицей *разве* (**Разве же ты не узнал свою работу?*). Заметим, что запрет на использование *же* в вопросах с частицей *разве* сформулирован излишне категорично: риторические вопросы, начинающиеся с *разве ж*, нередко используются в просторечии. Многим памятна сцена из «Архипелага ГУЛАГ»:

(18) *Так на истолчённой соломке пола стало нас восемь сапог к двери и четыре шинели. Они спали, я пылал. Чем самоуверенней я был капитаном полдня назад, тем больней было защемиться на дне этой каморки. Раз-другой ребята просыпались от затеклости бока, и мы разом переворачивались.*
К утру они отоспались, зевнули, крякнули, подобрали ноги, рассунулись в разные углы – и началось знакомство.
– А ты за что?
Но смутный ветерок настороженности уже опахнул меня под отравленной кровлею СМЕРШа, и я простосердечно удивился:
– Понятия не имею. Рази ж говорят, гады?

Просторечная форма частицы *разве* (*рази*) вполне согласуется с общей просторечной окраской высказывания.

В основном подкорпусе НКРЯ (новая версия) по данным на 1 марта 2021 нашлось 320 вхождений сочетания *разве же*, и, кстати, не всегда в них ощутим налет просторечия. Но особенно показательно, что три таких примера обнаруживаются и в англо-русском подкорпусе НКРЯ, причем во всех трех случаях переводчик стремится имитировать просторечную окраску высказывания:

(19) *Who's made appointments with him in the hold? Ain't that queer, now? – Кто это ему там в трюме свидания назначает? Ну разве ж это не странно? [Герман Мелвилл. Моби Дик (И. М. Бернштейн, 1961)]*

(20) *But I am not a brave man… – Так разве ж я храбрый человек… [Герман Мелвилл. Моби Дик (И. М. Бернштейн, 1961)]*

(21) *DOOLITTLE (remonstrating) Now, now, look here, Governor. Is this reasonable? Is it fairity to take advantage of a man like this? – ДУЛИТТЛ (протестующе). Но-но, хозяин, так не годится. Разве ж это по-честному? Разве так поступают с человеком? [Бернард Шоу. Пигмалион (Н. Рахманова, П. Мелкова, 1993)]*

В целом же можно согласиться с формулировкой, согласно которой *же* маркирует «вопрос, порожденный предшествующим контекстом» (Падучева 2018: 344)[6]. Конечно, в таком виде ее затруднительно использовать в словарном толковании рассматриваемого значения *же*. Общий смысл вопросов с частицей *же* можно было бы сформулировать, напр., так: ‘Ты понимаешь, что сложившаяся ситуация естественно вызывает данный вопрос’ (вопрос может быть и риторическим). Примеров, подтверждающих такое толкование, в англо-русском подкорпусе НКРЯ довольно много.

[6] В статье (Пекелис 2020), посвященной микроэволюции частицы *же* в вопросительных предложениях в языке XVIII–XIX вв., выделяется четыре разновидности вопроса с *же* в современном языке: недоумение, несогласие, непонимание, нетерпение.

Выходя за рамки НКРЯ, можно упомянуть историю, рассказанную в книге Лео Ростена *The Joys of Yiddish*:

(22) *Mr. Sokoloff has had dinner for twenty years in the same restaurant on the Second Avenue. This evening, as always, he orders bouillon. The waiter brings it, and wants to go back, but Mr. Sokoloff addresses him: "Waiter!" – "Yes, please?" – "Be so kind to taste this soup." – "But Mr. Sokoloff, you have come here for twenty years and you have never complained." – "Please", repeats Mr. Sokoloff obstinately, "taste this soup." – "But what is the matter, Mr. Sokoloff?" – "Please taste it." – "All right", the waiter says. "But… a moment. Where is the spoon?" – "*Aha!*", says Mr. Sokoloff.*

В (Шмелев 2015: 590) отмечалось, что почти все, кто прочел эту историю и хочет рассказать ее на русском языке, передают последние фразы официанта следующим образом: *Но… минуточку. Где же ложка?* Частица *же* почти неизбежно появляется в переводе и пересказе.

Иногда при помощи такого вопроса говорящий высказывает упрек адресату речи, напоминая ему о чем-то, что может служить основанием для упрека (*Что же ты не выполнил обещания*?). Но для англо-русского подкорпуса НКРЯ такие употребления почему-то совершенно не характерны. Возможно, такого рода упреки не характерны для англоязычного дискурса как такового.

Наконец, обратимся к самому частотному значению частицы *же*, которое обозначено в АС как *же 1*. Оно истолковано в АС следующим образом: «*А1 же* ‘Говорящий подчеркивает, что А1 должно быть известно адресату речи’». В комментарии к толкованию говорится: «Часто используется при выражении упрека, повторного требования, напоминания и выражает различные эмоции». Такие употребления характерны для диалогического режима, и в англо-русском подкорпусе НКРЯ они в основном встречаются в переводах диалогов (поэтому они не столь многочисленны, как можно было бы ожидать). Ср.:

(23) *I'd have gone anywhere. I said I'd go anywhere you wanted. – Я бы поехала куда угодно. Я же говорила, что поеду, куда только ты захочешь. [Эрнест Хемингуэй. Снега Килиманджаро (Н. Волжина, 1956)]*

(24) *I said I didn't want… – Я сказал же, что не хочу… [Уильям Голдинг. Повелитель мух (Е. Суриц, 1985)]*

(25) *And all the nicest things that Puzzle brought back were eaten by Shift; for as Shift said, "You see, Puzzle, I can't eat grass and thistles like you, so it's only fair I should make it up in other ways." – Все вкусное, что ослик приносил из города, съедал Хитр, и при этом говорил: «Ты же понимаешь, я не могу есть траву и чертополох как ты, и по справедливости я должен это чем-то возмещать». [Клайв Стейплз Льюис. Хроники Нарнии. Последняя битва (Г. А. Островская, 1991)]*

(26) *You know you're no good at thinking, Puzzle, so why don't you let me do your thinking for you? – Ты же знаешь, что не слишком хорошо умеешь думать, так уж позволь мне думать за тебя. [Клайв Стейплз Льюис. Хроники Нарнии. Последняя битва (Г. А. Островская, 1991)]*

Толкование АС представляется вполне адекватным. Возможно, впрочем, что глагол *подчеркивать* в толковании не вполне точен: говорящий скорее исходит из того, что А1 известно адресату речи, и использует А1 как обоснование чего-то, что, по его мнению, адресат речи не учитывает. При этом, как отмечалось в (Падучева 2018: 341), аргумент, выраженный при помощи предложения с *же*, «обычно слабый, как бы не решающий». Часто *же* выражает раздражение по поводу пренебрежения очевидностью.

Заметим, что в контекстах такого рода мы сталкиваемся с упомянутой выше разновидностью прагматической обязательности. В большинстве случаев *же 1* может быть заменено на частицу *ведь* (позиция которой в предложении более свободна). Однако отсутствие какого бы то ни было показателя сделало бы высказывание прагматически неадекватным: оно не могло бы восприниматься как аргумент, пусть даже слабый.

Итак, материал англо-русского подкорпуса НКРЯ позволяет уточнить описание частицы *же*, данное в АС. В частности, было высказано предположение, что в сочетаниях с *т*-местоимениями *же* следовало бы описывать как постфикс (а соответствующие сочетания – как производные местоимения). Разумеется, такие уточнения можно было бы сделать и без опоры на параллельный корпус, используя лишь оригинальные русские тексты, а также лингвистический эксперимент. Однако параллельный корпус позволяет представить некоторые свойства частицы максимально наглядно. Выявляя причины, которые побудили переводчиков использовать частицу *же* в своих переводах мы получаем шанс увидеть какие-то ее свойства, которые в противном случае могли бы ускользнуть от внимания. Остается вопрос: если бы подключить к анализу материалы переводов с других языков, насколько это повлияло бы на общие выводы? Этот вопрос нуждается в дальнейшем исследовании.

References

[1] Dobrovolskii D. O., Levontina I. B. (2012) Synonymous focus particles in German and Russian, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue'2012*, vol. 1, Moscow, pp. 138–149.

[2] Dobrovolskii D. O., Levontina I. B. (2014) Discourse words in general questions: Russian-German near-equivalents, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue'2014*, Moscow, pp. 138–149.

[3] Dobrovolskii D. O., Levontina I. B. (2015) Modal Particles and the Actualization of Forgotten Details (Based on the Materials of Parallel Corpora), *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue'2015*, vol. 1, Moscow, pp. 104–117.

[4] Dobrovolskii D. O., Levontina I. B. (2017) Discourse Particles and Their Translation: *Nu* in Vladimir Sorokin's Novel *The Queue* (*Ochered'*), *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue'2017*, vol. 2, Moscow, pp. 106–117.

[5] Kobozeva I. M., Orlova S. V. (2008) Unicellular organisms of communication under a microscope: German particle *ja* versus its Russian translation equivalents *ved'* and *že*, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue'2008*, Moscow, pp. 199–205.

[6] Levontina I. B. (2017) **Zhe**, V. Apresjan, I. Galaktionova, B. Iomdin (eds.) *Active Dictionary of Russian*, [*Aktivnyi slovar' russkogo iazyka*], vol. 3, Nestor-Istoriia, Moscow; Saint Petersburg.

[7] Levontina I. B. (1999), Strategies of coaxing: particles in repeated requests [Strategii ugovarivaniia: chastitsy v povtornykh pros'bakh], *Language. Culture. Humanities. Scientific heritage of G. O. Vinokur today* [*Iazyk. Kul'tura. Gumanitarnoe znanie. Nauchnoe nasledie G. O. Vinokura i sovremennost'*], Nauchnyi Mir Publ., Moscow, pp. 188–201.

[8] Levontina I., Shmelev A. (2005), The particles one cannot do without // *East—West Encounter: Second International Conference on Meaning ↔ Text Theory*, Slavic Culture Languages Publishing House, Moscow, pp. 258-267.

[9] Padučeva E. V. (1988) The particle *zhe*: semantics, syntax, and prosody [La particule ŽE: semantique, syntaxe et prosodie], *Narrative particles in modern Russian* [*Les particules énonciatives en Russe contemporain*], Institut d'études slaves, Paris, vol. 3, pp. 11–44.

[10] Paducheva E. V. (2018) *Egocentric linguistic items* [*Egotsentricheskie edinitsy iazyka*], Izdatel'skii Dom IaSK Publ., Moscow, 2018

[11] Pekelis O. E. (2020) A Case of Pragmaticalization in Russian: Micro-diachronic Analysis of the Particle že in Questions. *Slověne*, vol. 9, № 1, pp. 340–361.

[12] Shmelev A. D. (2009) "Void" and "unexpressed" negation ["Neznachashchee" i "nevyrazhennoe" otritsanie], *Logical Analysis of Language: Assertion and Negation* [*Logicheskii analiz iazyka: assertsiia i negatsiia*], Indrik Publ., Moscow, pp. 173–202.

[13] 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*, vol. 1, Moscow, pp. 584–594.

[14] Shmelev A. D., Zalizniak 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.

[15] Zalizniak Anna, Shmelev A. (2017) Russian discourse words in the light of parallel corpora [Russkie diskursivnye slova po dannym parallel'nykh korpusov], available at: https://www.aatseel.org/100111/pdf/abstracts/1419/Zalizniak.pdf

Литература

[1] Levontina I., Shmelev A. The particles one cannot do without // East—West Encounter: Second International Conference on Meaning ↔ Text Theory. — Moscow: Slavic Culture Languages Publishing House, 2005. — P. 258-267.

[2] Padučeva E. V. La particule ŽE: semantique, syntaxe et prosodie // Les particules énonciatives en Russe contemporain. — Paris : Institut d'études slaves, vol. 3, 1988. — P. 11–44.

[3] Zalizniak Anna, Shmelev A. Русские дискурсивные слова по данным параллельных корпусов // AATSEEL: Program of the 2017 Annual Meeting, https://www.aatseel.org/100111/pdf/abstracts/1419/Zalizniak.pdf

[4] Добровольский Д. О., Левонтина И. Б. О синонимии фокусирующих частиц (на материале немецкого и русского языков) // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2012. Т. 1. — Москва, 2012. — С. 138–149.

[5] Добровольский Д. О., Левонтина И. Б. Дискурсивные слова в общевопросительных предложениях: русско-немецкие соответствия // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2014. — Москва, 2014. — С. 138–149.

[6] Добровольский Д. О., Левонтина И. Б. Модальные частицы и идея актуализации забытого (на материале параллельных корпусов) // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015. Т. 1. — Москва, 2015. — С. 104–117.

[7] Добровольский Д. О., Левонтина И. Б. Дискурсивные частицы и способы их перевода: 'ну' в романе Владимира Сорокина «Очередь» // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2017. Т. 2. — Москва, 2017. — С. 106–117.

[8] Кобозева И. М., Орлова С. В. Одноклеточные организмы общения под микроскопом: немецкая частица *ja* в сопоставлении с ее переводными эквивалентами *ведь* и *же* // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2008. — Москва, 2008. — С. 199–205.

[9] Левонтина И. Б. ***Же*** // Активный словарь русского языка. — Т. 3. Редакторы тома В. Ю. Апресян, И. В. Галактионова, Б. Л. Иомдин. Под общим руководством акад. РАН Ю. Д. Апресяна. — Москва; СПб.: Нестор-История, 2017. — С. 374–375.

[10] Левонтина И. Б. Стратегии уговаривания: частицы в повторных просьбах // Язык. Культура, Гуманитарное знание. Научное наследие Г. О. Винокура и современность. — Москва: Научный мир, 1999. — С. 188–201).

[11] Падучева Е. В. Эгоцентрические единицы языка, — Москва: Издательский дом ЯСК, 2018.

[12] Пекелис О. Е. Об одном случае прагматикализации в русском языке: микродиахроническое исследование частицы *же* в составе вопроса // Slověne, 2020. — Vol. 9, № 1. — С. 340–361.

[13] Шмелев А. Д. «Незначащее» и «невыраженное» отрицание (когнитивные и коммуникативные источники энантиосемии) // Логический анализ языка. Ассерция и негация. Москва: Индрик, 2009. — С. 173–202.

[14] Шмелев А. Д. Русские лингвоспецифичные лексические единицы в параллельных корпусах: возможности исследования и «подводные камни» // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015. Т. 1. — Москва, 2015. — С. 584–594.

[15] Шмелев А. Д., Зализняк Анна А. Реверсивный перевод как инструмент лингвистического анализа дискурсивных слов // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2017. — Москва, 2017. — С. 370–380.

Automatic Detection of Implicit Aggression in Russian Social Media Comments

Shulginov V.A.
Higher School of Economics /
Laboratory of Linguistic Conflict
Resolution Studies and Contemporary
Communicative Practices
shulginov.val@yandex.ru

Mustafin R. Zh.
Higher School of Economics /
Laboratory of Linguistic Conflict
Resolution Studies and Contemporary
Communicative Practices
rmustafin.art@gmail.com

Tillabaeva A.A.
Higher School of Economics /
Laboratory of Linguistic Conflict
Resolution Studies and Contemporary
Communicative Practices
alinka99-t@mail.ru

Abstract

This article studies the characteristics of implicit and explicit types of aggression in the comments of a Russian social network with the means of machine learning. As it is hypothesized that expression of aggression depends on local norms, the dataset contains the comments collected from a single social media community. These comments were divided into three classes: polite communication, implicit aggression, and explicit aggression. Trying different combinations of data preprocessing, we discovered that lemmatization and replacement emojis with placeholders contribute to better results. We tested several models (Naive Bayes, Logistic Regression, Linear Classifiers with SGD Training, Random Forest, XGBoost, RuBERT) and compared their results. The study describes the misclassifications and compares the keywords of each class of comments. The results can be helpful while enhancing the algorithm of detection of implicit aggression

Keywords: politeness; verbal aggression; implicit aggression; machine learning; social media
DOI: 10.28995/2075-7182-2021-20-636-645

Автоматическое определение скрытой речевой агрессии в русскоязычных социальных сетях

Аннотация

В данной статье представлен принцип автоматического определения характеристик имплицитной и эксплицитной речевой агрессии в русскоязычных социальных сетях. Поскольку предполагается, что проявление агрессии зависит от локальных коммуникативных норм, датасет содержит комментарии, опубликованные в одном интернет-сообществе. Эти комментарии были разделены на три класса: кооперативная коммуникация, скрытая речевая агрессия и явная речевая агрессия. Используя различные комбинации признаков предобработки данных, мы обнаружили, что лемматизация и замена эмодзи на плейсхолдеры способствуют получению лучших результатов. Кроме того, мы протестировали несколько моделей машинного обучения (Naive Bayes, Logistic Regression, Linear Classifiers with SGD Training, Random Forest, XGBoost, RuBERT) и сравнили их результаты. В исследовании описываются ошибки классификации и сравниваются набор лексических маркеров для каждого класса комментариев. Полученные результаты могут быть полезны при усовершенствовании алгоритма обнаружения скрытой агрессии.

Ключевые слова: лингвистическая вежливость; речевая агрессия; скрытая агрессия; машинное обучение; социальные сети

1 Introduction

Aggressive behavior has a negative impact on participants of online communication [Warner & Hirschberg, 2012]. Social media such as Facebook and Twitter state in the usage policy [5], [12] that they are concerned about hateful user-generated content (abusive language, hate speech, cyberbullying, and trolling). Automation tools detecting aggression may be used to help moderators. For example, the Russian social network VKontakte introduced a new function available for the administrators and moderators of online communities: they will be able to filter the comments containing threats or hate speech.

It is important to note that the task of automatic detection of verbal aggression differs from the task of sentiment analysis [Cambria et al., 2017]; [Lukashevich, 2017]. It has not been studied properly because aggression is a complex sociocultural phenomenon. Verbal behaviour cannot be described by the dichotomy of aggression and politeness. It is more likely to be a continuum between two poles: completely rude interaction and polite respectful interaction [Locher, 2006]. In certain contexts, the words that are usually attributed to impolite behavior can be used either in cooperative or confrontational interaction. While detecting verbal aggression, ideally, we should consider both the intention of the speaker to conduct a face-threatening act and the hearer's perception of that. It might be traced if we consider the broad context of the message. However, even for human beings, it is difficult to determine the initial intention of the speaker and the internal state of the recipient, especially when the aggression is implicit and expressed with sarcastic, insincere politeness. There have already been attempts of automatic aggression detection in social media. The International Workshop on Semantic Evaluation SemEval-2019 motivated many researchers to study this topic. One of its tasks required identifying and categorizing offensive language in social media. Although our study is quite similar to that subtask, there are considerable differences. In addition to the basic differences in the language of texts and the source of data, a substantial difference in taxonomy must be mentioned. The works presented on SemEval-2019 included only two classes: not offensive and offensive. Offensive texts were characterized by the presence of obscene words. In contrast, we distinguish three classes of comments (polite, explicitly aggressive and implicitly aggressive) paying attention to the type of intention. Verbal aggression often contains vulgar lexis but it is not a definitive factor. This approach allows to detect the implicit aggression including sarcasm and irony which may not have vulgar lexis as their distinctive lexical features. Another workshop that should be considered is the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1). The taxonomy of TRAC included three classes just as the taxonomy of our study: overtly aggressive, covertly aggressive and non-aggressive texts [Aroyehun & Gelbukh, 2018]. However, the criteria of labelling were not described properly, and it was not clear what kind of texts were defined as covertly aggressive.

The aim of this paper is to study the features of implicit and explicit aggression using the means of machine learning. There are six tasks to be completed: to create a model detecting both explicit and implicit verbal aggression in Russian social media comments; to select and collect a corpus of comments; to preprocess the data obtained; to select the methods of data processing; to train the model; to test the model and to draw conclusions.

2 Methods

2.1 Data Collection

As stated above, we focus on social media comments. The data were collected in the social network Vkontakte, more specifically in the BORSCH (БОРЩ) community. The decision to select a particular community for data collection was motivated by the hypothesis that the way how aggression is expressed varies depending on the local norms and community standards.

Local norms are formed in communities of practice, which define the social engagements. Penelope Eckert states that "a community of practice is an aggregate of people who come together around mutual engagement in an endeavor" [Eckert, 2006]. The conventionalization of meaning is the result of the collective activity in which they share their experience. The three main criteria for identifying a community of practice, according to Wenger [1998], are (1) mutual engagement; (2) a joint enterprise; and (3) a shared repertoire. Thus, each community of practice has its own standards of aggressive communication.

The presumed consistency of the norm within a single community was supposed to contribute to better results. The total number of the comments collected was 28,272. Then a subset of the comments that were written in reply to another comment and containing more than three words was chosen. Such restrictions were meant to provide more detailed context helping the annotators label data more accurately, that was especially challenging in case of implicit aggression. The decision to select comments that were written in reply to other comments can be explained by the fact that annotators need the context to label them. The imbalanced dataset contains 7,225 comments in total: 5,058 comments in the training and 2,167 comments in the test dataset. The balanced dataset includes 5,687 comments in total: 3,984 comments in the training dataset (1,328 comments in each class) and 1,703 comments in the test dataset [https://github.com/alinatl/Implicit-Aggression].

2.2 Taxonomy and Labeling

The comments were divided into three classes of comments: polite (cooperative) comments, explicitly aggressive comments and the implicitly aggressive ones. Annotators marked up each comment using one of the three labels and considering the context of its use (previous and subsequent comment in the thread). Lexical markers and the strategy of each participant were taken into account:

0 - polite (cooperative) comments. It is the class of comments which correspond to the Grice's Cooperative Principle. According to Leech, politeness is "a constraint observed in human communicative behaviour, influencing us to avoid communicative discord or offence, and maintain communicative concord" [Grice, 2007]. These comments do not contain face-threatening acts towards another participant or any social groups.

1 - implicitly aggressive comments. The speaker performs a face-threatening act using politeness strategies which are clearly insincere. The speaker's intent is exhibited only in the context. There are the following key markers:

- vocatives that do not bear negative connotation by themselves (*старик — old man, сынок — boy, дружочек — little buddy, девушка — girl, оно — it*);
- markers of politeness and impoliteness intertwined (*Хорошая история. Жаль, что враньё. It's a good story. Too bad it's a lie*);
- question containing implicit aggression *(Ты глупый? Are you silly?*);
- offensive expressions exhibiting emotional state of the speaker (*Бля, как можно этого не видеть. Shit, how can you not see that?*);

2 - explicitly aggressive comments. In these comments, face-threatening acts are performed in a direct, clear, unambiguous and concise way in circumstances where face is not irrelevant or minimized. The comments contain different types of insults (personalized negative vocatives, personalised negative assertions, personalised negative references, personalised third-person references that are negative from the point of view of the target), name-calling, casting aspersions and pejorative speech.

Each comment was manually labelled by two annotators in order to minimize inaccuracy. When the labels did not match, the cases were discussed collectively and the disagreement was mitigated. Besides, the dataset was balanced, i.e. the number of comments in each class was the same.

2.3 Data Preprocessing

On purpose or unintentionally, people tend to change the graphic form of words in online communication. That is why one of the crucial tasks of preprocessing is to unify the data keeping the potential markers of verbal aggression.

The first stage of preprocessing includes tokenization [https://www.nltk.org/], casting all the words to lowercase and spelling correction with YandexSpeller [https://github.com/oriontvv/pyaspeller]. These operations were done in all variants of preprocessing because spelling mistakes can worsen the results. In the second stage, we tried all combinations of the methods related to six independent token-based features (lemmatization, emoji, punctuation, named entities, vulgar words, stopwords). All of them might be meaningful for aggression detection. During the actual experiments all possible combinations of preprocessing methods are tested in terms of the model score. There were 192 possible combinations in total. All possible characteristics can be seen in Table 1.

Lemmatization, punctuation removal, vulgar words removal and stopwords removal could be executed or not. On the one hand, participants of the community which was chosen for this study do not

usually follow the rules of punctuation. In order to mitigate the inconsistency of punctuation in the corpus, punctuation marks can be removed. Stopwords, including prepositions, conjunctions and particles, can be also removed, because they do not contain any meaningful information, which is a common NLP practice. On the other hand, such fields of study as stylometry do not exclude stopwords nor punctuation marks when the method based on N-grams is applied. That is why all possible variants should be tried.

High variability of word forms in the Russian language makes it reasonable to apply lemmatization. Due to the typological features of Russian as a fusional language, one word can have many forms. In order to allow an algorithm to consider all the forms as one word, they are lemmatized.

Replacement of named entities and emojis with placeholders is applied when the dataset size is small and the number of distinct named entities, vulgar words and emojis is insufficient for statistical analysis. When these tokens are replaced, the information about particular named entities and emojis is lost. Nevertheless, this preprocessing method makes it possible to examine whether their presence constitutes a significant feature or not. Another possible variant is to remove all named-entities and emojis. Besides, emojis can be also replaced with specific classifying labels instead of placeholders: positive, negative or neutral.

All the parameters of preprocessing and the methods can be seen in Table 1.

Lemmatization		**Emoji**				**Punctuation**		**Named entities**			**Vulgar words**		**Stopwords**	
Yes	No	Keep	Remove	Replace with placeholders	Replace with labels	Keep	Remove	Keep	Remove	Replace with placeholders	Keep	Remove	Keep	Remove

Table 1. Parameters of data preprocessing

We used TF-IDF vectorizer because it is characterized by the adequate balance between high quality of vectorization and computational complexity.

2.4 Training Aggression Detection Models

Five models were trained to determine the baseline: Naive Bayes, Logistic Regression, Linear Classifiers with SGD Training, Random Forest, and XGBoost. The baseline model using simple algorithms should be surpassed by the final model. This baseline is established because there were no examples of studies with absolutely identical research design. We include the class of implicit aggression and focus on intention rather than on lexis while labelling.

The performance metric used was weighted average f1. Thus, 960 models were trained: 192 combinations of preprocessing techniques for each of 5 classifiers. The detailed explanation of how the best preprocessing type was selected is provided below.

We selected the top 20 models with the highest f1-score for each classifier. It allowed us to select the most stable preprocessing pipeline and model kind (i.e. classifier) showing the highest results.

No.	**logreg**	**xgboost**	**bayes**	**forest**	**sgd**
83	0.596	0.558	-	0.566	0.586
19	0.594	0.562	-	0.572	0.587

131	0.591	0.556	-		0.588
51	0.595	0.559	-	0.562	-
80	-	-	0.578	0.558	0.587
87	0.590	0.557	-	0562	-
81	0.592	0.559	-	0.561	-
21	0.590	0.556	-	0.565	-

Table 2. Types of preprocessing and the top-ranked results

Table 2 exhibits the types of preprocessing that were included in the top 20 results with at least 3 model types showing top-ranked results. According to it, the most successful variants of preprocessing pipeline for the majority of the model types were the variants 83 and 19. The methods applied in the five best types of preprocessing are provided in Table 3.

No.	**emojis**	**lemmatization**	**NER**	**punctuation**	**stopwords**	**vulgar**
21	replace	yes	no	keep	keep	del
51	del	yes	del	keep	keep	del
19	del	yes	no	keep	keep	del
83	del	yes	replace	keep	keep	del
87	label	yes	replace	keep	keep	del
81	no	yes	replace	keep	keep	del
131	del	no	del	del	keep	del
80	no	yes	replace	keep	keep	keep

Table 3. A comparison of preprocessing methods

7 out of 8 preprocessing variants included lemmatization, vulgar words deletion also appeared to be successful (7/8), in half of the cases the named entities were replaced by placeholders and in half of the cases emojis were deleted.

Figure 1. Value of f1-score metrics for all model types based on the results of 10 iterations

We calculated average values of the f1 score for models in top 20. From the table obtained we can state that Bayes classifier (0.579), Logistic Regression (0.591), and Linear Classifiers with SGD Training (0.587) demonstrated the best performance. Thus, Logistic Regression classifier combined with relevant preprocessing achieved the highest score and was defined as the baseline.

In order to outperform the baseline, we fine-tuned a transformer neural network based on the RuBERT model. This is a BERT (Bidirectional Encoder Representations From Transformers) trained on the set of Russian texts from the corresponding Wikipedia branch. That transformer language model achieves state-of-the-art results in a broad range of NLP tasks [Devlin J. and al., 2019]. For RuBERT model training we use both balanced and imbalanced datasets.

2.5 Keywords

When the models are ready, the following task is to define the lists of keywords for each class among the comments that the model labelled correctly, to compare the lists and to examine if there are regular patterns in terms of lexis.

We used three algorithms for selecting the keywords of each class: RAKE, Text Rank and Summa. RAKE (Rapid Automatic Keyword Extraction) calculates the weight of keywords (word scores) for words and phrases split by stopwords and punctuation marks. Tokens are presented as arrays and then are split into sequences of contiguous words at phrase delimiters and stop word positions. That is why this algorithm often selects collocations. We used the Python RAKE module [https://github.com/fabianvf/python-rake] with the following parameters: maxWords = 3, minFrequency = 2, whereas maxWords is the maximum number of the keywords, and minFrequency is the minimum keywords occurrence. TextRank and Summa are graph-based ranking algorithms that function as a voting and recommendation system that takes into account the relationships between words (vertices). As a result, we detected the keywords for all the classes. 10 keywords for each class are presented in Table 3 (0 - politeness, 1 - implicit impoliteness, 2 - explicit impoliteness).

RAKE			**TextRank**			**Summa**		
Class 0	Class 1	Class 2	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2
крайняя мера last resort	общая генная цепочка common gene sequence	emoji	это it	это it	что what	всё every-thing	свой own	emoji
создать to create	emoji	своя мамаша own mom	что what	что если what if	emoji	emoji	твой yours	твой yours
случайность accident	достаточно самомнениее go enough	твой интернет your internet	все быть all be	emoji	твои your	весь entire	emoji	свой own
иметь to have	покупательская способность buying power	падаль старая old carrion	how	твои yours	own	year film	человек human	идти to go

Время time	религиозное чувство religious feeling	весь похуй all fuck	так so	свои own	идти to go	мочь понять человека to be able to understand a human	который which	который which
человек human	интеллект медузы jellyfish intelligence	мерзость filth	если они if they	как how	все all	просто just	мочь can	всё everything
хороший good	мочь повлиять to be able to affect	тупой Вася dumb Vasya	там there	они they	этот this	наш our	весь entire	тупой dumb
сильный strong	право твоё your right	свой own	можно can	так so	быть to be	хороший good	просто just	человек human
цена price	просто видеть just see	делать to do	только only	такой this	тупой dumb	большой large	хороший good	жопа ass
маска mask	признавать начало to recognize the beginning	жить to live	emoji	человек human	клоун clown	новый new	начать to start	ебать to fuck

Table 4. Keywords lists

Having analyzed the intersection of the words of three classes, we explored the main topics of the community “BORCH”. We also detected the keywords of each class. The words in the lists were categorized into several semantic groups. The results obtained are discussed in the next section.

3 Results

The baseline model was the model using Logistic Regression. The best variants of preprocessing were the variants number 19 and 83. They both included vulgar words and emojis deletion and lemmatization. In the 19th variant, named entities were not removed, while in the 83rd they were replaced with placeholders. Both in 19th and 83rd punctuation marks and stopwords were kept. This might indicate that users omit punctuation marks when using obscene vocabulary. The conclusion to be drawn is that the fact of the presence of named entities helps the model detect verbal aggression in social media comments.

№	Model	F1	Precision	Recall
19	Naive Bayes	0.56	0.57	0.56
	Log Reg	0.60	0.60	0.60
	SGD	0.59	0.59	0.59
	Random Forest	0.57	0.58	0.58

	XG Boost	0.56	0.57	0.57
83	Naive Bayes	0.56	0.57	0.56
	Log Reg	0.60	0.60	0.60
	SGD	0.59	0.58	0.59
	Random Forest	0.57	0.57	0.57
	XG Boost	0.56	0.58	0.57
80	Naive Bayes	0.58	0.58	0.58
	Log Reg	0.59	0.59	0.59
	SGD	0.58	0.58	0.58
	Random Forest	0.56	0.56	0.57
	XG Boost	0.54	0.56	0.56
RuBERT balanced		0.65	0.65	0.66
RuBERT imbalanced		0.66	0.66	0.67

Table 5. F1-score, precision and recall

The model based on RuBert demonstrated the highest score (0.66) on imbalanced dataset in automatic detection of aggression and overcame the baseline model Logistic Regression (0.60). At the same time, attention should be paid to the high accuracy of the explicit aggression detection (0.82).

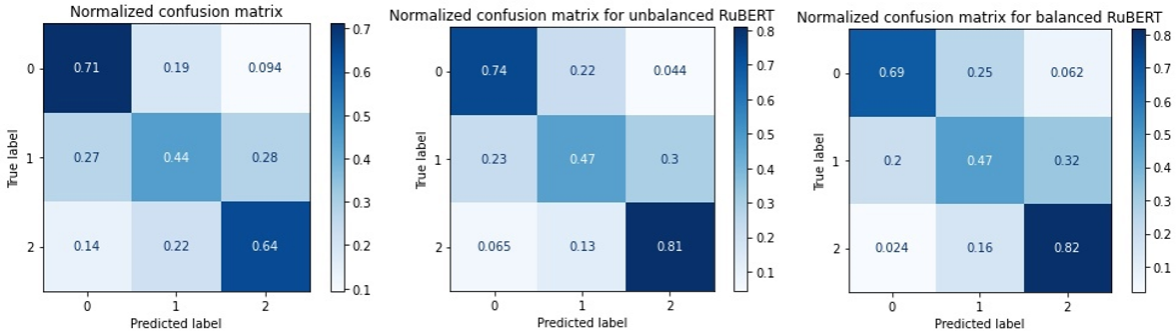


Figure 2.Confusion matrix of the Logistic Regression model with the 19th preprocessing variant/ the balanced RuBERT model/ the unbalanced RuBERT model

Having analysed the misclassifications of both models, we discovered that the model based on RuBERT attributed the comments of the class 1 correctly in 47% of the cases. It made a mistake and attributed the comments of that class to the class 0 in 36% of cases and to the class 2 in 17% of cases. The misclassifications between the classes 0 and 1 can be explained by the similarity in terms of their lexical features. It can be proved by the keywords representing each class. The keywords extracted for each class by all three algorithms (RAKE, Text Rank and Summa) are shown in Table 5.

Class 0	вода (water), значит (to mean), пользоваться (to use) место (place), кстати (by the way), говорить (to talk), время (time), маска (mask), бумажка (paper), классика (classics), вообще (at all), менять (to change), начать (to start), статистика (statistics), вариант (option), думать (to think), Россия (Russia), ситуация (situation), рубль (ruble), государство (state), просто (simply), посмотреть (to look), пример (example), остаться (stay), шина (tire), офигеть (be shocked), построить (to build), emoji, право (right), никто (nobody), курс (course), жить (to live),строить (to build)
Class 1	решить (to decide), месяц (month), ответ (answer), дело (case), жизнь (life), платить (to pay), значит (to mean), начало (beginning), заслужить (to deserve), никто (nobody), говорить (to speak), жопа (ass), точно (accurately), считать (to consider), хотеть (to want),

	видео (video), уровень (level), мочь (to be able to), мнение (opinion), перечитать (to reread), ждать (to wait), зарплата (salary), работать (to work), верить (to believe), человек (human), вообще (at all), сидеть (to sit), мама (mom), Россия (Russia), благодаря (due to), развивать (to develop), пора (it's time), невозможно (impossible), доказать (to prove), рубль (ruble), работа (job), весь (entire), видеть (to see), государство (state), молодец, давать (to give), Путин (Putin), пост (post), сказать (to say), посмотреть (to watch), Вася (Vasya). ясно (clear), жить (to live), показать (to show), таракан (cockroach), норма (norm), emoji, слово (word), покупать (to buy), понять (to understand), делать (to do), проблема (problem), понимать (to understand), продолжать (to continue), глаз (eye)
Class 2	параша (slop-pail), нормально (normal), дебил (moron), знать (to know), вместо (instead), шлюха (slut), говорить (to talk), жопа (ass), понятно (clear), хотеть (to want), читать (to read), ебать (to fuck), мразь (scum), мамкин (mom's), мочь (to be able), написать (to write), сосать (to suck), kremlebot, высер, говно (shit), ватник, лахта, пиздец, власть (authorities), почему (why), сказать (to say), смотреть (to watch), дурачок (fool), жить (to live), друг (friend), emoji, идти (to go), слово (word), делать (to do), работать (to work), понимать (to understand), сука (bitch)

Table 6. Keywords for each class

As demonstrated in Table 5, the class 2 contains the largest number of expressive lexis. There are several pejorative words marking political and ideological views (*(1) Kremlebot, (2) vatnik (3) lahta*), insults referred to promiscuity (*(4) slut, (5) to fuck, (6) to suck*), insults of family members (*(7) Mom's* (mamkin)) and scatological terms (*(8) shit*). The keywords of the comments with implicit aggression do not have offensive meaning by themselves and in many cases coincide with the words of class 0 which typify the main theme of the community (*(9) state, (10) right, (11) situation, (12) ruble, (13) mask*). It demonstrates that the lexical-based approach [Njagi et al, 2015] of aggression detection is not effective. Marked words in the class 1 can be used without addressee (*(14) ass*) or imply aggression only in particular contexts (*(15)Vasya* (this name is associated with a simpleton, a foolish person)*, (16) cockroach (*the Belarusian president's derogatory nickname)).

The keywords that are unique for the class of implicit aggression can contain substandard words (*(17) big head (bashka), (18) to get drunk, (19) to shit up*) or just name verbal aggression (*(20) boorish, (21) rudeness*) but they are not invective.

4 Conclusions and Future Work

The purpose of this study was to explore approaches to the automatic detection of implicit aggression in comparative perspective with the detection of explicitly aggressive and polite speech. The article discussed data collection, preprocessing and train modelling.

Several conclusions were drawn. First, we discovered that on the stage of data preparation lemmatization and keeping stopwords and punctuation marks contribute to better results. We also suppose that the comments with vulgar lexis do not contain punctuation marks more often. Second, certain similarities between polite communication and implicit aggression in terms of keywords make lexical features insufficient for accurate detection of implicit aggression. Third, the winning model was the model based on RuBERT. The f1 of this algorithm is 0.66, which is higher than the baseline and the best result presented for the similar task in TRAC-1 (f1 0.64). This result is still lower than the best result achieved by the participants of SemEval-2019 (f1 0.82), but it can be explained by a substantial difference in taxonomy. Our taxonomy includes not only polite and explicitly aggressive comments but also implicitly aggressive. This class is interjacent: it is at the same time closer to the explicitly aggressive class in terms of intention and to the polite class in terms of vocabulary. It allows to detect sarcasm and irony which do not bear any specific lexical markers. However, the absence of such markers worsen the results.

As for possible solutions to the problem of low accuracy, several solutions might be proposed. For instance, we could specify the taxonomy, analyse other linguistic features of implicit aggression (syntax,

POS), consider pragmatics and identify interjacent classes between polite and impolite communication. The size of the dataset is, probably, not ample to ensure the stable work of the model.

The topic of automatic detection of implicit aggression in social media has many paths for further research. This study can be used as a base for the future research of implicit aggression as a linguistic phenomenon and automatic detection of aggression in communication. For instance, it is possible to use crowd-sourcing and to create a larger dataset, to collect a corpus of other languages or use other social media as a source. The methods also can vary: further studies might classify comments differently, consider other linguistic features or choose alternative ways of data processing.

References

[1] Aroyehun Segun T., Gelbukh Alexander. Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling // Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC) 2018. — Santa Fe, USA, 2018. — P. 90–97.

[2] Cambria Erik, Poria Soujanya, Gelbukh Alexander, and Thelwall M. Sentiment analysis is a big suitcase. — IEEE Intelligent Systems, Vol. 32(6), pp. 74–80.

[3] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). — Minneapolis, USA, 2019.— P. 4171–4186.

[4] Eckert Penelope. Communities of Practice // Encyclopedia of Language and Linguistics / ed. by K. Brown. 2nd edition. Amsterdam: Elsevier, 2006, pp. 683–685

[5] Facebook's Policy & Usage Guidelines. Access mode: https://developers.facebook.com/docs/messenger-platform/policy/

[6] Grice Paul. Logic and conversation // Syntax and Semantics. Vol. 3: Speech Acts / ed. by P. Cole, J. Morgan. New York: Academic Press, 1975, pp. 41–58.

[7] Kecskes Istvan. Intercultural Pragmatics. — Oxford: Oxford University Press, 2014.

[8] Kumar Ritesh, Ojha Atul Kr, Malmasi Shervin, Zampieri Marcos. Benchmarking aggression identification in social media // Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC) 2018. — Santa Fe, USA, 2018. — P. 1–11.

[9] Leech G. N. Politeness: Is there an East-West divide? // Journal of Politeness Research. 2007. Vol. 3 (2), pp.167–206.

[10] Levin Y.I. About obscene expressions of the Russian language [Ob obscennih virajeniyah russkogo yazika]. Selected Works. Poetics. Semiotics. [Izbrannie trudi. Poetika. Cemiotika], Moscow, 1998, pp. 809-819.

[11] Locher M. A. Polite behavior within relational work: The discursive approach to politeness // Multilingua. Journal of Cross-Cultural and Interlanguage Communication. 2006. Vol. 25 (3), pp. 249–267.

[12] Lukashevich N.V. Automatic sentiment analysis methods [Avtomaticheskie metody analiza tonal'nosti], Automatic natural language text processing and data analysis [Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannyh], NIU HSE, Moscow, pp. 127-179

[13] Twitter Rules and policies. Access mode: https://help.twitter.com/en/rules-and-policies

[14] Warner W., and Hirschberg Julia. Detecting hate speech on the world wide web // Proceedings of the Second Workshop on Language in Social Media 2012. — Stroudsburg, USA,2012. — P. 19–26.

[15] Wenger, E. Communities of Practice: Learning, Meaning, and Identity. Cambridge: Cambridge University Press, 1998.

[16] Zampieri Marcos, Malmasi Shervin, Nakov Preslav, Rosenthal Sara, Farra Noura, Kumar Ritesh. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) // Proceedings of the 13th International Workshop on Semantic Evaluation. — Minneapolis, USA, 2019. — P. 75–86.

The Dynamics of Vocabulary in Russian Prose (Based on Frequency Dictionaries of the Corpus of Russian Short Stories 1900-1930)

Tatiana G. Skrebtsova
Saint Petersburg State University
Saint Petersburg, Russia
t.skrebtsova@spbu.ru

Alexander O. Grebennikov
Saint Petersburg State University
Saint Petersburg, Russia
a.grebennikov@spbu.ru

Tatiana Yu. Sherstinova
National Research University Higher School of Economics
Saint Petersburg, Russia
tsherstinova@hse.ru

Abstract

The paper presents the results of a study that is part of a large-scale project aimed at studying the changes that took place in the Russian language during the first three decades of the 20th century. In the history of Russia, this period was marked by stormy events that led to a radical change in the state system and the formation of a new society. To quantify the scale of changes that occurred in the language in the result of these dramatic events, it is necessary to analyze the representative volume of linguistic data and to compare different chronological periods in dynamics using quantitative methods. The research was carried out on the data of an annotated sample from the Corpus of the Russian Short Stories of 1900-1930, which contains texts by 300 Russian writers. All the texts in the Corpus are divided into three time frames: 1) the pre-war period (1900-1913), 2) the war and revolutionary years (1914-1922) and 3) the early Soviet period (1923-1930). Frequency distribution of significant vocabulary in dynamics was analyzed, which made it possible to identify the main tendencies in the change of individual words and lexical groups frequencies from one historical period to another and to correlate them with the previously identified dynamics of literary themes. The technique used allows to trace the influence of large-scale political changes on the vocabulary of literary language, to note the peculiarities and tendencies of the writers' worldview in a certain historical period, and also makes it possible to significantly supplement the analysis of the dynamics of literary themes in fiction.

Keywords: lexical studies; lexical changes in diachrony; language and style of literary texts; Russian short story; frequency dictionary; corpus linguistics; computational linguistics

DOI: 10.28995/2075-7182-2021-20-646-659

Динамика лексического состава русской художественной прозы (на материале частотных словарей корпуса русских рассказов 1900-1930)

Т. Г. Скребцова
Санкт-Петербургский государственный университет;
Санкт-Петербург, Россия
t.skrebtsova@spbu.ru

А. О. Гребенников
Санкт-Петербургский государственный университет;
Санкт-Петербург, Россия
a.grebennikov@spbu.ru

Т. Ю. Шерстинова
Научно-исследовательский университет «Высшая школа экономики»
Санкт-Петербург, Россия
tsherstinova@hse.ru

Аннотация

В работе представлены результаты масштабного проекта, направленного на изучение изменений, произошедших в русском языке в течение первых трех десятилетий XX века. В истории нашей страны этот период был отмечен бурными событиями, которые привели к радикальному изменению государственного строя и построению нового общества. Для количественной оценки масштаба изменений, которые произошли в языке в результате драматических событий рассматриваемого периода, необходим анализ представительного объема языкового материала и сравнение разных хронологических срезов в динамическом аспекте с применением количественных методов. Исследование проведено на материале аннотированной выборки из Корпуса русского рассказа 1900-1930 гг., в которой представлены тексты 300 русских писателей. Материалы Корпуса делятся на три временных среза: 1) довоенный период (1900–1913), 2) военно-революционные годы (1914–1922) и 3) советский период (1923–1930). Проанализировано частотное распределение знаменательной лексики в динамике, что позволило выявить яркие тенденции в изменении частотности отдельных слов и лексических групп от одного исторического отрезка к другому и соотнести их с ранее выявленной динамикой тем. Используемая методика позволяет проследить влияние крупномасштабных политических изменений на словарный состав языка художественной литературы, отметить особенности и тенденции мировосприятия авторов в определённый исторический период, а также дает возможность существенно дополнить анализ динамики тем художественных произведений.

Ключевые слова: лексика; изменение лексического состава в диахронии; язык и стиль художественных текстов; русский рассказ; частотный словарь; корпусная лингвистика; компьютерная лингвистика

1 О Корпусе русских рассказов (1900-1930) и его периодизации

Настоящее исследование является частью масштабного проекта, направленного на изучение изменений, которые происходили в русском языке в первую треть XX века, – возможно, самый драматический период его развития. Проект включает создание электронного текстового корпуса, содержащего тысячи русских рассказов, написанных в первые три десятилетия прошлого века, и их дальнейший комплексный филологический анализ [9; 10; 12].

В истории нашей страны этот период был отмечен бурными событиями, приведшими к радикальному изменению государственного строя и построению нового общества. Цепь исторических событий, охватывающих Первую мировую войну, Февральскую и Октябрьскую революцию и Гражданскую войну, обусловила масштабные языковые и стилистические сдвиги. Огромный пласт «отжившей» лексики сменился новыми словами, отражающими новые понятия и идеи, многие слова «из прошлой эпохи» приобрели новые значения или коннотации, произошла трансформация общепринятых речевых структур (в частности, поменялись функциональные частоты многих лексических единиц, сменился набор привычных коллокаций, появились новые модели сочетаемости, фразеологические обороты и т. д.). Помимо «естественного» процесса резких языковых изменений, неизбежно сопровождающих любой переломный период, следует отметить и сознательные действия новой власти, направленные на изменение языковых норм, с тем чтобы еще более размежеваться с уходящей эпохой и подчинить языковую политику государства решению новых актуальных задач.

Для количественной оценки масштаба изменений, которые произошли в языке в результате драматических событий первой трети XX века, необходим анализ представительного объема языкового материала и сравнение разных хронологических срезов в динамическом аспекте с применением количественных методов. Для этой цели создается Корпус русских рассказов первых трех десятилетий XX века, насчитывающий несколько тысяч единиц. Выбор жанра рассказа для изучения языковых и стилистических изменений обусловлен тем, что он принадлежит к числу наиболее распространенных жанров художественной литературы. Это позволяет охватить тексты максимального числа авторов, писавших в исследуемую эпоху, – не только ведущих, но и множества второстепенных, – что способствует репрезентативности коллекции и достоверности выводов. Другая причина выбора именно этого литературного жанра связана со способностью рассказов, в силу своего небольшого объема и предназначенности (как правило) для пуб-

ликации в периодических изданиях, чутко реагировать на текущие события и улавливать изменения в общественном сознании.

Историческим центром рассматриваемой эпохи, ее переломом, является Октябрьская революция. Все остальные события и процессы рассматриваются или как преддверие центрального события, или как его последствия. Материалы Корпуса делятся на три временных среза: 1) довоенный период: начало XX века до Первой мировой войны (1900–1913), 2) военно-революционные годы: Первая мировая война, Февральская и Октябрьская революция и Гражданская война (1914–1922) и 3) советский период (1923–1930).

Писатель может быть представлен одним рассказом в каждый временной отрезок, причем в Корпус не включаются рассказы, написанные в эмиграции. Так, покинувший Россию в 1920 г. И. А. Бунин представлен одним рассказом за довоенный период и еще одним – за военно-революционный период.

Аннотированная выборка из Корпуса, в которой представлены тексты 300 русских писателей (приблизительно по 100 за каждый период[1], более 1 млн. словоупотреблений) служит своеобразным полигоном для разностороннего изучения материала. Наряду с творчеством признанных мастеров пера – Чеховым, Буниным, Горьким, Куприным, Вересаевым, Булгаковым, Шмелевым, Грином, Тэффи, Замятиным, Зощенко, Катаевым, Пильняком, Кавериным, Гайдаром, Олешей, Бабелем, Платоновым, Пришвиным и др. – она включает произведения авторов, известных лишь узким специалистам. На основе этой выборки проводятся исследования, затрагивающие не только языковые изменения, но и динамику тем, а также композиционную специфику русских рассказов [7; 8; 12–17].

Настоящее исследование также базируется на данном подкорпусе. Оно посвящено анализу наиболее частотных знаменательных лексем, собранных как отдельно по каждому из трех периодов, так и совокупно по всему историческому отрезку длиной в три десятилетия. Следует подчеркнуть, что выделенные периоды принципиально отличны друг от друга: можно сказать, что второй (военный) противопоставлен первому и третьему (мирным), однако первый и третий кардинально различаются между собой по общественно-политическому устройству. Поэтому их попарное сравнение представляет несомненный интерес, в особенности на фоне выявленной ранее динамики тем [14; 15; 17].

В предыдущих исследованиях, выполненных на материале Корпуса русского рассказа, уже было отмечено, что масштабное сопоставление данных, позволяющее выявить лексическое своеобразие эпохи, обусловленное общественно-политической атмосферой и отражающее тенденции в языковом употреблении, возможно только путем сравнения верхних рангов знаменательной лексики с соответствующими рангами писательских словарей и словаря языка в целом, во-первых, и, что гораздо более желательно, с аналогичным электронным корпусом отечественных литературных произведений, относящихся к какому-либо другому историческому периоду, во-вторых.

В частности, в [11] были наглядно представлены результаты такого сопоставления для первого (довоенного) периода рассказов из нашей выборки. Одновременно установлено, что сравнение со словарями отдельных авторов обнаруживает значительную индивидуально-авторскую вариативность выделенных лексем, затрудняющую выделение общих тенденций [1–5]. Сравнение же со словарем языка в целом неизбежно сталкивается с проблемой представленности в нем множества жанров, и, хотя результаты зачастую интересны и показательны, они искажены значительной долей статистического «шума» [6; 7]. Поэтому в данном исследовании было принято решение от него отказаться.

Напротив, было установлено, что сопоставление рассказов 1900–1913 гг. с русскими рассказами начала XXI века, полученными на основе Национального корпуса русского языка (НКРЯ), обладает значительным стилеразличительным потенциалом [6]. Оно дает возможность подтвердить или опровергнуть справедливость сделанных предположений о том, что именно принадлежность к разным историческим периодам прежде всего обусловливает наблюдаемые различия в частотном распределении лексем.

[1] Было выбрано 300 авторов, но из-за того, что некоторые из них представлены более, чем в одном периоде, выборка составляет не 300, а 310 рассказов.

2 Методология исследования

На материале аннотированной выборки из Корпуса русских рассказов были построены расположенные в порядке убывания частот частотные словари как для выборки в целом, так и для каждого из исторических периодов объемом 24 316 лексем, 376 513 словоформ для первого периода; 24 617 лексем, 303 588 словоформ для второго периода; 30 560 лексем; 383 430 словоформ для третьего периода и 124 081 лексема, 1 077 970 словоформ для выборки в целом.

В качестве объекта исследования были выбраны знаменательные слова, расположенные в верхних зонах частотного распределения (с частотой выше 100). Для каждого периода их количество превысило 200, и составило около 800 для выборки в целом.

При сравнении частотного распределения лексики в разные периоды (см. раздел 3) используется такой параметр, как ранг лексемы (иначе говоря, ее статистический вес). Вследствие некоторой разницы в объемах полученных словарей учет именно рангов, а не абсолютных частот является корректным решением.

В тексте статьи для описания динамики частотности отдельных слов используется следующая нотация. В скобках при слове указывается его текущий ранг; если в предшествующий период оно также входило в рассматриваемую верхнюю зону рангового распределения, указывается и его прежний ранг, и два числа соединяются стрелкой. Например, при сравнении второго периода с первым запись *солдат* (116→62) означает, что данное слово имело 116-й ранг в первый период и переместилось на 62-ю позицию во второй период. Если слово в предшествующий период находилось вне верхней зоны, а в рассматриваемый период в нее вошло, первое число отсутствует, ср. *офицер* (→169). Если же, напротив, в более ранний период слово присутствовало в верхней зоне, а в рассматриваемый период ее покинуло, отсутствует второе число, ср. *красивый* (143→). При сопоставлении данных за три периода возможна ситуация, когда некоторое слово присутствует в верхней зоне в первый и третий периоды, но отсутствует во второй – в таком случае запись выглядит следующим образом: *смех* (257→→234).

Несколько слов следует сказать о специфике работы автоматической программы лемматизации[2]. В частности, парные глаголы совершенного и несовершенного вида рассматриваются в качестве отдельных лексем (что представляется методологически правильным в силу частых расхождений в наборе их значений). То же относится к супплетивным формам (так, ниже отдельно фигурируют формы *ребенок* и *дети*). Имена собственные по понятным причинам были исключены из частотных списков. Некоторые из них из-за графического совпадения с именами нарицательными могут влиять на статистику последних, ср. *Вера – вера*. В подобных случаях соответствующие общие имена также не принимались во внимание.

Заметим, что при автоматической обработке неизбежны некоторые погрешности, связанные с неправильным определением леммы (*большой* и *больший*, *стоять* и *стоить*, *пол* и *пола*, *лес* и *леса* и пр.), омонимией (*мир*, *язык*) и грамматической неоднозначностью (*стать*). Во избежание ошибочных заключений подобные случаи также исключены из рассмотрения.

При интерпретации полученных данных полисемичные слова рассматриваются в совокупности всех своих лексико-семантических вариантов, что, разумеется, ведет к некоторым погрешностям, но является неизбежным следствием применения автоматических методов. Представляется, что эта погрешность невелика при рассмотрении небольших и компактно расположенных исторических отрезков.

3 Динамика частотных лексем по периодам

Примечательно, что шесть самых верхних рангов во всех периодах занимают одни и те же слова (различаясь лишь позициями): *говорить*, *сказать*, *один*, *глаз*, *рука*, *мочь*. Остальной материал демонстрирует как сходства, так и существенные различия. Посредством сравнения более позднего периода с более ранним(и) мы стремимся выделить: 1) новые слова, вошедшие в верхнюю зону частотного распределения; 2) слова, которые ушли из нее и 3) слова, демонстрирую-

[2] Словари строились при помощи программы "UNILEX" (разработка Институт русского языка им. Виноградова). См. Аношкина Ж.Г. Тексто-ориентированная компонента АЛС УНИЛЕКС (УНИЛЕКС-Т) // Альманах «Говор», № 5, 1995, стр. 7-29.

щие выраженную динамику частоты. Мы анализируем эти явления, пытаясь связать их с общественно-политической обстановкой соответствующего времени и выявленной ранее динамикой тем [7–9].

Сравнение второго (военно-революционного) периода с первым (довоенным)

На фоне довоенных рассказов в произведениях второго периода закономерно увеличивается доля военной тематики [17, с. 50] и становится актуальной национальная самоидентификация, что соответственно проявляется во вхождении в верхнюю зону частотного распределения слов *офицер* (→169) и *русский* (→182). Военным временем, по-видимому, объясняется и вхождение слов *дьякон* (→83) и *писать* (→181) – в связи с возросшей ролью церкви, вынужденным разделением семей, тревогой за близких и потребностью в переписке. Указанные корреляции подкрепляются динамикой роста у слов *бог* (89→65), *солдат* (116→62) и *письмо* (190→117), которые были в верхней зоне и в первый период, но переместились на более высокие позиции рангового распределения.

Еще одним индикатором смены эпохи можно считать уход слова *можно* (80→), с одной стороны, и появление в верхней зоне слов *должный* (→76) и *нельзя* (→147) – с другой, что сигнализирует о наступлении более сурового и жесткого времени, связанного с ограничениями, запретами и принуждениями (глагол *мочь*, впрочем, сохраняет свой шестой ранг). В целом, можно сказать, что модальность долженствования вытесняет модальность возможности. В такой период важной оказывается концентрация на текущем моменте – отсюда вхождение слов *сегодня* (→172), *сейчас* (→71), *теперь* (→20), *наконец* (→186).

Вполне закономерным выглядит уход из верхней зоны подавляющего большинства слов, обозначающих положительные эмоции, а именно: *праздник* (268→), *добрый* (265→), *светлый* (236 →), *красивый* (143→), *веселый* (152→), *весело* (182→), *смех* (257→), *счастье* (228→), *чувство* (126→), *улыбка* (219→), *улыбаться* (173→), *тихий* (128→), *тишина* (249→). (Некоторые из них, а именно *веселый*, *смех*, *тихий* и *тишина*, возвращаются в третий период.) Этот факт коррелирует с уменьшением значимости широкого круга тем, связанных с любовью, семьей, помощью ближнему, благотворительностью [7, с. 54, 56]. В связи с этим обращает на себя внимание уход слов *ребенок* (213→), *играть* (227→) и снижение частоты слова *дети* (69→107), которое все-таки удерживается в верхней зоне.

Из прочих слов, покинувших верхнюю зону, отметим *пить* (267→) и *пьяный* (241→) – этот факт можно объяснить введением сухого закона и соответственным снижением темы пьянства, которая, очевидно, повышает частоту данных слов [7, с. 53]. Уход слов *барин* (258→), *хозяин* (226→), *работать* (218→), *рабочий* (202→) связан, по-видимому, с актуализацией темы войны, вытесняющей мирный труд. По аналогичной причине из верхних рангов пропало слово *студент* (217→).

Заметим, что большинство «пропавших» слов так и не вернулось в верхние ранги частоты в третий, советский, период.

Справедливости ради следует упомянуть и то, что можно назвать контрпримерами, а именно слова, сохранение или исчезновение которых не получается объяснить выявленной ранее динамикой тем. К примеру, в верхней зоне частотного распределения осталось слово *смеяться*, хотя ранг его и понизился (131→183). Напротив, ушли слова *страшный* (135→), *страх* (253→), *ужас* (166→), *дрожать* (185→), *умереть* (191→), *тоска* (207→), *больной* (210→), которые, казалось бы, гораздо более востребованы в эпоху войн и революций, чем в мирное время. Несмотря на заметный и вполне объяснимый рост «мистических» тем, связанных с видениями, предчувствиями, снами, мечтами [7, с. 56], верхнюю зону покинули слова *показаться* (194→), *похожий* (174→), *странный* (155→).

Сравнение третьего (советского) периода с предшествующими

Характерной особенностью третьего периода является вхождение в верхнюю зону частотного распределения большого числа конкретных существительных, связанных с сельской жизнью: *дед* (→114), *старуха* (→166), *ребята* (→183), *хлеб* (→152), *поле* (→238), *куст* (→255), *трава*

(→245), *собака* (→165), *конь* (→204), *птица* (→230) – и техническим прогрессом: *машина* (→250), *поезд* (→256), *вагон* (→173), *ход* (→177). Это коррелирует с выраженным ростом соответствующих тем [17, с. 51-52]. Число абстрактных существительных, напротив, сокращается.

Список частей тела человека, и так широко представленных в верхней зоне частотного распределения, в советский период увеличивается чуть ли не вдвое. Так, к уже имеющимся единицам *рука*, *глаз*, *голова*, *лицо*, *губа*, *зуб*, *нога*, *тело*, *плечо*, *палец*, *волос* добавились слова *нос* (→121), *ухо* (→184), *лоб* (→216), *шея* (→275), *щека* (→264), *борода* (→268), *бок* (→252), *колено* (→205).

Расширился также набор числительных – к *один*, *два*, *три*, *первый* добавились лексемы *четыре* (→262), *пять* (→218), *второй* (→258). На фоне общего уменьшения числа прилагательных возвращается слово *синий* и появляются такие цветообозначения, как *желтый* (→221) и *зеленый* (→222). (*Белый* и *черный* стабильно занимают высокие позиции, а по поводу *красный* см. ниже).

Появилось слово *вперед* (→170), что, по-видимому, объясняется актуализированными темами технического прогресса и светлого будущего [17, с. 51]. Политика ликвидации безграмотности и культпросвета обусловила вхождение в верхнюю зону слова *книга* (→208).

Вновь появилось пропавшее во второй период слово *можно* (80→→95). *Нельзя* (→147→) ушло, но *должный* (→76→67) сохранилось. Слово *хотеться* (101→150→227) демонстрирует последовательное снижение ранга, при том что *хотеть* (19→23→23) на протяжении всех трех периодов занимает примерно одинаковое высокое положение в частотном распределении.

Показателем наступившего мирного времени можно считать возвращение в верхнюю зону частотного распределения слов *работать* (218→→109), *рабочий* (202→→112), *веселый* (152→→232), *смех* (257→→234), *тихий* (128→→196), *тишина* (249→→231), *играть* (227→→207), *разговор* (186→→209).

Обращает на себя внимание вхождение в верхнюю зону слов *ружье* (→278), *рота* (→218) и *кровь* (→105). На первый взгляд, более естественным казалось бы их появление в предыдущий, военно-революционный, период. Возможное объяснение заключается в том, что в нашей выборке за третий период рассказов про Гражданскую войну оказалось примерно в два раза больше, чем за второй, т. е. тогда, когда эта война шла. Это «отставание» литературы от жизни обусловлено рядом факторов. Из наиболее очевидных упомянем тот банальный факт, что на написание рассказа требуется время, затем проходит еще какое-то время до его публикации, которая сама по себе затруднена в условиях затянувшейся войны, политических волнений и экономической разрухи. К тому же, для осознания столь масштабных событий, приведших к радикальному изменению общественной жизни, необходима дистанция («большое видится на расстоянии»). Отсюда своеобразный отсроченный эффект: чем крупнее историческое событие, тем дольше оно сохраняет свою значимость, в том числе в литературе и искусстве.

Вполне закономерно верхнюю зону частотного распределения покинули слова *офицер* (→169→) (в Красной Армии воинские звания были упразднены), *солдат* (116→62→) и *дьякон* (→83→). Слово *бог* (89→65→211) осталось, но его ранг заметно снизился, что обусловлено антирелигиозной политикой советской власти.

Примечателен уход из верхней зоны слов *гость* (165→) и *знакомый* (193→), присутствовавших там и в первый, и во второй периоды. Социальные отношения теперь сводятся в основном к семейным и трудовым, причем семейные связи угасают: у слов *муж* (146→166→236), *жена* (52→84→139), *дети* (69→107→198) наблюдается снижение ранга, а слово *ребенок* так и не вернулось в верхнюю зону.

В числе прочих слов, присутствовавших в верхней зоне в оба предшествующих периода, но ушедших в советское время, отметим *господин* (145→146→) (дореволюционная тематика практически сошла на нет), *толпа* (79→103→), *милый* (122→176→), *любовь* (134→112→), *письмо* (190→117→).

Общая динамика частот на протяжении трех периодов

В этом разделе мы сосредоточимся на словах, присутствующих в верхней зоне частотного распределения на протяжении всех трех периодов. Прежде всего нас интересуют случаи ярко выраженного последовательного изменения ранга. Они разделяются на 1) случаи поступательного роста ранга и 2) случаи поступательного снижения ранга. Эти примеры, как мы предполагаем, обусловлены внешним контекстом – коренным переворотом политического строя, сломом прежних социальных отношений и построением нового общества.

Возможно, наиболее ярким примером первого типа может служить заметная активизация слова *товарищ* (198→105→38) как индикатора новой советской власти. Здесь же уместно упомянуть схожую динамику слова *красный* (163→110→54). Помимо этого, наблюдается существенное повышение ранга у слов, которые можно прямо или опосредованно связать с жизнью на селе, ср. *деревня* (250→192→150), *народ* (224→198→164), *мужик* (252→144→61), *баба* (→143→81), *сын* (183→161→126), *брат* (255→136→120), *утро* (100→81→55), *лошадь* (176→139→93), *бежать* (169→134→85), *дорога* (209→168→134), *ветер* (216→149→113).

Примеры второго типа включают слова, обозначающие внутреннюю жизнь человека, ср. *любить* (39→43→69), *чувствовать* (54→152→197), *смеяться* (131→183→189), *душа* (42→35→151), *мысль* (62→102→130), а также его связи с близкими, ср. *жена* (52→84→139), *муж* (146→166→236), *дети* (69→107→198). Эта же динамика характерна для слов *молодой* (58→59→147), *казаться* (14→51→58) и *смерть* (106→163→223).

Последовательное уменьшение ранга зафиксировано у слова *деньги* (133→178→249), что обусловлено снижением покупательной способности денег в военно-революционный период, эмиссией разнообразных бумажных знаков, имевших сомнительную ценность, а затем денежными реформами советской власти (деноминациями). Заметим, что соответственно снижается частота темы, связанной с оппозицией богатства и бедности: после Гражданской войны в Советской России просто не осталось богатых людей [17, с. 53].

Кроме случаев поступательного изменения ранга (будь то рост или падение) имеются слова, у которых динамика употребления может быть представлена в виде ломаной линии. Иными словами, они имеют точку перелома во втором периоде, а показатели первого и третьего периодов достаточно схожи. Однако их перечень, как нам кажется, не дает основания для каких-либо обобщений и корреляций с исторической ситуацией. В связи с этим мы опускаем данные о динамике рангов, ограничиваясь простым перечислением. Так, возрастание ранга во второй период фиксируется у слов *живой*, *читать*, *легкий*, *просить*, *высокий*, *поднять*, *бояться*, *ждать*, *дать*. Напротив, падение ранга во второй период наблюдается у слов *девушка*, *длинный*, *тяжелый*, *угол*, *воздух*, *подумать*, *свет*.

Достаточно большое число слов вообще не имеют значительных колебаний ранга, образуя своеобразные «инварианты». Они расположены преимущественно в пределах верхних 80 рангов – ниже разброс, как правило, довольно велик. (Хотя и в этих рамках иногда случаются резкие колебания: достаточно указать на слова *красный* и *товарищ*, см. выше.)

К словам, стабильно характеризующимся высокой частотой, относятся основные глаголы движения (*идти*, *ходить*, *выйти*, *пойти*, *уйти*), позы (*сидеть*, *стоять*, *лежать*), речи (*говорить*, *сказать*, *спросить*, *молчать*), чувственного восприятия (*видеть*, *смотреть*), а также глаголы *жить*, *взять*, *спать*, *хотеть*, *любить*, *казаться*. Из высокочастотных прилагательных упомянем *последний*, *большой*, *маленький*, *старый*, *новый*, *белый*, *черный* и примыкающие к ним местоимения *другой* и *каждый*. Существительные верхних рангов включают такие группы слов, как *город – улица – дом*, *комната – стол*, *человек – люди*, *время – год – час*, названия частей тела (*голова*, *лицо*, *глаз*, *нога*, *рука*), времен суток (*день*, *ночь*, *утро*, *вечер*), а также лексемы *жизнь*, *место*, *сторона*, *земля*, *дело*, *отец*, *сердце*.

4 Заключение

В настоящей статье проанализировано частотное распределение знаменательной лексики на материале выборки из Корпуса русских рассказов (1900-1930). Внимательное рассмотрение верхней зоны частотного распределения позволило выявить яркие тенденции в изменении ча-

стотности отдельных слов и лексических групп от одного исторического отрезка к другому и соотнести их с ранее выявленной динамикой тем.

Среди возможных направлений дальнейшего анализа наиболее интересным и перспективным представляется сопоставление полученных данных с частотностью лексических единиц в русских рассказах начала XXI века. Оно позволяет в более широком ракурсе оценить языковые и стилистические изменения в языке литературных произведений и задуматься об их причинах. Мы рассматриваем это в качестве самостоятельного направления дальнейших исследований.

В целом, анализ наших данных показывает, что частотные распределения, построенные на материале представительных выборок из масштабного корпуса текстов, могут служить хорошим индикатором динамики лексического состава художественной прозы произведений отдельной эпохи. Используемая методика позволяет проследить влияние крупномасштабных политических изменений на словарный став языка художественной литературы, отметить особенности и тенденции мировосприятия авторов в определённый исторический период, а также позволяет существенно дополнить анализ динамики тем произведений.

Acknowledgements

Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 17-29-09173 «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

References

[1] Grebennikov A.O. (2019), Measures of lexical similarity of frequency dictionaries [Mery leksicheskogo skhodstva chastotnykh slovarey], Structural and Applied Linguistics [Strukturnaya i Prikladnaya Lingvistika], No. 12, Saint Petersburg, pp. 61–68.

[2] Grebennikov A.O., Martynenko G.Ya. (Ed.) (1999), Frequency Dictionary of the Short Stories by Anton P. Chekhov [Chastotnyy slovar rasskazov A.P. Chekhova], Izdatelstvo Sankt-Peterburgskogo Universiteta, Saint Petersburg.

[3] Grebennikov A.O., Martynenko G.Ya. (Ed.) (2003), Frequency Dictionary of the Short Stories by Leonid N. Andreev [Chastotnyy slovar rasskazov L.N. Andreeva], Izdatelstvo Sankt-Peterburgskogo Universiteta, Saint Petersburg.

[4] Grebennikov A.O., Martynenko G.Ya. (Ed.) (2006), Frequency Dictionary of the Short Stories by Alexander I. Kuprin [Chastotnyy slovar rasskazov A.I. Kuprina], Izdatelstvo Sankt-Peterburgskogo universiteta, Saint Petersburg.

[5] Grebennikov A.O., Martynenko G.Ya. (Ed.) (2011), Frequency Dictionary of the Short Stories by Ivan A. Bunin [Chastotnyy slovar rasskazov A.I. Bunina]. Izdatelstvo Sankt-Peterburgskogo universiteta, Saint Petersburg.

[6] Grebennikov A.O., Marusenko N.M. (2020), Corpus of the Russian story of the early XX century. An example of linguistic statistical analysis [Korpus russkogo rasskaza nachala XX veka. Primer lingvostatisticheskogo analiza], Computational Linguistics and Computational Ontologies: Proceedings of the XXIII Joint Conference "Internet and Modern Society" (IMS–2020) [Komp'yuternaya Lingvistika i Vychislitel'nye Ontologii. Trudy XXIII Mezhdunarodnoj Ob"edinennoj Konferencii «Internet i Sovremennoe Obshchestvo» (IMS–2020)], Saint Petersburg, pp. 21–29.

[7] Grebennikov A.O., Skrebtsova T.G. (2019), Yazykovaya kartina mira v russkom rasskaze nachala XX veka [Linguistic picture of the world in the Russian story of the early XX century], Philosophy and Humanities in the Information Society [Filosofija i gumanitarnyie nauki v informatsionnom obschestve], No. 3, pp. 82–92.

[8] Martynenko G., Sherstinova T. (2018), Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture, Digital Transformation and Global Society. Communications in Computer and Information Science, Vol. 859, Saint Petersburg, pp. 299–309.

[9] Martynenko G.Ya., Sherstinova T.Yu, Popova T.I., Melnik A.G. (2018), Methodological Issues Related with the Compilationof DigitalAnthology of Russian Short Stories (the First Third of the 20th Century) [Metodologicheskie problemy sozdaniya Komp'yuternoj Antologii Russkogo Rasskaza kak Yazykovogo Resursa Dlya Issledovaniya Yazyka i Stilya Russkoj Khudozhestvennoj Prozy v Ehpokhu Revolyucionnykh Peremen (Pervoj Treti XX veka)], Computational Linguistics and Computational Ontologies: Proceedings of the

XXI Joint Conference "Internet and Modern Society" (IMS–2018) [Komp'yuternaya Lingvistika i Vychislitel'nye Ontologii. Trudy XXI Mezhdunarodnoj Ob"edinennoj Konferencii «Internet i Sovremennoe Obshchestvo» (IMS–2018)], Saint Petersburg, pp. 99–104. Access mode: https://openbooks.itmo.ru/ru/file/8421/8421.pdf

[10] Martynenko G.Ya., Sherstinova T.Yu., Popova T.I., Melnik A.G., Zamirajlova E.V. (2018), On the principles of creation of the Russian short stories corpus of the first third of the 20th century [O printsipakh sozdaniya korpusa russkogo rasskaza pervoy treti XX veka], Proceedings of the XV International Conference on Computer and Cognitive Linguistics “TEL 2018”, Kazan, pp. 180–197.

[11] Sherstinova T., Grebennikov A., Skrebtsova T., Guseva A., Gukasian M., Egoshina I., Turygina M. (2020), Frequency word lists and their variability (the case of Russian fiction in 1900-1930), Proceedings of the 27th Conference of FRUCT Association, Helsinki, № 27, pp. 366–373.

[12] Sherstinova T., Martynenko G. (2020) Linguistic and stylistic parameters for the study of literary language in the Corpus of Russian short stories of the first third of the 20th century. R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia. CEUR Workshop Proceedings, Vol. 2552, pp. 105–120. Access mode: http://ceur-ws.org/Vol-2552/.

[13] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. (2020), Topic modelling with NMF vs. expert topic annotation: the case study of Russian fiction, Advances in Computational Intelligence. MICAI 2020. Lecture Notes in Computer Science, vol. 12469, Springer, Cham, pp. 134–151. Access mode: https://doi.org/10.1007/978-3-030-60887-3_13.

[14] Sherstinova T., Skrebtsova T. (2020), Russian literature around the October revolution: A quantitative exploratory study of literary themes and narrative structure in Russian short stories of 1900-1930, Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, pp. 117–128. Access mode: http://ceur-ws.org/Vol-2813/rpaper09.pdf

[15] Skrebtsova T. (2020) Thematic tagging of literary fiction: the case of early 20th century Russian short stories, Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, pp. 265–276. Access mode: http://ceur-ws.org/Vol-2813/rpaper20.pdf

[16] Skrebtsova T.G. (2019), The structure of the narrative in the Russian story of the early XX century [Struktura narrativa v russkom rasskaze nachala XX veka], Proceedings of the International Conference “Corpus linguistics–2019” [Trudy Mezhdunarodnoy Konferentsii “Korpusnaya Lingvistika–2019”], St. Petersburg, pp. 426–431.

[17] Skrebtsova T.G. (2020), Dynamics of themes of Russian stories of the early XX century [Dinamika tem russkikh rasskazov nachala XX veka], Philosophy and Humanities in the Information Society [Filosofija i gumanitarnyie nauki v informatsionnom obschestve], No. 3, pp. 45–60.

Приложение. Ранги знаменательной лексики по частоте употребления в каждом из трех исторических периодов, верхняя зона частот (250 слов).

СЛОВО	1-й ПЕРИОД (1900-1913)	2-й ПЕРИОД (1914-1922)	3-й ПЕРИОД (1923-1930)
	РАНГ	РАНГ	РАНГ
сказать	1	2	3
один	2	3	4
глаз	3	4	2
говорить	4	1	5
рука	5	5	1
мочь	6	6	6
лицо	7	9	12
знать	8	7	8
другой	9	12	13
голова	10	10	7
идти	11	8	9
жизнь	12	19	33
человек	13	14	11
казаться	14	51	58
думать	15	22	20
люди	16	15	29
время	17	37	24
голос	18	26	30
хотеть	19	23	23
видеть	20	30	22
дом	21	32	37
большой	22	16	17
смотреть	23	38	31
два	24	17	14
ночь	25	21	26
раз	26	34	28
дело	27	27	18
сидеть	28	29	21
слово	29	28	40
пойти	30	31	19
нога	31	25	15
земля	32	40	27
друг	33	42	88
комната	34	56	56
окно	35	45	42
дверь	36	33	25
есть	37	36	46
белый	38	39	32
любить	39	43	69
стоять	40	50	44
черный	41	46	36
душа	42	35	151
первый	43	60	43
темный	44	118	101
спать	45	58	53
спросить	46	49	57
сердце	47	54	77
час	48	70	71

место	49	44	34
молчать	50	85	59
маленький	51	48	65
жена	52	84	139
новый	53	61	66
чувствовать	54	152	197
жить	55	53	64
минута	56	101	146
сторона	57	78	35
молодой	58	59	147
старый	59	77	60
уйти	60	94	62
выйти	61	57	45
мысль	62	102	130
стена	63	126	80
сделать	64	108	117
стол	65	64	51
взять	66	52	39
делать	67	100	87
отец	68	55	75
дети	69	107	198
леса	70	47	73
каждый	71	90	79
город	72	79	47
ходить	73	73	76
улица	74	74	70
свет	75	148	115
лежать	76	72	72
нужный	77	68	49
вечер	78	69	48
толпа	79	103	
можно	80		95
последний	81	86	83
тяжелый	82	157	143
вода	83	123	52
небо	84	92	116
слышать	85	116	129
стоить	86	104	92
женщина	87	67	91
понять	88	125	100
бог	89	65	211
понимать	90	135	127
почтить	91	142	156
сила	92	128	90
лета	93	106	145
ответить	94	122	111
самый	95	120	86
три	96	98	68
мать	97	113	102
год	98	80	74
воздух	99	164	137
утро	100	81	55
хотеться	101	150	227
глядеть	102	82	123
губа	103	115	82

тело	104	133	84
больший	105	99	63
смерть	106	163	223
конец	107	96	89
кричать	108	129	94
начало	109	145	144
ряд	110	119	106
слушать	111	93	118
длинный	112	188	158
угол	113	167	110
вера	114		
грудь	115	124	103
солдат	116	62	
ждать	117	87	142
подумать	118	170	122
далекий	119	121	125
остаться	120	131	108
взгляд	121	153	213
милый	122	176	
увидеть	123	91	107
солнце	124	97	78
остановиться	125	155	133
чувство	126		
волос	127	173	155
тихий	128		196
дать	129	63	98
плечо	130	109	96
смеяться	131	183	189
шаг	132	158	179
деньга	133	178	249
любовь	134	112	
страшный	135		241
прийти	136	89	157
дорогой	137	132	97
подойти	138	130	128
холодный	139		246
слеза	140	162	182
бояться	141	95	149
девушка	142	191	171
красивый	143		
целый	144	154	200
господин	145	146	
муж	146	166	236
отвечать	147		
хороший	148	127	172
работа	149	114	50
высокий	150	138	186
пройти	151	199	176
веселый	152		232
становиться	153		
посмотреть	154	177	160
странный	155		
войти	156		185
встать	157	187	178
давать	158	174	124

звук	159		
сильный	160		
сон	161	184	175
иногда	162		
красный	163	110	54
огромный	164		201
село	165	171	104
ужас	166		
крикнуть	167		180
широкий	168		168
бежать	169	134	85
вид	170		254
продолжать	171		
плакать	172	156	237
улыбаться	173		
похожий	174		
уходить	175	151	154
лошадь	176	139	93
огонь	177		135
заметить	178		190
знакомый	179	193	
иметь	180		167
полный	181		169
весело	182		
сын	183	161	126
выходить	184	185	243
дрожать	185		276
разговор	186		209
найти	187	137	132
спрашивать	188	189	265
тонкий	189		229
письмо	190	117	
умереть	191		253
палец	192	159	99
серый	193	194	119
показаться	194		
снег	195		140
бледный	196		
поднять	197		181
товарищ	198	105	38
черта	199		192
забыть	200		
мир	201		174
рабочий	202		112
синий	203		136
стараться	204		
двор	205	180	153
начинать	206		273
тоска	207		
помнить	208		
дорога	209	168	134
больной	210		
сталь	211		161
живой	212	175	199
ребенок	213		

сень	214	140	
бросить	215		141
ветер	216	149	113
студент	217		
работать	218		109
улыбка	219		
батюшка	220		
близкий	221		
радость	222	196	191
движение	223		
народ	224	198	164
дерево	225		225
хозяин	226		274
играть	227		207
счастье	228		
ехать	229		159
почувствовать	230		
чужой	231		272
оставаться	232		
оставить	233		
поднять	234	141	
придти	235		
светлый	236		
гореть	237		
десять	238		
доктор	239		233
река	240		224
пьяный	241		
яркий	242		
вернуться	243		187
рубль	244		
крик	245		
легкий	246	195	226
сестра	247		
тень	248		
тишина	249		231
деревня	250	192	150

On Slavic cognate recognition in context

Irina Stenger
Saarland University / Campus C5.3,
66123 Saarbrücken, Germany
ira.stenger@mx.uni-saarland.de

Tania Avgustinova
Saarland University / Campus C7.2,
66123 Saarbrücken, Germany
avgustinova@coli.uni-saarland.de

Abstract

This study contributes to a better understanding of reading intercomprehension as manifested in the intelligibility of East and South Slavic languages to Russian native speakers in contextualized cognate recognition experiments using Belarusian, Ukrainian, and Bulgarian stimuli. While the results mostly confirm the expected mutual intelligibility effects, we also register apparent processing difficulties in some of the cases. In search of an explanation, we examine the correlation of the experimentally obtained intercomprehension scores with various linguistic factors, which contribute to cognate intelligibility in a context, considering common predictors of intercomprehension associated with (i) morphology and orthography, (ii) lexis, and (iii) syntax.

Keywords: Slavic intercomprehension, online experiments, cognate recognition, linguistic context
DOI: 10.28995/2075-7182-2021-20-660-668

О распознавании славянских слов-когнатов в контексте

Ирина Штенгер
Университет земли Саар / Кампус С5.3, 66123 Саарбрюккен, Германия
ira.stenger@mx.uni-saarland.de

Таня Августинова
Университет земли Саар / Кампус С7.2, 66123 Саарбрюккен, Германия
avgustinova@coli.uni-saarland.de

Аннотация

Данное исследование способствует лучшему пониманию межъязыковой понятности славянских языков в ситуации, когда читателю необходимо извлечь информацию из текста при чтении незнакомого, но (близко)родственного языка. В частности, мы изучаем понятность восточнославянских и южнославянских языков носителями русского языка. В настоящей статье основное внимание уделяется вопросам спонтанного понимания белорусских, украинских и болгарских слов-когнатов в контексте. Результаты, полученные в ходе проведенных онлайн-экспериментов, свидетельствуют в целом о высокой степени распознавания славянских стимулов носителями русского языка, а также о некоторых трудностях в процессе понимания. В поисках объяснения полученных результатов мы рассматриваем лингвистические факторы, которые могут предсказать степень понятности славянских слов-когнатов в контексте. Результаты онлайн-экспериментов сравниваются со следующими потенциальными параметрами, оценивающими понятность незнакомого, но (близко)родственного языка при чтении: (i) морфологические и орфографические факторы, (ii) лексические факторы и (iii) синтаксические факторы.

Ключевые слова: славянская межъязыковая понятность, онлайн-эксперименты, распознавание когнатов, лингвистический контекст

1 Introduction

Multiple studies have evidenced that reading is a complex and structured process that involves not only familiarity with linguistic elements, but also the entire knowledge of the reader [Frost 2012], [Lutjeharms 2004]. In an intercomprehension scenario, readers have the potential to understand a message encoded in an unknown (stimulus) language if they are speakers of or have in their linguistic repertoire some genetically related language(s). The previous research on reading intercomprehension indicates a remarkably good performance of Russian speaking subjects in spontaneous out-of-context guessing of isolated cognate words from other Cyrillic-script Slavic languages [Stenger 2019]. Context-free cognate recognition is, of course, quite far from the real-world situation, inasmuch as inferences based on contextual assumptions represent a central technique in intercomprehension. On the other hand, for a context to be useful it needs to be understandable, too. We shall focus on the contextualized cross-lingual cognate recognition, looking for predictors of the human performance in intelligibility tests.

The assumption that cognate words across genetically related languages are better recognizable in a context than as isolated items is quite intuitive and may sound trivial. To model and predict this phenomenon, however, we need to differentiate the linguistic factors that facilitate in-context recognition from those that do not. The shape of cognate words may change in the course of time to the extent of being no longer transparent to a reader with no historical-linguistic background. This poses the question of morphological and orthographic similarities regarding their potential to predict and explain the results of intelligibility tests. On the other hand, lexical and syntactic (dis)similarities appear to influence more directly the use of context in intercomprehension. We shall consider all these linguistic factors to find out to what extent a particular context facilitates the understanding of target words, and why this is not always the case.

After presenting the experimental setup and results, we consider potential predictors of cross-lingual contextualized cognate intelligibility, analyze the influence of individual linguistic factors on human performance, and draw conclusions.

2 Intercomprehension experiment

The stimuli material represents a collection of parallel sentences from the parallel text corpus of the Russian National Corpus[1], which contains the words we are interested in here and which we have previously tested as stimuli[2] for online single-word free translation tasks. In such a way, we can obtain intercomprehension scores for native speakers of Russian (RU). To reveal the inherent intercomprehension, we include in the analysis only people who do not know the stimulus language. To avoid possible learning effects, we consider the results of the participant's initial experiment, and exclude subjects[3] who have already completed other experiments at the intercomprehension website.[4] Thus, the number of participants is 87, aged between 17 and 49 years (i.e. average age 22) with 79 women and 8 men. After completing a background questionnaire, they were introduced to 60 Belarusian (BE), 60 Ukrainian (UK) and 120 Bulgarian (BG) sentences[5] in 16 sessions (15 sentences per session). The number of subjects for each stimulus sentence ranges from 35 to 46 (i.e. in average per session: 40 participants). In a session on contextualized cognate recognition, the subjects see a randomized stimulus sentence on their screen (see Figure 1) and have 10 seconds to translate the marked target word.[6]

[1] https://ruscorpora.ru/new/

[2] The stimuli items come from parallel lists consisting of internationalisms, Pan-Slavic vocabulary, and cognates from Slavic Swadesh lists – for more details cf. [Stenger 2019].

[3] In this case, 4 out of 91 participants were excluded from the analysis.

[4] For our online experiments, we use the INCOMSLAV platform. 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 (https://intercomprehension.coli.uni-saarland.de/en/).

[5] The mean length of BE sentences: 9 tokens, UK sentences: 8 tokens, BG sentences: 8 tokens.

[6] In addition, 3 seconds per item in a sentence are given for reading the whole stimulus.

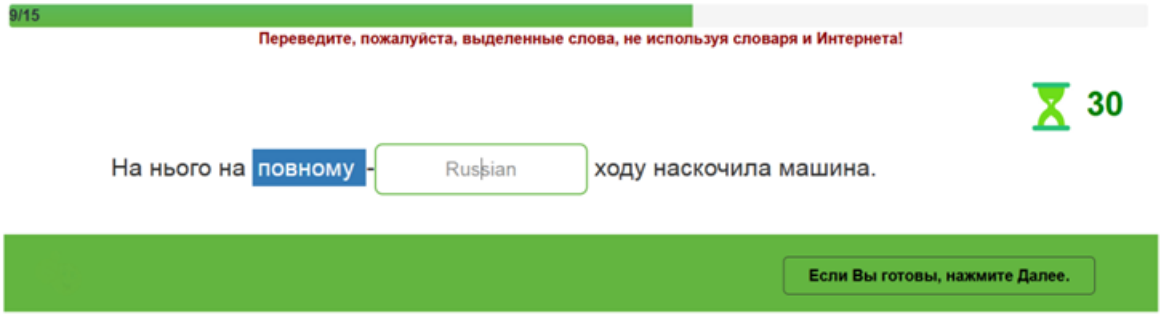


Figure 1: Contextualized cognate translation from UK into RU of 'He was run over by a car at full speed'. Instruction on top: "Please translate the marked words without a dictionary or the internet!"

The target words occupy different sentence positions equally distributed in the respective stimuli sets: BE – 22% initial, 40% middle, 38% end; UK – 20% initial, 47% middle, 33% end; BG – 23% initial, 41% middle, 36% end. After the first session (25 subjects started with UK stimuli, 26 with BE, and 36 with BG), each participant may continue the experiment by completing the task for the remaining stimulus languages offered in a random order. Their responses are categorized automatically via pattern matching with pre-defined correct answers and acceptable alternatives, and checked manually in the final analysis. The mean percentage of successfully translated items constitutes the achieved intercomprehension score for a given constellation (Table 1). The overall cognate intelligibility order attested here for the in-context condition resembles what [Stenger 2019] has observed for the out-of-context condition: UK > BE > BG. While the differences in the comprehension of UK and BG are in favor of the out-of-context condition, the role of context appears more prominent for BE.

stimuli–subject	in-context (this study)	out-of-context [Stenger 2019]
BE–RU group	81.32%	72.56%
UK–RU group	84.93%	85.61%
BG–RU group	69.78%	71.33%

Table 1: Cross-lingual intelligibility of cognates for RU speaking subjects

Regarding the number of word pairs with higher intercomprehension scores in the respective condition, observe that more cognates are successfully translated in-context than out-of-context, e.g., BE 48 > 11 (1)[7], UK 25 > 23 (12), BG 70 > 46 (4). The results are correspondingly visualized as a percentage of correctly produced translations for BE (Figure 2), UK (Figure 3), and BG (Figure 4). In these representations, cognate pairs better recognized in-context are presented to the left, those more successfully guessed out-of-context – to the right, and stimuli yielding identical scores for both conditions – in the middle.

As human performance depends not only on context (un)availability, but also on the individual forms, here are some examples. The UK stimulus *дім* (dim) 'house' is correctly understood to 51% in-context and only to 11% without context. However, the UK stimulus *літо* (lito) 'summer' is better understandable to RU subjects as a single word (95%) rather than in context: *влітку* (vlitku) 'in summer' (17%). The BE stimulus *дзіця* (dzicja) 'child' is to 95% correctly translated in-context and only to 35% without context. Yet, the BE stimulus *бераг* (berah) 'bank' is to 71% successfully guessed as a single word in comparison to its only 8% recognition in context: *на бераze* (na beraze) 'on the bank'. The BG stimulus *пет* (pet) 'five' is much better understood with context than without (76% intelligibly vs. 15%), while *славей* (slavej) 'nightingale' causes less difficulties as a single word (90%) than in context (5%) *славеят* (slavejat) 'the nightingale'.

[7] The number of cognate pairs yielding identical scores for both conditions is given in brackets.

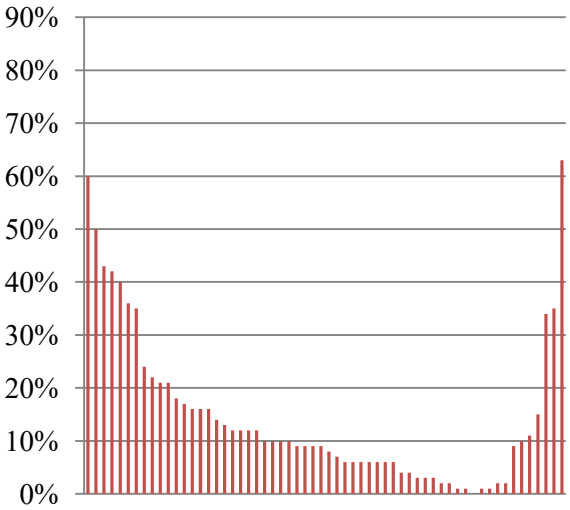


Figure 2: Number of successfully guessed cognates in context (left) vs. without context (right): BE

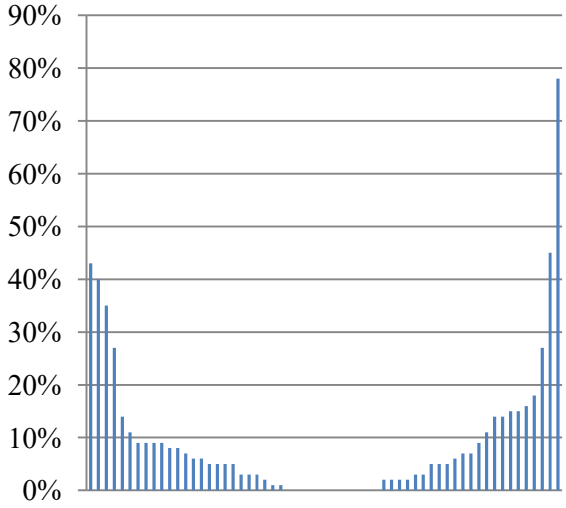


Figure 3: Number of successfully guessed cognates in context (left) vs. without context (right): UK

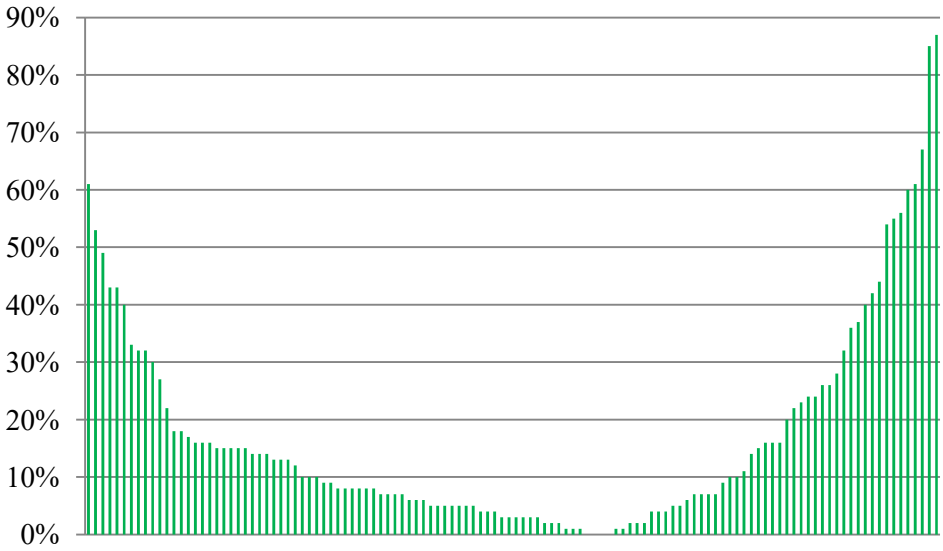


Figure 4: Number of successfully guessed cognates in context (left) vs. without context (right): BG

3 Measuring contextualized cognate intelligibility

Morphological and orthographic factors: A string similarity measure [Levenshtein 1965] known as Levenshtein distance (LD) approximates synchronically observable morphological and orthographic characteristics of cross-linguistic correspondences. We perform objective LD computations between word-forms automatically, using the *incom.py* tool [Mosbach et al. 2019]. In a previous study [Stenger and Avgustinova 2020] have successfully used the normalized LD to measure orthographic and phonetic distances between related languages. Here, we assume that the larger the distance between a stimulus and its correspondence in subjects' native language is, the more difficult intercomprehension would be. Figure 5 shows a sample LD calculation between contextualized word-forms for 'head' (in 'head downwards') – cf. the BG stimulus *с главата надолу* (s glavata nadolu,) and its corresponding RU translation *вниз головой* (vniz golovoj). To normalize, the sum of all costs for the character-by-character transformation of one string into another (1+1+1+1+1=5) we divide the number of alignment slots (8), which results in a normalized LD (nLD) of 0.63 or 63%. Applying this method, we obtain the mean nLD (%) between stimulus word-forms and their target cognates in the subjects' language (Table 2).

	1	2	3	4	5	6	7	8
BG	г		л	а	в	а	т	а
RU	г	о	л	о	в	о	й	
Cost	0	1	0	1	0	1	1	1

Figure 5: LD between in-context cognates

We further apply the information-theoretical notion of surprisal [Shannon 1948] to model the predictability of a particular cross-lingual correspondence for a given pair of source and target languages. In particular, we assume that higher word adaptation surprisal (WAS) values cause more intercomprehension difficulties, so that it would be harder to recognize the actual cognate stimulus (BE, UK, BG) from the viewpoint of subjects' language (RU). WAS corresponds to the sum of the character adaptation surprisal (CAS) values, and is calculated in bits according to the character transformation probabilities[8]. Again, we use the *incom.py* tool [Mosbach et al. 2019] for calculating the mean normalized WAS (nWAS) values (in bits) for the stimulus word-forms w.r.t. their target cognates in the subjects' language (Table 2).

in-context word-forms	mean nLD to RU cognates	mean nWAS w.r.t. RU
BE	34%	0.64
UK	29%	0.59
BG	37%	1.06

Table 2: Mean normalized Levenshtein distance and mean normalized word adaptation surprisal between BE, UK, BG and RU

Lexical factors: As non-cognates, i.e. historically non-related words, tend to be unintelligible to readers with no prior knowledge of the stimulus language, we expect their large proportion to impede intercomprehension [Gooskens 2019]. As a rule, the percentage of non-cognates indicatively determines the lexical distance between closely related languages. The so-called *false friends* need special attention here, since they may cause even larger difficulties than non-cognates. For marking cross-lingual lexical differences, we give points to cognate word-form pairs of aligned sentences, namely: a non-cognate (incl. false friends) obtains one point, a partial cognate[9] obtains half a point, and a cognate (i.e. with a common root and similar meaning) obtains zero points. In some cases the cognate correspondence may consist of non-cognates, for example, BG *око* (oko) 'eye' translates into RU *глаз* (glaz), forming a pair of words that are non-cognate. Nevertheless, such a word pair obtains zero points too, as there is a synonym *око* (oko) in RU, which makes the written BG word *око* (oko) understandable. Our hypothesis is that with an increasing percentage of non-cognates, partial cognates and false friends, the subjects perceive less similarity, which makes it more difficult for them to understand an unknown even though related language. In Table 3 we present for each language pair the mean lexical distances.[10]

[8] $\mathrm{CAS}\,(\mathrm{L1}=\mathrm{c1}|\mathrm{L2}=\mathrm{c2}) = -\log_2 \mathrm{P}(\mathrm{L1}=\mathrm{c1}|\mathrm{L2}=\mathrm{c2})$

L1 – native language, c1 – character of the native language

L2 – stimulus language, c2 – character of the stimulus language

[9] Partial cognates are words with a similar meaning, for example, BG *жена* (žena) 'woman' and RU *жена* (žena) 'wife' both point the reader towards the correct category of adult female (cf. [Golubović 2016: 244]).

[10] This calculation is based on lexical words (i.e. nouns, adjectives, verbs, adverbs, and numerals), as they are more important for intelligibility than function words (cf. [van Bezooijen and Gooskens 2005], [Gooskens 2006]).

language pair	mean lexical distances
BE–RU	9%
UK–RU	12%
BG–RU	13%

Table 3: Lexical distance between BE, UK, BG and RU

Syntactic factors: When processing a sentence in a (closely) related language, the readers may experience that some items are displaced, missing or superfluous. To approximate this situation, we use the respective syntactically relevant measures of movement, insertion and deletion [Heeringa et al. 2017]. As an illustration, let us take the BG–RU alignment in Figure 6.

	1	2	3	4	5	6	7	8	9	10	11
BG		Искам		и	моят	мъж	да	бъде	свободен	като	вятъра
RU	Я	хочу	чтобы	и	мой	муж		был	волен	как	ветер

Figure 6: BG–RU alignment for the sentence ‘I want my husband to wander as free as the wind.’

Our assumption is that the more positions a word moves and the more words need to be added or deleted in order to achieve a perfect cross-lingual correspondence, the more negative the effect on intercomprehension tends to be. Here, the RU *чтобы* (čtoby ‘in order to’: position 3) corresponds to the BG *да* (da: position 7), even if they occur in different alignment slots and are not cognates of each other, so the number of movements here is 7–3=4. With the InDel measure, we register insertions and deletions in the alignment slots. From the perspective of RU subjects, we need one insertion, at the first position, which results in InDel value 1. Normalizing the InDel distance by the number of the alignment slots (11), we obtain nInDel: 0.09 or 9% (cf. [Gooskens and Swarte 2017]). Table 4 summarizes the mean nInDel and movement distances from the stimuli sentence alignments to their RU targets.

stimulus language	subjects’ language: RU	
	mean nInDel distance	mean movement distance
BE	11% (1.15)	1.38 (0.33)
UK	7% (0.87)	1.67 (0.35)
BG	16% (1.50)	2.04 (0.58)

Table 4: Syntactic distances between BE, UK, BG and RU

The average number of words to be inserted or deleted per sentence in the stimulus from a RU perspective are presented (in brackets) after the respective nInDel value. Similarly, the average number of words to be moved to another sentence position from a RU perspective are given (in brackets) after the respective movement value.

4 Potential predictors of human performance

We can now examine to what extent the presented linguistic factors (cf. Section 3) can predict and explain the experimental results (cf. Section 2) on cross-lingual in-context recognition of cognates. Regarding the impact of morphological and orthographic factors on human performance in intelligibility tests, we correlate the obtained intelligibility scores with the calculated nLD and nWAS between their stimulus word-forms and their target cognates in RU[11]. To understand the role of the context in cognate recognition, we correlate the obtained intercomprehension scores with the

[11] The correlations with the nLD and nWAS for the language group BG–RU are based on 118 BG–RU cognate pairs, since two word pairs consist of non-cognates, compare BG–RU: *киселец–щавель* (kiselece–ščavel’) ‘dock’ and *барут–порох* (barut–poroch) ‘gun powder’.

calculated lexical and syntactic distances. Table 6 presents the correlation coefficients (Pearson's *r*) and the *p*-values. Negative correlations can be expected for all linguistic factors.

The correlation of the intercomprehension scores with the nLD is higher than with the nWAS, and it is significant for all groups. This reveals a clear relationship between morphological and orthographic similarities and successful in-context cognate recognition. Yet, the nWAS is likely to be the predictor of intelligibility only for UK and BG stimuli, but it shows no significant correlation with the intelligibility of BE stimuli.

The effect of lexical distances may be more difficult to predict. The correlation of the intelligibility scores for BG with the lexical distance is negative but very small and not significant. For BE and UK cognates the correlations are not negative, but insignificant.

As far as the syntactic distances are concerned, the highest negative correlation is found for the nInDel distance in BE cognate recognition. For UK and BG cognates, the correlations are negative but smaller. However, none of them is significant. Furthermore, the movement distance seems not to play any role for RU readers. For all three languages, the correlations are not negative as assumed, but insignificant.

in-context cognate recognition	linguistic factors				
	nLD	nWAS	lex. distance	nInDel	movement
BE–RU group	$r = –.342$ $p < .01$	$r = –.178$ $p = .17$	$r = .036$ $p = .78$	$r = –.210$ $p = .11$	$r = .169$ $p = .20$
UK–RU group	$r = – .536$ $p = 1e\text{-}05$	$r = –.447$ $p < .0005$	$r = .055$ $p = .67$	$r = –.075$ $p = .57$	$r = .011$ $p = .93$
BG–RU group	$r = –.413$ $p = 3.40e\text{-}06$	$r = –.191$ $p < .05$	$r = –.171$ $p = .06$	$r = –.113$ $p = .22$	$r = .041$ $p = .66$

Table 6: Correlations between linguistic factors and intercomprehension scores of RU subjects in three stimuli groups

5 Discussion and conclusion

The observed human performance validates the intuition that Russian-speaking subjects understand better East Slavic, to which RU belongs together with UK and BE, than South Slavic. Comparing "out-of-context guessing" with "in-context recognition", we see that the latter is much better only for BE–RU, while the former shows slightly higher intercomprehension scores for UK–RU and BG–RU. However, a closer look at individual cognate pairs reveals that in all three stimuli-subject combinations, more cognates are successfully recognized in a context than without a context. The slightly better performance in single word recognition in UK–RU and BG–RU is due to the higher intelligibility score of some correctly translated single words in the two groups.

stimuli–subject	**"free translation" out-of-context guessing**	**"contextualized translation" in-context recognition**
UK–RU group	slightly higher	
BE–RU group		much better
BG–RU group	slightly higher	
overall intelligibility		more successful
subject performance	UK>BE>BG	UK>BE>BG

The present investigation has shown that morphological and orthographic factors play an important role in correct cognate recognition with context. The nLD is a significant predictor of intelligibility of in-context cognates in all three Slavic language groups. In addition to morphological and orthographic similarities, the predictability of correspondences by means of nWAS seems to influence the successful human performance too, but the significance of this variable holds only for UK and BG cognates in context.

The relationship between lexical distances and intelligibility of cognates in context is less clear. Investigating linguistic predictors of inter-Scandinavian intelligibility, [Gooskens 2006: 111] points

out that the impact of lexical differences depends on the nature of the lexical deviances and in some cases one single deviant word can be very disturbing for the comprehensibility of the whole context while in other cases a number of non-cognates is hardly disturbing, because they are not important concepts. Examining the mutual intelligibility of some West and South Slavic languages, [Golubović 2016: 123] found out that lexical distance (on the basis of all words) is a good predictor of mutual intelligibility in individual word translation tasks and cloze tests. In our experiments, the quantitative effect of lexical differences seems not to play so important role. The readers seem to pay more attention to morphological and orthographic similarities and differences of the particular cognate pair in order to succeed in intelligibility tests. We assume that lexical distance may be a better predictor in non-cognate recognition (cf. the role of predictive context in [Jágrová and Avgustinova 2019]).

We have also compared the experimental results to the two presented syntactic distances and found negative correlations between the nInDel distance and the intelligibility scores for all stimulus languages. However, the correlations are low and none of the correlations is significant. According to [Heeringa et al. 2017] the InDel distance highly depends on the way sentences are translated. Indeed, in a reading intercomprehension scenario a reader is trying to match words in a stimulus sentence to the words in his or her native language. According to our results, it seems that the number of words that should be added or deleted in comparison to the closest possible sentence which the reader would have used him/herself, can be seen only as a tendency that has a negative effect on correct cognate recognition in context.

Additionally we found that the movement distances do not correlate negatively with the intelligibility scores, as assumed, but all of them were insignificant. This means that the assumption that the further a word is moved the more negatively it will affect intelligibility cannot be confirmed in this study. [Heeringa et al. 2017] pointed that [Swarte 2016] measured mutual intelligibility between five Germanic languages by means of a spoken and a written cloze test and correlated the intelligibility scores of the experiments with the movement distance. She found a significant correlation between the movement measure and written and spoken intelligibility at the 0.05 level. The reason that the movement distance does not explain the intelligibility of cognates in our experiments might be that the word order is not so rigid in Slavic languages, as, for example, in Germanic languages.

Linguistic similarity is a multidimensional phenomenon [van Heuven 2008]. In this investigation, we focused on linguistic factors that predict inherent intelligibility between related languages. New testing methods have been established in the last decade [Gooskens 2018] to define certain breakdown points at which language varieties become unintelligible. The goal is to provide a more solid and experimentally grounded foundation for the classical and traditional claims made by linguists about genealogical relatedness among languages. In future research, we shall extend our approach to include extra-linguistic factors with obvious impact on mutual intelligibility across languages [Gooskens and van Heuven 2020] in order to adequately model what is known as receptive multilingualism.

Acknowledgements

We thank Marius Mosbach for the implementation of our data in the *incom.py* tool, Philip Georgis for assistance in calculating of lexical and syntactic distances, and Hasan Alam for support with the implementation of online experiments. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

[1] van Bezooijen Renée, Gooskens Charlotte. How easy is it for speakers of Dutch to understand spoken and written Frisian and Afrikaans, and why? — Linguistics in the Netherlands 22. — John Benjamins, Amsterdam, 2005. — P. 13–24.

[2] Frost Ram. Towards a universal model of reading. — Behavioral and Brain Sciences 35(5). — Cambridge University Press, 2012. — P. 263–329.

[3] Golubović Jelena. Mutual intelligibility in the Slavic language area. — University of Groningen, PhD thesis, 2016.

[4] Gooskens Charlotte. Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility. — Linguistics in the Netherlands 23. — John Benjamins, Amsterdam, 2006. — P. 101–113.

[5] Gooskens Charlotte. Dialect intelligibility. — Handbook of dialectology. — Wiley-Blackwell, Oxford, 2018. — P. 204–218.
[6] Gooskens Charlotte. Receptive multilingualism. — Multidisciplinary perspectives on multilingualism: The fundamentals. — De Gruyter Mouton, Berlin, 2019. — P. 149–174.
[7] Gooskens Charlotte, van Heuven Vincent J. How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables*?* — Linguistic Approaches to Bilingualism 10(3). — John Benjamins, Amsterdam, 2020. — P. 351–379.
[8] Gooskens Charlotte, Swarte Femke. Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages. — Nordic Journal of Linguistics 40(2). — Cambridge University Press, 2017. — P. 123–147.
[9] Heeringa Wilbert, Swarte Femke, Schüppert Anja, Gooskens Charlotte. Measuring syntactical variation in Germanic texts. — Digital Scholarship in the Humanities 33(2). — Oxford University Press, 2017. — P. 279–296.
[10] van Heuven Vincent J. Making sense of strange sounds: (Mutual) intelligibility of related language varieties. A review. — International Journal of Humanities and Arts Computing 2(1-2). — Edinburgh University Press, 2008. — P. 39–62.
[11] Jágrová Klára, Avgustinova Tania. Intelligibility of highly predictable Polish target words in sentences presented to Czech readers. — CICLing 2019, Springer's Lecture Notes in Computer Science: preprint, 2019.
[12] Levenshtein Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. — Doklady of the Soviet Academy 163(4), 1965. — P. 845–848.
[13] Lutjeharms Madeline. Verarbeitungsebenen beim Lesen in Fremdsprachen. — Neuere Forschungen zur Europäischen Interkomprehension. — Shaker Verlag, Aachen, 2004. — P. 67–82.
[14] Mosbach Marius, Stenger Irina, Avgustinova Tania, Klakow Dietrich. incom.py – A Toolbox for Calculating Linguistic Distances and Asymmetries between Related Languages // Proceedings of the International Conference Recent Advances in Natural Languages Processing 2019. — Varna, Bulgaria, 2019. — P. 811–819.
[15] Shannon Claude E. A mathematical theory of communication. — Bell System Technical Journal 27(3), 1948. — P. 623–656.
[16] Stenger Irina. Zur Rolle der Orthographie in der slavischen Interkomprehension mit besonderem Fokus auf die kyrillische Schrift. — Universaar, Saarbrücken, PhD thesis, 2019.
[17] Stenger Irina, Avgustinova Tania. Visual vs. auditory perception of Bulgarian stimuli by Russian native speakers // Proceedings of the Annual International Conference ‘Dialogue’ 2020. — Moscow, Russia, 2020. — P. 623–656.
[18] Swarte Femke. Predicting the Mutuall Intelligibility of Germanic languages from linguistic and extra-linguistic factors. — University of Groningen, PhD thesis, 2016.

What have I seen? On the meaning and distribution of an experiential discourse marker

Tatevosov S. G.
Lomonosov Moscow State University,
Moscow, Russia
tatevosov@gmail.com

Kisseleva X. L.
Vinogradov Russian Language Institute of the Russian Academy of Sciences,
Moscow, Russia
xkisseleva@gmail.com

Abstract

The paper explores the discourse marker *ja vižu* (lit. 'I see') and its cross-linguistic counterparts. We argue that it presents its scope proposition as the product of abduction, a logical inference that derives the optimal explanation for the observed state of affairs. This view is supported by the set of observations suggesting that restrictions on the distribution of *ja vižu* are mostly derivable as restrictions on abuctive reasoning, which involve informativeness, likelihood and parsimony considerations.

Keywords: abduction, non-monotonic reasoning, evidentiality, discourse markers
DOI: 10.28995/2075-7182-2021-20-669-680

Что я видел? Некоторые особенности значения и употребления экспериентивного дискурсивного показателя[1]

Татевосов С. Г.
МГУ им. М. В. Ломоносова,
Москва, Россия
tatevosov@gmail.com

Киселева К. Л.
ИРЯ им. В. В. Виноградова РАН,
Москва, Россия
xkisseleva@gmail.com

Аннотация

В статье обсуждается дискурсивный маркер я вижу и его межъязыковые аналоги. Предлагается анализ, согласно которому пропозиция в его сфере действия представляет собой продукт абдукции, операции, которая обеспечивает вывод наилучшего объяснения для наблюдаемого положения вещей. Этот анализ подкрепляется эмпирическими фактами, которые указывают, то ограничения на дистрибуцию я вижу представляют собой ограничения на абдуктивный логический вывод, обусловленные соображениями информативности, вероятности и экономии.

Ключевые слова: абдукция, немонотонные рассуждения, эвиденциальность, дискурсивные маркеры

1 *Я вижу*: несколько ограничений

Цель этой статьи — предложить семантическое описание выражения *я вижу* в примерах типа (1)-(2), где оно сообщает некоторый особый статус пропозиции, составляющей основное содержание предложения. В (1) это происходит с пропозицией 'ты бежала', а в (2) — с 'вы засыпаете'. Наша основная задача — прояснить, какой информацией *я вижу* снабжает эти пропозиции.

[1] Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Сохранение мирового культурно-исторического наследия»

(1) — Сядь. Ты, *я вижу*, бежала. — Катерина чуть отошла, послушалась, села на пень. [И. А. Новиков. Жертва (1921)]

(2) Позвольте вам подложить подушку, вы, *я вижу*, засыпаете. [А. Пятигорский. Древний Человек в Городе // «Октябрь», 2001]

Морфосинтаксическая реализация пропозиции, выступающей аргументом *я вижу*, ничем принципиально не ограничена. Предложения с *я вижу* могут иметь любую временную референцию — к настоящему, как в (2), прошедшему, как в (1), или будущему, как в (3), — а также любую видовую характеристику, ср. НСВ в (1)-(2) и СВ в (3)-(4).

(3) — Без таких … технологичных комплексов, нам, конечно, не обеспечить … собственными продуктами питания, мясом жителей России… И это, *я вижу*, будет сделано в Приморском крае вот такими молодыми амбициозными людьми. [https://vestiprim.ru/news/]

(4) — Садитесь. Вы, *я вижу*, удрали с работы? — А я и вообще не ходил. [И. Л. Солоневич. Россия в концлагере (1935)]

Употребления *я вижу*, насколько можно судить, встречаются в утвердительных предложениях (1) и общих вопросах (5). В частных вопросах и императивах *я вижу* невозможно, ср. (6)-(7).

(5) — Вчерашняя знакомая, Александр, узнаешь? Вы, *я вижу*, подружились? — сказал он Кириллу. — Да. [К. А. Федин. Первые радости (1943-1945)]

(6) (**Я вижу*) Куда (**я вижу*) Володя (**я вижу*) пошел?

(7) (**Я вижу*) Дай (**я вижу*) мне (**я вижу*) эту книжку!

Я вижу используется практически исключительно в независимых предложениях, как (1). Во вложенных клаузах, например, в (8a-b), *я вижу* неграмматично; повышение приемлемости для некоторых носителей регистрируется лишь в конфигурациях с глаголами речи и частицей *мол*:

(8) a. *Алеша подумал / представил себе / поверил / сказал, что ты, *я вижу*, бежала.

*Алеша подумал / представил себе / поверил / сказал, что она, *я вижу*, бежала.

(9) ?Алеша сказал, что, мол, ты, *я вижу*, бежала.

Семантическое содержание, которое привносит *я вижу*, невозможно отрицать или ставить под вопрос:

(10) — *Я вижу*, Володя пришел.
— *Нет, ты ошибаешься. На самом деле ты этого не видишь
— *А ты уверен? Ты действительно это видишь?

Для *я вижу* эти ограничения, как кажется, отказываются более жесткими, чем для многих других единиц со сходной семантической функцией — установление отношения между говорящим и пропозицией. (11)-(13) иллюстрируют *кажется* и *похоже*:

(11) — Кажется, Володя пришел.
— ?Нет, ты ошибаешься. Тебе это не кажется.

(12) — Похоже, Володя пришел.
— ?Нет, ты ошибаешься. На это не похоже.

В той степени, в которой ответные реплики в (11)-(12) приемлемы, они, по-видимому, представляют собой языковую игру с переключением между двумя прочтениями *кажется* и *похоже* — «модальным» и «буквальным». Однако такая же игра, судя по всему, невозможна с *я вижу*.

Предложения с *я вижу* предполагают полную ответственность говорящего за истинность пропозиции. Любое выражение, снимающее эту ответственность, создает противоречие.

(13) *Я вижу*, Володя пришел. #Но на самом деле нет / #Но, может быть, и нет / #Но я думаю, что нет / #Но я сомневаюсь, что это так.

Можно, наконец, заметить, что *я вижу* интерпретируется за пределами сентенциального отрицания и, вероятно, других операторов сентенциального уровня.

(14) А ты человеческих слов, *я вижу*, не понимаешь. [В. Шукшин. Други игрищ и забав (1974)]

Единственно возможное прочтение (14) — то, при котором предполагается узкая сфера действия отрицания (*я вижу* > ¬). Узкая сфера действия *я вижу* (‘неверно, что я нахожусь в отношении *я вижу* с пропозицией-аргументом’) для (14) невозможна.

Все перечисленные свойства хорошо укладываются в семантический профиль показателей эвиденциальности, которые характеризуются неоспоримостью ([Faller 2002, 2006], [Korotkova 2016], [Murray 2010, 2012], [Portner 2018]), ответственностью говорящего за истинность пропозиции ([Murray 2010], [Korotkova 2016], [Koev 2011]) и широкой сферой действия по отношению к отрицанию ([Faller 2002, 2006]). Разные теории дают разное объяснение этим явлениям; мы надеемся, что сказанное ниже окажется совместимо с любым анализом, который в конечном итоге будет признан верным. Соответственно, далее мы не обсуждаем эти вопросы и сосредоточимся на свойствах я вижу, о которых сказано ниже.

2 *Я вижу* и инферентивность

Эвиденциальные показатели описывают источник информации, на который опирается утверждение. В частности, в довольно широком классе контекстов *я вижу* ведет себя как показатель инферентивности, т.е. сообщает, что пропозиция создается в результате логического вывода.

Грамматические единицы, которые допускают прочтение, похожее на *я вижу*, засвидетельствованы в значительном числе ареально и генетически независимых языков. (15) — иллюстрация из языка лиллуэт (салишская семья):

(15) Контекст: У говорящего было несколько кусков тсвана[2], и накануне он проверял, что они на месте. Сегодня он обнаружил, что тсван исчез. Говорящий, который знает, что Джон, сосед по квартире, — большой любитель тсвана, обнаруживает объедки в его комнате.

ts’aqw-an’-ás-**an’**	i	ts’wán-a	kw	s-John
есть-DIR-3ERG-PERC.EVID	DET.PL	тсван-EXIS	DET	NOM-Джон

‘*Я вижу*, Джон съел тсван.’ (Matthewson at al. 2007)

В (15), видя объедки и зная о любви соседа к этому деликатесу, говорящий заключает, что пропозиция ‘Джон съел тсван’ истинна:

(16) Я наблюдаю q, q = объедки тсвана находятся у комнате Джона, и, имея q, а также то, что известно в текущем контексте о Джоне и тсване,
я делаю вывод, что p, p = Джон съел тсван, верно описывает мироздание.

Ровно то же самое происходит и в примерах типа (1), а также в (4) и (17).

[2] Тсван — вяленый лосось.

(17) Порывшись в бумагах, он достает из недр стопки один лист.
—— Вот, – повторяет он и, показав мне усеянный буквами лист, начинает читать. – Утверждена комплексная программа сохранения церкви Успения Богородицы до 2018 года, включающая такие пункты, как: организация и проведение ежегодной конференции по спасению церкви; создание общенационального фонда по спасению памятников архитектуры Молдавии…, — заканчивает Кокша и смотрит на меня просветленным взглядом.
— Нда, — уважительно качаю головой я, — подготовились вы, *я вижу*, основательно.
[Сергей Дигол. Три фута под землей // «Волга», 2012]

В (17) утверждение ‘вы подготовились основательно’ опирается, в частности, на наблюдаемое поведение собеседника, который уверенно и подробно излагает обстоятельства, имеющие отношение к теме разговора.

Наличие в ближайшей временной окрестности момента речи воспринимаемого положения вещей, позволяющего осуществить логический вывод утверждаемой пропозиции, — необходимое условие употребления *я вижу*. *Я вижу* невозможно, если вывод опирается только на доступные в контексте общие знания и соображения.

(18) — Ты думаешь, Маша уже дома?
— Семь часов, рабочий день закончился, она никуда больше не собиралась, до дома ехать сорок минут… #*Я вижу*, она уже дома.

Показатели, которые подчиняются ограничению, иллюстрируемому в (18), часто называются экспериенциальными инферентивными показателями. Впервые этот термин употреблен, насколько нам известно, в [Anderson 1986].

Положение вещей, необходимое для осуществления логического вывода, как можно предположить исходя из лексического состава *я вижу*, чаще всего воспринимается визуально. Однако возможно использование и других каналов восприятия:

(19) Говорящий слышит, что в квартире соседей, где происходит ремонт, возобновился звук перфоратора.
— Их работники, *я вижу*, пришли с обеда.

Все эти примеры как будто указывают на многообещающий параллелизм в дистрибуции *я вижу* и описанных в литературе показателей экспериенциальной инферентивности. У показателей инферентивности типа лиллуэтского *–an′* в (15) есть, однако, два ограничения, отсутствующих у *я вижу*. Во-первых, описываемое утверждаемой пропозицией положение вещей (‘Джон съел тсван’) должно иметь место в прошлом. Во-вторых, оно не может восприниматься говорящим непосредственно. В любом контексте, в котором говорящий наблюдает ситуацию поедания тсвана Джоном и при этом отдает себе отчет в происходящем, а затем рассказывает об этом, предложение (15) невозможно.

Исключение составляет очень специфический класс контекстов, когда от осознания говорящего ускользает или идентичность участников ситуации или ее дескриптивные свойства, как, например, в (20):

(20) Говорящий вошел в комнату и увидел темный силуэт человека, доедающего тсван. Прежде чем говорящий успел включить свет, вор покинул комнату. Позже говорящий нашел крошки тсвана в комнате Джона.
‘Я вижу, [Джон]$_F$ съел тсван.’

В (20) информация об агенсе недоступна в момент осуществления ситуации и восполняется посредством логического вывода, как и в исходном предложении (15), а соответствующую составляющую в (20) можно опознать по обязательному фокусному выделению.

Такие же ограничения засвидетельствованы для других показателей, допускающих, как лиллуэтский *–an’*, инферентивное употребление в огромном количестве языков.

Я вижу имеет более широкую дистрибуцию, чем показатели типа *–an'*. Как мы видели в (1)-(4), утверждаемая пропозиция совместима с любой временной локализацией. А с точки зрения непосредственной засвидетельствованности *я вижу* имеет неодинаковый набор возможностей для разных типов временной референции.

Самый простой случай — временная референция к будущему.

(21) — Да, он таки талантливый человек, — удивленно возразил профессор, — и, *я вижу*, сумеет извлечь толк из всех своих тополей. [Ю. Домбровский. Рождение мыши (1951-1956)]

Ситуации в будущем еще не наступили, и непосредственно наблюдать их тем самым невозможно. В (21) утверждение о будущем по необходимости опирается на наблюдения над происходящим вокруг момента речи: статья упоминаемого персонажа позволяет профессору предполагать для него блестящее будущее.

При временной локализации описываемой ситуации в прошлом, *я вижу* ведет себя так же, как и *–an'* и его аналоги в других языках: если описываемая ситуация наблюдаема и говорящий идентифицирует все ее компоненты, *я вижу* невозможно.

(22) Говорящий видел, как Володя взял со стола конфету и съел ее.
— ??*Я вижу*, Володя съел конфету.

Параллелизм с классическими инферентивными показателями типа *–an'* подчеркивается и тем, что прямая засвидетельствованность совместима лишь с контекстами неполного восприятия типа (20).

Иначе обстоит дело в предложениях, где описываемая ситуация имеет место в настоящем.

(23) Перессорился со всеми, а теперь, *я вижу*, хочешь и с Соколовым поссориться. [В. Гроссман. Жизнь и судьба, часть 2 (1960)]

(24) Контекст: Говорящий встречает в кафе старого знакомого.
— Ты, *я вижу*, сидишь за тем же столиком. [http://www.lib.ru/inproz/fisher_t/]

(25) Тетя Клава заглянула в комнату к Арине.
— Ты, *я вижу*, дома... [М. Капранова. Ландыши, приносящие смерть (2020)]

(23) описывает ментальное состояние участника ситуации, которое недоступно для непосредственного наблюдения. Утверждение о таком состоянии — результат умозаключения, опирающегося на наблюдения за их внешними проявлениями. В этом отношении (23) устроено так же, как предложения с временной референцией к прошлому. В (24) и (25), однако, инферентивный характер употребления *я вижу* отнюдь не очевиден. Обе ситуации наблюдаются непосредственно и создают контекст прямой засвидетельствованности. Если показатели инферентивности всегда предполагают непрямую засвидетельствованность, использование *я вижу* в такой конфигурации должно приводить к аномальности. Однако этого не происходит.

Таким образом, диапазон интерпретаций *я вижу* описывается обобщением в (26):

(26) Засвидетельствованность и временная референция
При временной референции к прошлому и будущему прямая засвидетельствованность для *я вижу* невозможна. При временной референции в настоящему ограничений на засвидетельствованность нет.

В следующем разделе мы обсудим возможный анализ *я вижу* для случаев с временной референцией к прошлому, а затем наметим, как распространить его на остальные случаи.

3 Абдукция

В [Tatevosov 2019] предлагается анализ экспериенциальной инферентивности, описывающий ее как **вывод наилучшего объяснения**.

Почему, видя крошки тсвана в комнате Джона, мы заключаем, что утверждение ‘Джон съел тсван’ в (15) истинно? Логически это утверждение не следует из ‘В комнате Джона есть крошки тсвана’. Однако в том контексте, о котором идет речь, оно выступает **наилучшим объяснением** наблюдаемого положения вещей.

Нахождение наилучшего объяснения — логическая операция, известная как **абдукция** ([Josephson & Josephson 1996], [McIlraith 1998], [Hobbs 2004], **[**Douven 2016]). В первом приближении ее можно описать следующим образом:

(27) При наличии наблюдаемого положения вещей *E* и возможных объяснений $H_1, \ldots H_n$, следует заключить, что истинно то H_i, которое объясняет *E* наилучшим образом.

Абдукция — важнейшая часть обыденной логики. Эту операцию, например, проделывает врач, когда, изучив симптомы заболевания, ставит диагноз. Абдукция широко представлена в естественнонаучных рассуждениях, когда требуется дать каузальное объяснение наблюдаемому явлению, а в распоряжении исследователя имеется более одной гипотезы.

В предельно схематичном виде абдукция, а также ее отличие от дедукции иллюстрируется в (28):

(28)

	Абдукция	Дедукция
Наблюдение	q	p
Теория	$p \rightarrow q$	$p \rightarrow q$
Вывод	p	q

Оба типа рассуждений опираются на «теорию», образуемую законоподобными утверждениями с импликацией вида $p \rightarrow q$. При дедуктивном рассуждении, имея это утверждение и убедившись в истинности посылки p, мы можем прийти к выводу q: *все вороны черные*, *Володя ворон*, следовательно, *Володя черный*. Так устроен, например, классический силлогизм. При абдуктивном рассуждении мы убеждаемся в истинности q, и опираясь на теорию, выдвигаем гипотезу об истинности p: *при кори у пациента возникает характерная сыпь*; *мы наблюдаем именно такую сыпь*, следовательно, *пациент болен корью*.

Таким образом, опираясь на параллелизм *я вижу* и показателей экспериенциальной инферентивности, можно предположить следующую семантику этого выражения:

(29) *Я вижу* используется для утверждения пропозиции p, если p — наилучшее объяснение для пропозиции q, причем истинность последней обеспечивается непосредственным наблюдением, происходящим в окрестностях момент речи.

Согласно (29), при использовании *я вижу* пропозиция, входящая в сферу действия этого выражения, абдуцируется (выводится) из некоторого наблюдаемого положения вещей, причем время наблюдения содержит в себе момент речи.

Помимо двух пропозиций — «наблюдаемой» и «абдуцируемой», критически важную роль в дистрибуции *я вижу* играет фоновая информация, доступная в текущем контексте. Ср. минимальную пару в (30):

(30) Контекст 1. Известно, что Надя никому не дает поносить свое пальто и, придя домой, оставляет его в прихожей на вешалке. Говорящий видит в прихожей Надино пальто.
Контекст 2. *Известно, что Надя дает младшим сестрам поносить свои вещи, и те охотно этим пользуются. Все они, придя домой, оставляют пальто в прихожей на вешалке. Говорящий видит в прихожей Надино пальто.
— *Я вижу*, Надя пришла.

В (30) Контекст 1 и Контекст 2 не различаются с точки зрения наблюдаемой и абдуцируемой пропозиций. Это соответственно ‘Надино пальто в прихожей’ и ‘Надя пришла’. Однако ‘Надя пришла’ абдуцируется только в первом контексте. Если верно, что Надя всегда носит пальто сама, то его появление в прихожей однозначно указывает на приход владелицы. Если пальто носят несколько человек, ‘Надя пришла’ перестает быть наилучшим объяснением, поскольку альтернативные объяснения, в отсутствие дополнительной информации, не менее вероятны.

В пользу того, что в семантику *я вижу* встроена абдуция, свидетельствуют факты, указывающие, что предложения с *я вижу* подчиняются независимо засвидетельствованным ограничениям на абдукцию. Ограничения на абдукцию в конечном итоге вытекают из идеи хорошего / лучшего / наилучшего объяснения, а именно из соображений сравнительной вероятности, информативности, экономичности, общности, эмпирической широты объяснения. Для наших целей существенны два ограничения, которые показаны в (31) и (38).

(31) Абдукция как вывод наиболее информативного каузального объяснения
Пусть q — пропозиция, описывающая наблюдаемое положение вещей, а p и p' — ее возможные объяснения, причем p’ — логическое следствие p. В этом случае из q абдуцируется более сильная p, а не более слабая p'.

(33)-(36) иллюстрирует пропозицию ‘Володя унес маркер’ из предложения (32), а также несколько альтернативных пропозиций, связанных отношением асимметричного следования.

(32) Контекст. Единственный маркер, которым можно писать на доске, исчез с кафедры. На кафедре, кроме говорящего, работают преподаватели Володя, Лева и Инесса. На следующий день говорящий видит Володю, у которого маркер торчит из кармана.
— *Я вижу*, Володя унес маркер.

(33) Володя унес маркер
унес(маркер)(Володя)

(34) ?? Кто-то унес маркер
∃x унес(маркер)(x)

(35) *Кто-то что-то сделал с маркером
∃R ∃x R(маркер)(x)

(36) *Кто-то что-то с чем-то сделал
∃R ∃x ∃y R(y)(x)

Пропозиции в (34)-(36), будучи логическими следствиями (33), истинны, если истинна (33); все они верно описывают мироздание, в котором Володя унес маркер. Однако использование любой из них в (32) вместо (33) семантически аномально, причем тем более аномально, чем более слабая пропозиция используется. Это показывает, что *я вижу* подчиняется принципу в (31).

Рассмотрим другой контекст, который получается из (32) удалением информации, что у Володи торчит из кармана маркер:

(37) Контекст. Единственный маркер, которым можно писать на доске, исчез с кафедры. На кафедре, кроме говорящего, работают Володя, Коля и Маша. Говорящий, не найдя маркер на обычном месте:
a. — *Я вижу*, кто-то унес маркер.
b. — **Я вижу*, Володя унес маркер.

(37a-b) показывают, что в отличие от (32), самая информативная из доступных пропозиций — это ‘Кто-то унес маркер’. ‘Володя унес маркер’, будучи более информативной, тем не менее недоступна. Эту недоступность можно объяснить принципом в (38), который требует, чтобы абдуцируемая пропозиция была наиболее вероятным объяснением наблюдаемого положения вещей.

(Начиная с этого момента, под объяснением мы будем понимать каузальное объяснение; никакие существенные рассуждения, впрочем, на это допущение не опираются.)

(38) Абдукция как вывод наиболее вероятного каузального объяснения
Пусть q — пропозиция, описывающая наблюдаемое положение вещей, а p и p' — ее возможные объяснения, не связанные отношением логического следования, причем p cause q более вероятно в данном контексте, чем p' cause q. В этом случае из q абдуцируется более вероятная p, а не менее вероятная p'.

В (37) есть несколько альтернативных объяснений исчезновения маркера: Володя унес маркер, Лева унес маркер, Инесса унесла маркер. Если помимо информации в (37), больше ничего неизвестно, оснований рассматривать одно из этих трех объяснений как более вероятное, чем другие, не появляется. (38) предсказывает, что ни одна из этих пропозиций не лучше для абдукции чем другие. Соответственно, для утверждения можно выбрать лишь более слабую пропозицию в (37a).

Если же обогатить контекст, доступный в (37), информацией из (38), одна из возможностей делается существенно более вероятной, чем две другие, и использование более сильной пропозиции в комбинации с *я вижу* снова становится возможным:

(39) Контекст: Известно, что за последний год Володя уносит маркер с кафедры почти после каждого занятия. Коля и Маша за этим не замечены ни разу.
— *Я вижу*, Володя (опять) унес маркер.

Наконец, последнее свойство естественно-языковой абдукции, проявляющееся в ограничениях на дистрибуцию *я вижу*, состоит в том, что абдуцируемая пропозиция должна быть не просто вероятнее альтернатив, но и достаточно вероятна сама по себе[3]. В (40) создается контекст, когда у наблюдаемого положения вещей нет лучшего объяснения, чем *маркер унесли инопланетяне*, однако это объяснение, будучи невероятным, не оставляет места для *я вижу*.

(40) Контекст. Два дня назад говорящий убедился, что маркер на месте. Точно известно, что с этого момента на кафедру никто не заходил. Тем не менее говорящий обнаруживает, что маркер исчез.
#—*Я вижу*, маркер унесли инопланетяне.

(40) возможно лишь как языковая игра, когда заведомо невероятное объяснение подается как единственно возможное. Соответственно, (40) сообщает слушающему, что у говорящего нет разумной гипотезы, описывающей происшедшее.

С учетом всего сказанного семантику абдуктивного отношения между двумя пропозициями p и q и множеством пропозиций P, которое задействуется в интерпретации *я вижу*, можно более формально описать так, как показано в (41):

(41) Абдукция
Пусть q — это пропозиция, описывающая наблюдаемое положение вещей; p — абдуцируемая пропозиция, которая выступает ее объяснением, P — множество пропозиций, которые полагаются истинными в текущем контексте и которые тем самым соответствуют доступной в этом контексте информации; ∩P — пропозиция, образуемая множеством миров, в которых истинны все пропозиции из P, или контекстное множество.

Тогда пропозиция p абдуцируется из пропозиции q относительно множества пропозиций P, ABDUCT(p, q, P), ровно в том случае, когда выполняются следующие условия (ср. [Brachman & Levesque 2004]):

1. Пропозиция q не является логическим следствием P; ∩P ⫋ q

2. Пропозиция p совместима с P, т.е. ¬p не является следствием P; ∩P ⫋ ¬p

[3] Мы признательны П. Портнеру за обсуждение этого компонента абдуктивной семантики.

3. Пропозиция q является логическим следствием $P \cup \{p\}$; $(\cap P \cap p) \subseteq q$

4. Множество пропозиций P — самое информативное из тех, которые поддерживают абдукцию p из q
$ABDUCT(p, q, P) \rightarrow [\forall Q\ ABDUCT(p, q, Q) \rightarrow P \subseteq Q]$ [4]

5. Для любой пропозиции p′, удовлетворяющей условиям 1.-4., вероятность p′ cause q ниже вероятности p cause q; p′ cause $q \prec_{Pr}$ p cause q

6. Вероятность p cause q превосходит контекстно-зависимый порог правдоподобия r_C; p cause $q \succ_{Pr} r_C$

Согласно условию (41.1) наблюдаемое положение вещей не вытекает из информации, заключенной в контексте, то есть нуждается в объяснении. (41.2) требует, чтобы объяснение не противоречило известным фактам. В соответствии с (41.3) абдуцируемая пропозиция в комбинации с контекстной информацией дает объяснение наблюдаемому положению вещей.

Свойство наибольшей информативности, описанное ранее в (31), обеспечивается экзостификацией контекста в (41.4).[5] (42) предотвращает абдукцию пропозиции в более сильном контексте, чем необходимо. Если контекст поддерживает более абдукцию более сильной пропозиции, мы не используем более слабую. Например, пропозиция в (34)(=(37a)) ∃x унес(маркер)(x) абдуцируется в контексте (37). Согласно (42), она не абдуцируется в более сильном / информативном контексте (32). Соответственно, в (32) она не конкурирует с более сильной пропозицией унес(маркер)(Володя) за статус наилучшего объяснения отсутствия маркера.

Далее, (41.5) описывает обобщение в (38): абдуцируемая пропозиция должна быть самым вероятным объяснением наблюдаемого. Наконец, (41.6) требует, чтобы объяснение было минимально правдоподобным и делает невозможным (40) и аналогичные примеры.

Таким образом, перечисленные в (41) условия дают достаточно точные предсказания о том, какого типа пропозиции и в каких контекстах абдуцируются в предложениях с *я вижу*.

В следующем разделе мы обратимся к двум другим свойствам предложений с *я вижу* из раздела 1 — временной референции, возможности прямой засвидетельствованности, а также отношению между этими свойствами.

4 Временная референция и прямая засвидетельствованность

Существенное эмпирическое обобщение о *я вижу* из раздела 2 состоит в том, что прямая засвидетельствованность описываемой ситуации допускается, если предложение с *я вижу* имеет временную референцию к настоящему. В двух других случаях — в предложениях о прошлом и о будущем — это невозможно.

Эти свойства не вытекают из абдуктивной семантики, описанной в предыдущем разделе, и должны получить отдельное объяснение. Наше предположение по поводу их источника сформулировано в (42):

(42) Эффекты прямой засвидетельствованности и временная интерпретация
a. Выражение *я вижу* вводит в семантическое представление событие получения говорящим информации об описываемой ситуации, имеющее временную локализацию в настоящем.
b. Условием прямой засвидетельствованности выступает пересечение описываемой ситуации и события получения информации о ней во времени.

Предположение (42a) помещает анализ *я вижу* в семейство теорий, которые усматривают в значении эвиденциальных показателей указание на событие получения информации о ситуации, описываемой пропозицией-аргументом ([Maisak, Tatevosov 2007], [Koev 2011], [Smirnova 2011,

[4] $P \subseteq Q$: множество пропозиций P не менее информативно, чем множество пропозиций Q.
[5] Мы признательны В. Лехнеру за обсуждение этого компонента абдуктивной семантики.

2012]). Можно, далее, предполагать, что временная морфология в выражении *я вижу* имеет регулярную временную интерпретацию, то есть локализует это событие в настоящем. Настоящее, далее, естественно понимать как интервал, включающий момент речи и его контекстно-релевантные окрестности, составляющие единый с моментом речи сегмент истории мироздания. Если все это верно, семантику *я вижу* можно уточнить так, как показано в (43):

(43) *Я вижу p* сообщает, что на некотором интервале t, включающем момент речи, имеет место событие e, в котором говорящий непосредственно наблюдает положение вещей q. Опираясь на q и контекстно-доступную информацию CB, говорящий производит абдукцию пропозиции p, ABDUCT(p, q, CB).

Далее в игру вступает условие в (42b): оно утверждает, что непосредственно наблюдать ситуации возможно, только если время наблюдения пересекается со временем ситуации. Если верно, что грамматические показатели разных времен вводят непересекающиеся временные интервалы, то единственная конфигурация, в которой возможна прямая засвидетельствованность, — это совпадение времени *я вижу* и времени пропозиции аргумента, то есть настоящее. В случае с прошедшим и будущим интервал, на котором происходит наблюдение, не пересекается с интервалом, занятым описываемой ситуацией. Соответственно, предметом наблюдения выступает не сама описываемая ситуации, а некоторое другое положение дел, которое затем используется при абдукции. Это объясняет невозможность прямой засвидетельствованности для предложений с настоящим и будущим грамматическим временем.

Рассмотрим более подробно случай, когда пропозиция-аргумент представлена предложением в настоящем времени. Соответствующие примеры показаны в (23)-(25) выше. В этом случае есть две возможности:

(44) a. Пропозиция, описывающая наблюдаемое положение вещей, отлична от утверждаемой пропозиции
b. Наблюдаемая и утверждаемая пропозиция совпадают

Первая возможность ничем принципиально не отличается от других случаев абдукции, обсуждавшихся выше. На основании наблюдаемого положения дел q и контекстной информации CB говорящий абдуцирует ненаблюдаемую пропозицию p. Единственная особенность: p и q обе содержат в себе момент речи и тем самым пересекаются во времени. Ровно это наблюдается в (23) выше: наблюдая поведение собеседника (q), говорящий заключает, что тот намерен поссориться с Соколовым (p).

Вторая возможность состоит в том, что наблюдаемая и утверждаемая пропозиции совпадают. Последняя, соответственно, абдуцируется сама из себя: «наблюдая p, я заключаю, что наилучшее для этого объяснение — то, что p имеет место». Это случай примеров (24)-(25): *ты, я вижу, сидишь за тем же столиком* и *ты, я вижу, дома*.

В этом месте возникает закономерный вопрос: насколько такую «абдукцию из себя», ABDUCT(p, p, CB), уместно признавать абдукцией? Безусловно, любое событие можно рассматривать как каузально зависимое от самого себя и тем самым дающее каузальное объяснение самому себе. Более того, такое объяснение имеет хорошие шансы оказаться наилучшим — например, оно имеет 100-процентую вероятность оказаться верным. Однако совпадение наблюдаемого и объясняемого полностью тривиализует абдукцию: независимо от содержания p и фоновой информации CB ABDUCT(p, p, CB) — это истинная пропозиция. Тем самым *я вижу* оказывается неинформативным.

Здесь открываются два пути. Первый: вывести случаи типа (24)-(25) из-под определения абдукции (41), внеся в это определение необходимые уточнения, рассматривать употребления *я вижу* в таких предложениях отдельно от всех остальных и разрабатывать для них отдельный анализ. Второй: допустить, что тривиализация абдукции возможна, если предложения с *я вижу* информативны в каком-то другом отношении.

Не исключая полностью первый путь, мы тем не менее предпочитаем второй. С одной стороны, признание отдельного значения *я вижу* в случаях типа (24)-(25) влечет за собой весь набор неудобств, связанных с постулированием многозначности и дальнейшим ее описанием и

объяснением. Во-вторых, с эмпирической точки зрения тривиализация абдуктивного компонента семантики *я вижу* не закрывает возможности интерпретировать такие предложения как информативные. Источник информативности в этом случае — компонент значения, введенный в (43): ‘на интервале, включающем момент речи, имеет место событие получения информации о p’. Он эксплицитно указывает, что до самого момента речи пропозиция отсутствовала в блоке контекстной информации. *Ты, я вижу, дома* в (25), тем самым, сообщает, что до момента речи говорящий не предполагал истинности пропозиции ‘ты дома’. Так возникает эффект новизны и неожиданности у предложений такого типа. Дополнительным свидетельством в пользу правдоподобности такого сценария выступают отмеченные в литературе эффекты «миративности» ([Meydan 1996], [Lazard, 1996], [DeLancey 1997]) у эвиденциальных показателей в контекстах, похожих на (24)-(25).

Это соображение завершает наше обсуждение семантики и дистрибуции *я вижу*. Мы готовы суммировать основные наблюдения, изложенные выше.

5 Сумма наблюдений

Я вижу — выражение, которое сообщает, что зависимая от него пропозиция есть результат абдукции, логического вывода, который находит для наблюдаемого положения вещей наилучшее объяснение. Ограничения на дистрибуцию *я вижу* в значительной степени представляют собой ограничения на абдукцию, продиктованные соображениями правдоподобия, информативности и вероятности объяснения.

Кроме того, *я вижу* вводит в рассмотрение событие получения информации о зависимой пропозиции. Это событие имеет фиксированную временную локализацию — интервал, содержащий в себе момент речи. Если этот интервал не пересекается с интервалом, вводимым зависимой пропозицией (когда последняя реализована прошедшим или будущим временем), для *я вижу* возникает эффект непрямой засвидетельствованности.

Особый случай — совпадение пропозиции, описывающей наблюдаемое положение вещей, и абдуцируемой пропозиции, приводящее к тривиальной абдукции. В этом случае единственным нетривиальным компонентом содержания высказывания становится указание на получение информации о пропозиции.

Анонимный рецензент «Диалога» хотел бы видеть в разделе выводов «хотя бы какие-то попытки оценить практическую применимость и масштабируемость полученных результатов». С удовольствием выполняем пожелание рецензента.

Эта работа — типичный пример масштабируемости результатов, но только не такой масштабируемости, когда находки из области лингвистической семантики получают применение в компьютерной науке, а такой, когда движение происходит в обратном направлении. Исследования абдукции, как и других типов немонотонных рассуждений (non-monotonic reasoning), исходно развивались в работах по вычислительной математике, машинному обучению и ИИ, начиная с 1980-х годов (см., например, [Gabbay, Smets (eds.) 2000]) по нынешнее время (например, [Poole, Mackworth (eds.) 2017]. Благодаря таким работам специалистам по семантике естественного языка пришло осознание того факта, что логическая операция, которую они всегда проводили по департаменту «компьютерщиков», оказывается, не просто применима, а критически необходима для анализа определенных естественноязыковых выражений. Случилось это совсем недавно — в последние пять-шесть лет (характерный пример — Gyarmathy, Altshuler 2020); эта статья — движение в том же направлении.

Нам остается лишь выразить надежду, что именно так и должен выглядеть диалог дисциплин, представители которых каждый год видят друг друга на «Диалоге».

Литература

[1] Anderson L.B. (1986), Evidentials, paths of change, and mental maps: typologically regular asymmetries, Evidentiality: the linguistic coding of epistemology, ed. by W. Chafe, J. Nichols, Ablex, Norwood, pp. 273-312.

[2] Brachman R. J., Levesque H. J. (2004), Knowledge Representation and Reasoning, Morgan Kaufmann, San Francisco.

[3] DeLancey S. (1997), Mirativity: The grammatical marking of unexpected information, Linguistic Typology, Vol. 1, No 1, pp. 33-52.

[4] Douven I. (2016), Abduction. The Stanford Encyclopedia of Philosophy (Summer 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/ sum2017/entries/abduction/>.
[5] Faller M. (2002), Semantics and Pragmatics of Evidentials in Cuzco Quechua, Ph.D. dissertation, Stanford University, Stanford.
[6] Faller M. (2011), A possible worlds semantics for Cuzco Quechua evidentials, Proceedings of SALT 20, pp. 660–683.
[7] Gabbay D. M., Smets Ph. (eds.) (2000), Handbook of Defeasible Reasoning and Uncertainty Management Systems. Algorithms for Uncertainty and Defeasible Reasoning, Springer, Amsterdam.
[8] Gyarmathy Z., Altshuler D. (2020), (Non)culmination by abduction, Linguistics, Vol. 58, pp. 1373-1411
[9] Hobbs J. R. (2004), Abduction in Natural Language Understanding, The Handbook of Pragmatics, ed. by L. Horn and G. Ward, Blackwell, Oxford, pp. 724–741.
[10] Josephson J. R., Josephson S. G. (eds.) (1996), Abductive inference: Computation, philosophy, technology, Cambridge University Press, Cambridge.
[11] Koev T. (2011), Evidentiality and temporal distance learning, Proceedings of SALT 21.
[12] Koev T. (2013), Apposition and the structure of discourse, Ph. D. thesis, Rutgers University, New Brunswick..
[13] Korotkova N. (2016), Heterogeneity and uniformity in the evidential domain, Ph. D thesis, University of California, Los Angeles.
[14] Lazard G. (1996), Le médiatif en persan, L'énonciation médiatisée, ed. by Z. Guentchéva, Peeters, Paris and Louvain, pp. 21-30.
[15] Matthewson L., Davis H., Rullmann H. (2007), Evidentials as epistemic modals: evidence from St'˝at'imcets, Linguistic Variation Yearbook 7.
[16] McIlraith Sh. (1998), Logic–based abductive inference, Technical Report Number KSL-98-19, Knowledge Systems Laboratory, Stanford University, Stanford.
[17] Meydan M. (1996), Les emplois médiatifs de –miş en turc, L'énonciation médiatisée, ed. by Z. Guentchéva, Peeters, Paris and Louvain, pp. 125-143.
[18] Murray S.E. (2010), Evidentiality and the structure of speech acts. Ph.D. thesis, Rutgers University, New Brunswick.
[19] Murray S. E. (2016), Evidentiality and illocutionary mood in Cheyenne, International Journal of American Linguistics.
[20] Poole D., Mackworth A. (eds.) (2017), Artificial Intelligence: Foundations of Computational Agents, Cambridge University Press, Cambridge.
[21] Portner P. (2018), Mood, Oxford University Press, Oxford.
[22] Smirnova A. (2011), Evidentiality, mood, and epistemic modality: evidence from Bulgarian. Ph.D. thesis, The Ohio State University.
[23] Smirnova A. (2012), Evidentiality in Bulgarian: Temporality, epistemic modality, and information source, Journal of Semantics, Vol. 30, pp. 479–532.
[24] Maisak T, Tatevosov S. Beyond evidentiality and mirativity: Evidence from tsakhur, L'Énonciation médiatisée II. Le traitement épistémologique de l'information: illustrations amérindiennes et caucasiennes, Vol. 63 of Bibliothèque de l'Information Grammaticale, Leuven, pp. 377–406.
[25] Tatevosov S. (2019), Evidentiality and abduction [Evidentsialnost' i abductsija], 85th Birthday Festschrift in honor of V.S. Khrakovsky [Sbornik statej k 85-letiju V.S. Khrakovskogo], ed. by D. Gerasimov, S. Dmitrenko, N. Zaika, Jazyki slavjanskikh kultur, Moscow, pp. 463–495.
[26] Willett T. (1988). A cross-linguistic survey of the grammaticization of evidentiality, Studies in Language, Vol. 12, No 1, pp. 51–97.
[27] Woodbury A.C. (1986), Interactions of tense and evidentiality: a study of Sherpa and English, Evidentiality: the linguistic coding of epistemology, ed. by W. Chafe, J. Nichols, Ablex, Norwood, pp. 188-202.

Meta-Embeddings in Taxonomy Enrichment Task

Tikhomirov M. M.
Lomonosov Moscow State University
Moscow, Russia
tikhomirov.mm@gmail.com

Loukachevitch N. V.
Lomonosov Moscow State University
Moscow, Russia
louk_nat@mail.ru

Abstract

In this paper we consider the taxonomy enrichment task based on a recently appeared dataset, called Diachronic wordnets, created on the basis of English and Russian wordnets. We study meta-embeddings approaches, which combine several source embeddings, to the hypernym prediction of novel words and show that meta-embedding approaches obtain the best results for this task if compared to other methods based on different principles. When combining with automatically extracted features from the Wiktionary online dictionary, the joint approach improves the results.

Key words: taxonomy enrichment, WordNet, meta-embeddings
DOI: 10.28995/2075-7182-2021-20-681-691

Мета-эмбеддинги в задаче пополнения таксономии

Тихомиров М. М.
МГУ имени М. В. Ломоносова
Москва, Россия
tikhomirov.mm@gmail.com

Лукашевич Н. В.
МГУ имени М. В. Ломоносова
Москва, Россия
louk_nat@mail.ru

Аннотация

В данной статье рассматривается задача обогащения таксономии на базе недавно появившегося набора данных, называемого Diachronic wordnets, созданного на основе английских и русских тезаурусов типа WordNet. Исследуется подход с использованием мета-эмбедингов, которые объединяют в себе разные векторные представления, для решения задачи предсказания гиперонимов. В статье показано, что подходы на основе мета-эмбедингов дают лучшие результаты на данной задаче по сравнению с другими методами, основанными на иных принципах. При комбинировании с автоматически извлеченными признаками из онлайн-словаря Викисловарь совместный результат становится еще лучше.

Ключевые слова: обогащение таксономии, ворднет, мета-эмбедингы

1 Introduction

Various algorithms in natural language processing use vector representations of words, so the quality of the corresponding vectors is essential. There are many different word embeddings, which performance vary for various tasks. Different methods for constructing vectors capture the context in different ways, and can be trained on different datasets, resulting in a wide variety of vector models available. It has been shown that combining word embeddings can improve the accuracy of dependency parsing [2], classification in healthcare [20], named-entity recognition [31, 21], sentiment analysis [21].

Among lexical semantics settings, meta-embeddings were tested in word analoqy and similarity tasks [33, 5, 4] but they were not applied to hypernym detection and taxonomy enrichment tasks, to the best of our knowledge.

Hyponym-hypernym relations ("is_a relations") constitute the backbone structure of many different ontological and lexical-semantic resources. Therefore numerous studies are devoted to the task of extracting hypernym relations from text collections. Hypernyms can be extracted from scratch, without any target resource or taxonomy. But also the hypernym extraction task can be set as a task of searching hypernyms for new words in an existing taxonomy, that is as a taxonomy enrichment task.

IN 2016, the taxonomy enrichment task was organized as a shared task at SemEval workshop (task 14) [15]. At this task, the participants should to attach words with definitions to correct hypernyms in WordNet [19]. However, in real applications definitions of novel words and their senses are most likely absent. In 2020, a new open evaluation on taxonomy enrichment of the Russian wordnet RuWordNet [6] RUSSE'2020 was organized [24]. The task was to find correct hypernyms from an older RuWordNet version for words described in a newer RuWordNet version. The work [29] describes new datasets called Diachronic wordnets[1] created on the basis of English and Russian wordnets. These datasets contain new words added to later versions of wordnets in comparison to earlier version together with their hypernyms in older versions. In such a way it is possible to use different versions of wordnets in their historical development to evaluate methods for the taxonomy enrichment task.

In this paper we show that meta-embedding approaches obtain the best results for the taxonomy enrichment task on the Diachronic-wordnets dataset if compared to other methods based on different principles. When combining with automatically extracted features from the Wiktionary online dictionary, the joint approach improves the results.

2 Related Work

2.1 Hypernym Detection Approaches

Traditional methods for hypernym detection include pattern-based methods, searching for specific hypernym patterns in sentences [14, 25, 26], methods based on similarity of word vector representations [17, 9], and also combined approaches integrating various context and similarity features of words [28, 3, 27].

In the RUSSE-2020 evaluation [24], in which the task was to predict RuWordNet hypernym synsets for new words, the participants used various word embeddings (static – fastText [10], word2vec [8], and contextualized - BERT [1]), the available RuWordNet taxonomy structure, hypernym and co-hyponym patterns, definitions of words from Wiktionary, and global search engines results [32, 7, 30, 24].

Recent methods to hypernym extraction exploit graph-based representations of taxonomy structure. Liu et al. [23] use node2vec embeddings of graph structures [13] for taxonomy induction. Aly et al. [12] use hyperbolic Poincare embeddings [22] for automatic generation of taxonomies. In [11], the authors study graph-based representation methods on the Diachronic-wordnets dataset.

2.2 Meta-Embeddings

The most simple vector combining approaches are concatenation or averaging of some source embeddings. The authors of [5] reported that even such simple combined embeddings could significantly improve the overall performance for several tasks. It has been shown [33] that using singular value decomposition (SVD) can also show good results with the ability to control the final dimension of vectors. Autoencoders [4], called Autoencoded Meta-Embeddings (AEME), became a further development of the idea of creating meta-embeddings. The authors of [4] proposed several different algorithms (CAEME, AAEME and etc) for combination various word vectors in one vector by encoding initial vectors in some meta-embedding space and then decoding backward.

At the first step in the CAEME approach, all word vectors are encoded into meta-vectors and then concatenated. Each vector is encoded into a vector of the same size, and the overall shape of the meta-vector is the sum of the original vector dimensions. Then, the decoding step uses a concatenated representation to predict the original vector representations. The AAEME approach is very similar to CAEME, except that each vector is mapped to a fixed-size vector and all encoded representations are averaged, but

[1] https://github.com/skoltech-nlp/diachronic-wordnets

not concatenated. An obvious advantage of this approach is the ability to control the meta-embedding dimension.

For any AEME approach, different loss functions can be used at the decoding stage: MSE loss, KL-divergence loss, cosine distance loss and also their combinations. In [21] the authors investigated the performance of the autoencoders depending on the loss function. They found that there is no evident winner across tasks and that different loss functions should be chosen for different applications.

3 Datasets and Evaluation Measure

The Diacronic wordnets collection consists of two diachronic datasets: one for English, another one for Russian based respectively on Princeton WordNet [18] and RuWordNet taxonomies [6]. Each dataset contains a taxonomy and a set of novel words to be added to this resource. The statistics are provided in Table 1.

Table 1: Datasets statistics.

Dataset	Nouns	Verbs
WordNet1.6 - WordNet3.0	17 043	755
WordNet1.7 - WordNet3.0	6 161	362
WordNet2.0 - WordNet3.0	2 620	193
RuWordNet1.0 - RuWordNet2.0	14 660	2 154
RUSSE'2020	2 288	525

To compile the English dataset, two versions of WordNet were chosen and then words, which appear only in a newer version, were selected. For each novel word, its hypernyms from the newer WordNet version were extracted and considered as gold standard hypernyms. Novel words were added to the dataset if only their hypernyms appear in both versions. Only nouns and verbs were considered. Several datasets by skipping one or more WordNet versions were created (Table 1).

As gold standard hypernyms, not only the immediate hypernyms of each lemma were considered but also the second-order hypernyms: hypernyms of the hypernyms. The aim was to make the evaluation less restricted but to keep it quite precise.

In order to create an analogous version to English dataset for Russian, a Russian WordNet counterpart, the RuWordNet taxonomy [6] was used. The current version of RuWordNet and also the extended version of RuWordNet, which has not been published yet, were taken to compile the dataset (cf. Table 1).

Another variant of the Russian dataset includes the dataset created for RUSSE'2020 Dialog evaluation [24]. The RUSSE'2020 can be viewed as a restricted subset of the Russian dataset, because of exclusion from it several word categories (short words, diminutive forms, geographic and personal names, and others).

The task of thesaurus enrichment is treated as a ranking task where the correct answers should be in the top of a candidate list. For evaluation, a traditional measure for ranking tasks Mean Average Precision measure is used.

$$MAP = \frac{1}{N}\sum_{i=1}^{N} AP_i;$$
$$AP_i = \frac{1}{M}\sum_{i}^{n} prec_i \times I[y_i = 1], \tag{1}$$

where N and M are the number of predicted and ground truth values, respectively, $prec_i$ is the fraction of ground truth values in the predictions from 1 to i, y_i is the label of the i-th answer in the ranked list of predictions, and I is the indicator function.

To account for second-order hypernyms, the modified MAP measure is used. It transforms a list of gold standard hypernyms into a list of connected components. Each of these components includes hypernyms (both direct and second-order) which form a connected component in a taxonomy graph. To obtain the highest score, the model should guess a hypernym from each connectivity component [24].

4 Creating Meta-Embeddings

In our work we compare simple meta-embeddings such as concatenation of source embeddings and SVD over the concatenation and two varians of autoencoders generating meta-embeddings: Concatenated Autoencoded Meta-Embeddings (CAEME) and Averaged Autoencoded Meta-Embeddings (AAEME), which have shown good results in previous works [4].

Suppose we have two source embeddings $s_1(w)$ and $s_2(w)$, their encoders $E_1(w)$ and $E_2(w)$ and their decoders $D_1(w)$ and $D_2(w)$. Meta-embedding $m(w)$ in CAEME is constructed as the L_2-normalised concatenation of two encoded source embeddings $E_1(s_1(w))$ and $E_2(s_2(w))$:

$$m(w) = \frac{E_1(s_1(w)) \oplus E_2(s_2(w))}{||E_1(s_1(w)) \oplus E_2(s_2(w)||_2} \tag{2}$$

where $\oplus$ is the concatenation operation.

In CAEME, the dimensionality of the meta-embedding space is the sum of the dimensions of the source embeddings. The AAEME encoder can be seen as a special case of the CAEME encoder, where the meta-embedding is computed by averaging the two encoded sources in (2) instead of their concatenation. Averaging gives the possibility to avoid increasing the dimensionality of the meta-embedding.

The AAEME encoder computes the meta-embedding of a word w from its two source embeddings $s_1(w)$ and $s_2(w)$ as the L_2-normalised sum of two encoded versions of the source embeddings $E_1(s_1(w))$ and $E_2(s_2(w))$.

$$m(w) = \frac{E_1(s_1(w)) + E_2(s_2(w))}{||E_1(s_1(w)) + E_2(s_2(w))||_2} \tag{3}$$

The CAEME and AAEME decoders reconstruct the source embeddings from the same meta-embedding $m(w)$, thereby implicitly using both common and complementary information in the source embeddings. The constructed source embeddings look as follows:

$$\begin{aligned} \hat{s}_1(w) &= D_1(m(w)) \\ \hat{s}_2(w) &= D_2(m(w)) \end{aligned} \tag{4}$$

The overall objective of autoencoder training is given below. Function f can be any distance or similarity measure as MSE, KL-divergence, or cosine distance.The coefficients λ_1 and λ_2 can be used to give different emphasis to the reconstruction of the two sources.

$$Loss_w(E_1, E_2, D_1, D_2) = \sum_w (\lambda_1 f(s_1(w), \hat{s}_1(w)) + \lambda_2 f(s_2(w), \hat{s}_2(w))) \tag{5}$$

Jointly learning of E_1, E_2, D_1, D_2 minimises the total reconstruction error given by Equation 5.

To obtain meta-embedding representations after training, only the encoders are applied, which convert the input source embeddings into a meta representation. Further, these meta-embedding vectors are used as vector representations of words.

5 Meta-Embeddings in Taxonomy Enrichment Task

In our approach, we use embeddings (source embeddings or meta-embeddings) to generate a list of most similar taxonomy entries (words or phrases from the taxonomy) to the target word according to cosine similarity. For each target word, the top 20 taxonomy entries are considered. The number of elements for consideration was chosen experimentally. For each entry in the similarity list, all correspondings synsets, their direct and second-order hypernyms are extracted from the taxonomy. They are considered as candidate synsets to be hypernyms of the target word.

For candidate hypernyms synsets, several features are calculated. Logistic regression is used to predict the probability of a candidate to be a hypernym of the target word. The calculated features are as follows:

- maximal, minimal and average similarity between the target word and synonyms in a candidate synset;

- similarity values between the the target word and synonyms in hyponym synsets of the candidate synset. At first, maximal, minimal and average similarity values are calculated for the target and synonyms in each hyponym synset. Then maximal, minimal and average similarity values are calculated over all hyponym synsets for a given candidate synset;
- positional feature (0, 1, 2): if a candidate is one of the synset of a taxonomy entry or it is its direct or second-order hypernym.
- the number of occurrences of the synset in the candidate list.

In total, 17 features were calculated.

Training data were generated randomly and automatically from the published RuWordNet version and WordNet-1.6 (thus, the data do not contain test data).

Table 2: Training Datasets sizes

Language	Part of speech	
	Nouns	Verbs
English	2931	1532
Russian	2990	814

Table 3: MAP scores for the taxonomy enrichment methods for the English datasets.

method	nouns			verbs			vector
	1.6-3.0	1.7-3.0	2.0-3.0	1.6-3.0	1.7-3.0	2.0-3.0	dim
fastText	0.300	0.346	0.396	**0.290**	0.224	**0.280**	300
word2vec	0.226	0.242	0.265	0.091	0.114	0.150	300
GloVe	0.261	0.290	0.326	0.182	0.145	0.175	300
concat	0.308	0.344	0.387	0.273	0.206	0.247	900
SVD	0.308	0.358	0.406	0.286	0.222	0.271	600
CAEME	0.309	0.347	0.395	0.252	0.189	0.260	900
CAEME triplet loss	0.322	0.367	**0.416**	0.287	0.218	0.270	900
AAEME	0.318	0.354	0.401	0.283	0.218	0.254	600
AAEME triplet loss	**0.333**	**0.373**	**0.416**	0.**289**	**0.227**	0.274	600
Previous results based on embeddings:							
fastText [29]	–	–	0.339	–	–	0.213	–
Poincaré embeddings[12, 11]	0.059	0.066	0.101	0.126	0.066	0.109	–
node2vec [13, 11]	0.194	0.219	0.155	0.151	0.109	0.147	–
GCN autoencoder [16, 11]	0.157	0.175	0.168	0.109	0.094	0.117	–

The quality of the approach was evaluated using different source vector representations: fastText[2], word2vec[3], GloVe[4]. Also different meta-embedding approaches were investigated: concatenation, SVD over concatenation, CAEME, AAEME. The standard loss function for AEME approaches we used was cosine distance loss. We have tried variations and combinations between MSE loss, KL divergence loss and cosine distance loss, and last one works best in our case.

A specific feature of the English datasets is the presence of a significant number of multi-word expressions. In order to obtain vectors for such cases, the following procedures were carried out:

- For FastText, embeddings if not in the vocabulary, were obtained in a natural way, by calculating vectors by the model itself,
- For Word2Vec and GloVe, embeddings were calculated by averaging the vectors of maximum prefixes for the constituent words of a multi-word expressions. There is a limitation on the minimum length of a prefix word, which is 4 characters,

[2]Common Crawl English and Russian versions from https://fastText.cc/docs/en/crawl-vectors.html
[3]Araneum for Russian and Gigaword for English from http://vectors.nlpl.eu/repository/
[4]Common Crawl 840b tokens from https://nlp.stanford.edu/projects/GloVe/

- For meta-embeddings approaches, if there is no vector for a word in any model, the corresponding source vector was initialized with zeros.

Another direction of experiments relates to the additional restrictions to the generated meta-embeddings in the AEME algorithms such as the triplet loss. We suppose that the word should be closer to its semantically related words according to the taxonomy than to a randomly chosen word. The algorithm of calculating the triplet loss is as follows:

1. For each word that is present in the taxonomy, a list of semantically related words is compiled from synonyms, hyponyms and hypernyms,
2. In each epoch, we randomly select K positive words from this related set and K negative words from the vocabulary,
3. If the word is not presented in the taxonomy, then the noisy variations of the original vector are positive vectors,
4. Next, triplet margin loss is calculated: the triplet loss is combined with the original loss as $\alpha * loss + (1 - \alpha) * triplet\_loss$, we used $\alpha = 0.5$,

The results of the experiments are given in Table 3 and Table 4. It can be seen that the hypernym prediction for verbs is much worse than for nouns in all datasets. SVD applied to the concatenation of initial embeddings always improves the results compared to the concatenation. It obtains better results if compared to source embeddings in most datasets, except English verbs. SVD achieves the best quality of hypernym prediction among all considered meta-embeddings on the Russian verb datasets.

Table 4: MAP scores for the taxonomy enrichment methods for the Russian datasets.

method	nouns		verbs		vector
	non-restricted	restricted	non-restricted	restricted	dim
fastText	0.416	0.537	0.318	0.418	300
word2vec	0.276	0.526	0.231	0.272	600
concat	0.401	0.563	0.337	0.423	900
SVD	0.435	0.579	**0.382**	0.442	600
CAEME	0.468	0.579	0.352	0.433	900
CAEME triplet loss	0.470	**0.580**	0.350	0.420	900
AAEME	0.466	0.577	0.350	0.432	600
AAEME triplet loss	**0.474**	**0.581**	0.375	0.439	600
Previous results based on embeddings:					
RUSSE Top-1 for verbs: [7]	0.288	0.418	0.340	**0.448**	–
Poincaré embeddings [12, 11]	0.143	0.252	0.105	0.140	–
node2vec [13, 11]	0.266	0.366	0.168	0.252	–
GCN autoencoder [16, 11]	0.183	0.261	0.095	0.141	–

The AAEME autoencoder with the cosine loss is comparable to the CAEME autoencoder in the Russian data, and better on the English datasets. Among meta-embeddings based on autoencoders, the best results of hypernym prediction are obtained with the AAEME autoencoder with triplet loss. This approach improves the results of hypernym prediction in almost all cases compared to initial fastText embeddings, except English verbs. Triplet loss autoencoders in most cases achieve better results than corresponding cosine loss autoencoders.

If compared to other approaches based only on various vector representations, our applied methods obtained the best results on all datasets, except Russian restricted verbs where the results are slightly worse.

6 Improving Meta-Embeddings Results Using Wiktionary

Some previous approaches [32, 24] were tested in hypernym prediction combining embeddings with automatic Wiktionary analysis.

Each Wiktionary page usually comprises a definition and lists of hypernyms, hyponyms and synonyms, which could be useful for our task. Similar to other approaches, we implement the following Wiktionary

features:

- the candidate is present in the Wiktionary hypernyms list for the input word (binary feature),
- the candidate is present in the Wiktionary synonyms list (binary feature),
- the candidate is present in the Wiktionary definition (binary feature),

We do not use the definitions directly, as their texts are too noisy. They often include example usages of words, which cannot be separated from the definitions and can distort their vector representations. This processing generate additional above-mentioned features, which are added to the described logistic regression classifier.

Tables 5 and 6 shows the results of combined approaches. Adding Wiktionary features improved the achieved results on all datasets. If compared to previous approaches used Wiktionary, we obtained the best results on the Diachronic wordnet datasets.

Table 5: MAP scores for the taxonomy enrichment methods for the English datasets with wiktionary features.

method	nouns			verbs		
	1.6-3.0	1.7-3.0	2.0-3.0	1.6-3.0	1.7-3.0	2.0-3.0
fastText	0.319	0.373	0.424	**0.296**	0.231	**0.288**
word2vec	0.236	0.259	0.288	0.090	0.12	0.149
GloVe	0.277	0.315	0.357	0.189	0.153	0.190
concat	0.326	0.370	0.425	0.275	0.213	0.259
SVD	0.324	0.380	0.437	0.289	0.228	0.276
CAEME	0.327	0.373	0.430	0.262	0.205	0.264
CAEME triplet loss	0.339	0.39	0.44	0.292	0.225	0.277
AAEME	0.334	0.379	0.432	0.286	0.225	0.267
AAEME triplet loss	**0.345**	**0.394**	**0.445**	0.289	**0.239**	0.272
Previous results with wiki:						
fastText[29]	0.337	0.380	0.344	0.267	0.200	0.237
fastText + node2vec [11]	0.313	0.380	0.340	0.259	0.195	0.200
fastText+node2vec + Poincaré [11]	0.311	0.350	0.300	0.251	0.177	0.248

Table 6: MAP scores for the taxonomy enrichment methods for the Russian datasets with wiktionary features (wiki).

method	nouns		verbs	
	non-restricted	restricted	non-restricted	restricted
fastText	0.436	0.579	0.366	0.445
word2vec	0.292	0.574	0.258	0.295
concat	0.421	0.598	0.388	0.474
SVD	0.452	**0.612**	0.423	0.458
CAEME	0.485	**0.614**	0.400	0.472
CAEME triplet loss	0.486	0.607	0.402	0.463
AAEME	0.484	**0.611**	0.403	**0.486**
AAEME triplet loss	**0.490**	**0.611**	**0.427**	0.471
Previous results with wiki:				
RUSSE Top-1 for nouns [29, 24]	0.393	0.552	0.293	0.436
fastText+wiki [29]	0.413	0.551	0.297	0.389
fastText + node2vec + Poincaré+wiki [11]	0.414	0.560	0.306	0.391

7 Analysis of Results

It can be seen from the achieved results that having a large taxonomy, automatic methods are able to predict correct taxonomy hypernyms for novel words within three first positions of the ranked list of candidates on average. For the Russian dataset, the correct predictions could be found among the top-2 candidates on average. For the majority of novel words, the correct hypernyms appear within the top-10 candidates.

To analyze the usefulness of the obtained hypernym predictions, we calculated the proportions of words having at least one correct hypernym withing N first positions of the candidate list. Table 7 shows the obtained proportions of words with the first correct answer at Top-N positions.

It can be seen that current results do not allow using for fully automatic taxonomy enrichment because the predictions of correct answers on the first position are still quite low. But the predicted results can significantly facilitate the work of lexicograhers or knowledge engineers.

We analysed hypernym predictions for words not included in the Top-10 of correct answers and found the following cases:

- Predicted hypernyms correspond to senses missed in the taxonomy. For example, word *vechernitsa* is described in RuWordNet only in the sense of 'student of evening education', but also this word can mean 'bat' (noctule bat) or 'flowering plant'. Predicted hypernyms include synset corresponding to the flower synset at the 3d position of the candidate list;
- in many cases predictions are very semantically close but not correct. For example, for word 'gugl' (Google) the correct answers are synsets 'search engine' and 'global search engine'. The predicted hypernyms are 'computer program', 'internet-technology', 'internet-site', 'IT-technology', etc. For word "datacentr" (data-center), which is a specialized building, the prediction at the first position is the 'computer" synset.
- there are also numerous examples when too general hypernyms are predicted;
- in some cases predicted hypernyms are very far from reasonable answers and are difficult for explanation.

Table 7: First correct answer at top N positions. The results are give for nouns based on AAEME triplet-loss embeddings with Wiktionary.

	English %			Russian %	
	1.6-3.0	1.7-3.0	2.0-3.0	non-restricted	restricted
Top-1	28	32	38	42	52
Top-3	41	45	51	57	75
Top-5	47	52	56	64	81
Top-10	54	59	63	70	87

8 Conclusion

In this paper we considered methods for combining different word embeddings in a single meta-embedding. We studied the meta-embedding approach in the taxonomy enrichment task based on several versions of English and Russian wordnets. The meta-embedding methods included concatenation of initial embeddings, SVD over the concatenation, two variants of autoencoders aimed to learn better word embeddings from initial vectors.

We showed that the use of meta-embeddings improves the performance of the system for almost all datasets, except English verbs. SVD always improves the results compared to concatenation. Autoencoder-based meta-embeddings achieve the best results in most cases. It can also be seen that adding the triplet loss improves the results.

We also experimented with a method of joint using meta-embeddings and information about word described in the Wiktionary electronic dictionary, which improved the results.

Acknowledgements

The participation of M. Tikhomirov in the reported study was funded by RFBR, project number 19-37-90119. The work of Natalia Loukachevitch in the current study (preparation of data for the experiments) is supported by the Russian Science Foundation (project 20-11-20166).

References

[1] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // 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). — Minneapolis, Minnesota : Association for Computational Linguistics, 2019. — Jun. — P. 4171–4186. — Access mode: https://www.aclweb.org/anthology/N19-1423.

[2] Bansal Mohit, Gimpel Kevin, Livescu Karen. Tailoring continuous word representations for dependency parsing // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — 2014. — P. 809–815.

[3] Bernier-Colborne Gabriel, Barriere Caroline. Crim at semeval-2018 task 9: A hybrid approach to hypernym discovery // Proceedings of the 12th international workshop on semantic evaluation. — 2018. — P. 725–731.

[4] Bollegala Danushka, Bao Cong. Learning word meta-embeddings by autoencoding // Proceedings of the 27th international conference on computational linguistics. — 2018. — P. 1650–1661.

[5] Coates Joshua, Bollegala Danushka. Frustratingly Easy Meta-Embedding–Computing Meta-Embeddings by Averaging Source Word Embeddings // arXiv preprint arXiv:1804.05262. — 2018.

[6] Creating Russian wordnet by conversion / Natalia V Loukachevitch, German Lashevich, Anastasia A Gerasimova et al. // Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue. — 2016. — P. 405–415.

[7] Dale David. A simple solution for the Taxonomy enrichment task: Discovering hypernyms using nearest neighbor search // Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”. — 2020.

[8] Distributed Representations of Words and Phrases and their Compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in Neural Information Processing Systems 26 / Ed. by C. J. C. Burges, L. Bottou, M. Welling et al. — Curran Associates, Inc., 2013. — P. 3111–3119.

[9] Do supervised distributional methods really learn lexical inference relations? / Omer Levy, Steffen Remus, Chris Biemann, Ido Dagan // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2015. — P. 970–976.

[10] Enriching Word Vectors with Subword Information / Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — P. 135–146.

[11] Evaluation of Taxonomy Enrichment on Diachronic WordNet Versions . / Irina Nikishina, Alexander Panchenko, Varvara Logacheva, Natalia Loukachevitch // Proceedings of the 11th Global WordNet conference GWC-2021. — 2021.

[12] Every Child Should Have Parents: A Taxonomy Refinement Algorithm Based on Hyperbolic Term Embeddings / Rami Aly, Shantanu Acharya, Alexander Ossa et al. // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4811–4817. — Access mode: https://www.aclweb.org/anthology/P19-1474.

[13] Grover Aditya, Leskovec Jure. node2vec: Scalable feature learning for networks // Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. — 2016. — P. 855–864.

[14] Hearst Marti A. Automatic acquisition of hyponyms from large text corpora // Coling 1992 volume 2: The 15th international conference on computational linguistics. — 1992.

[15] Jurgens David, Pilehvar Mohammad Taher. SemEval-2016 Task 14: Semantic Taxonomy Enrichment // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). — San Diego, California : Association for Computational Linguistics, 2016. — Jun. — P. 1092–1102. — Access mode: https://www.aclweb.org/anthology/S16-1169.

[16] Kipf Thomas N, Welling Max. Semi-supervised classification with graph convolutional networks // arXiv preprint arXiv:1609.02907. — 2016.

[17] Learning semantic hierarchies via word embeddings / Ruiji Fu, Jiang Guo, Bing Qin et al. // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2014. — P. 1199–1209.

[18] Miller George A. WordNet: a lexical database for English // Communications of the ACM. — 1995. — Vol. 38, no. 11. — P. 39–41.

[19] Miller George A. WordNet: An electronic lexical database. — MIT press, 1998.

[20] Mixed Pooling Multi-View Attention Autoencoder for Representation Learning in Healthcare / Shaika Chowdhury, Chenwei Zhang, Philip S Yu, Yuan Luo // arXiv preprint arXiv:1910.06456. — 2019.

[21] Neill James O', Bollegala Danushka. Meta-embedding as auxiliary task regularization // arXiv preprint arXiv:1809.05886. — 2018.

[22] Nickel Maximilian, Kiela Douwe. Poincar\'e embeddings for learning hierarchical representations // arXiv preprint arXiv:1705.08039. — 2017.

[23] On interpretation of network embedding via taxonomy induction / Ninghao Liu, Xiao Huang, Jundong Li, Xia Hu // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. — 2018. — P. 1812–1820.

[24] RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian Language / Irina Nikishina, Varvara Logacheva, Alexander Panchenko, Natalia Loukachevitch // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". — 2020.

[25] Roller Stephen, Kiela Douwe, Nickel Maximilian. Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — 2018. — P. 358–363.

[26] Sabirova Kristina, Lukanin Artem. Automatic Extraction of Hypernyms and Hyponyms from Russian Texts. // AIST (Supplement). — 2014. — P. 35–40.

[27] Shwartz Vered, Dagan Ido. Path-based vs. Distributional Information in Recognizing Lexical Semantic Relations // COLING 2016. — 2016. — P. 24.

[28] Snow Rion, Jurafsky Dan, Ng Andrew Y. Semantic taxonomy induction from heterogenous evidence // Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. — 2006. — P. 801–808.

[29] Studying Taxonomy Enrichment on Diachronic WordNet Versions / Irina Nikishina, Alexander Panchenko, Varvara Logacheva, Natalia Loukachevitch // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain : Association for Computational Linguistics, 2020. — December.

[30] Tikhomirov Mikhail, LOukachevitch Natalia, Ekaterina Parkhomenko. Combined approach to hypernym detection for thesaurus enrichment // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". — 2020.

[31] Winata Genta Indra, Lin Zhaojiang, Fung Pascale. Learning multilingual meta-embeddings for code-switching named entity recognition // Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). — 2019. — P. 181–186.

[32] Word2vec not dead: predicting hypernyms of co-hyponyms is better than reading definitions / Nikolay Arefyev, Maksim Fedoseev, Andrey Kabanov, Vadim Zizov // Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”. — 2020.

[33] Yin Wenpeng, Schütze Hinrich. Learning word meta-embeddings // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2016. — P. 1351–1360.

Russian News Similarity Detection with SBERT: pre-training and fine-tuning

Vatolin A. S.
SberBank / Moscow, Russia
vatolinalex@gmail.com

Smirnova E. Y.
SberBank / Moscow, Russia
kate15511@gmail.com

Shkarin S. S.
SberBank / Moscow, Russia
kouki.sergey@gmail.com

Abstract

Computation of text similarity is one of the most challenging tasks in NLP as it implies understanding of semantics beyond the meaning of individual words (tokens). Due to the lack of labelled data this task is often accomplished by means of unsupervised methods such as clustering. Within the DE2021: "Russian News Clustering and Headline Selection" we propose a method of building robust text embeddings based on Sentence Transformers architecture, pretrained on a large dataset of in-domain data and then fine-tuned on a small dataset of paraphrases leveraging GlobalMultiheadPooling.

Keywords: text similarity; clustering; sentence embedding; BERT; Sentence Transformers; paraphrase detection
DOI: 10.28995/2075-7182-2021-20-692-697

Поиск похожих новостей на русском языке с помощью SBERT: предварительное обучение и тонкая настройка

Ватолин А. С.
СберБанк / Москва, Россия
vatolinalex@gmail.com

Смирнова Е. Ю.
СберБанк / Москва, Россия
kate15511@gmail.com

Шкарин С. С.
СберБанк / Москва, Россия
kouki.sergey@gmail.com

Аннотация

Вычисление схожести текстов - одна из самых сложных задач в сфере автоматической обработки естественного языка, поскольку подразумевает понимание семантики всего текста, что выходит за рамки значений отдельных слов (токенов). Из-за недостатка размеченных данных эта задача часто решается методами обучения "без учителя", такими как кластеризация. В рамках соревнования DE2021: "Russian News Clustering and Headline Selection" мы предлагаем метод построения векторных представлений (эмбеддингов) текста на основе архитектуры Sentence Transformers, предварительно обученной на большом наборе данных заданного домена (новостные тексты), а затем тонко настроенных на небольшом наборе данных парафразов с использованием GlobalMultiheadPooling.

Ключевые слова: семантическая близость текстов; кластеризация; векторное представление текстов; BERT; Sentence Transformers; нахождение парафразов

1 Introduction

Computation of text similarity is a common, yet challenging task that plays an important role in a variety of Natural Language Processing (NLP) applications, such as search engines, plagiarism detectors, question answering systems, etc.

The main difficulty stems from the variability of human language, as the same meaning can be conveyed with different language units. As a result, we cannot rely on superficial resemblance of texts, such as sharing common words or expressions. Instead, we need to go beyond individual words, capture the semantic meaning of the texts and then evaluate this semantic similarity with some measurable score.

One method to create such a semantic similarity score between texts is based on weighted relations between words in linguistic resources, such as WordNet and Semantic nets (see [10], [12]). This approach has its constraints in the limited size and rigidity of manually constructed thesauri and ontologies.

Another approach is to create special vector representations, or embeddings, that can capture the semantics of a text. One of the successful solutions for creating such vectors for individual words was word2vec, introduced in 2013 [14]. The idea behind it is based on the famous assumption "you shall know a word by the company it keeps". Thus, the initial values of the embeddings are replaced during training in such a way that the words used in the same contexts (and hence similar) are located closer in the vector space.

As an extension of this idea, the doc2vec algorithm was developed in 2014 [9], which added a vector with the document (e.g., text) ID to the word2vec model. This made it possible to obtain embeddings of sentences and texts, which inherit the main feature of word2vec: similar texts are located closer in the vector space. Thus, the similarity between texts can be calculated as the distance between their embeddings (for example, cosine, Euclidean, Manhattan distance, etc.).

Recent advances in deep learning have allowed for the creation of text embeddings that are more powerful in representing the semantics. The Skip-Thought method [8] trains a model to predict surrounding sentences based on the encoder-decoder architecture. InferSent [5] trains a bi-directional Long-Short Term Memory (LSTM) model on the labelled Natural Language Inference (NLI) data, achieving better performance. ELMo [11] introduces contextualized word embeddings, created by training a bi-directional LSTM. Taking the weighted mean of ELMo word embeddings to get a text embedding results in great performance on sentence similarity benchmarks as shown in [11].

Finally, the current state-of-the-art method is to pre-train Transformer models on language modelling (LM) and then fine-tune it for downstream tasks. The best results are achieved by Universal Sentence Encoder models [3], fine-tuned on the NLI task, and Sentence Transformers [15], fine-tuned on the NLI and Semantic text similarity (STS) benchmarks.

There are several pre-trained Sentence Transformer models, as well as the RuBERT model by DeepPavlov, that can be exploited for calculating similarity of texts in Russian. However, to obtain the best results it is necessary to pre-train the transformer model on the in-domain data and then fine-tune the model for a specific task.

2 Shared task description

This paper is the result of the research done for the first track of the Dialogue-2021 shared task 'Russian News Clustering and Headline Generation' [6]. The main purpose of this track is to investigate different approaches to clustering similar news texts in Russian.

Data for this competition was sourced from the Telegram Data Clustering Contest and annotated via Yandex.Toloka. The main dataset represents couples of sentences marked according to whether they are related to the same story (OK) or not (BAD). The same story is considered to be a text in different media aka newspapers, web sources, etc. about a certain event that happened with certain people and at a certain time. Such dataset design essentially turns the clustering task into paraphrase detection. Though a common solution would be to build a classifier on top of BERT model (see Cross-Encoder in [7]), methods such as Bi-Encoders that produce text embeddings instead of classification label were considered more preferable (but not required) by the contest organizers as they are more consistent with the initial idea of the clustering task.

The participants' results were evaluated using f1-score, calculated for positive (OK) class.

3 System description

Transformer-based models, such as BERT, show start-of-the-art performance on most NLP tasks. In the original paper [7], the BERT model uses the cross-encoder approach: two texts are fed to the input and the target value is predicted from them. However, this approach is not optimal when predicting the target value for all text pairs from a large dataset. The complexity can be estimated as $O(n(n-1)/2)$.

Another approach is bi-encoder: the model projects the text into a dense vector space and then uses similarity metrics such as cosine similarity or Manhattan/Euclidean distance to calculate the semantic similarity between the two texts.

What is more, the latter approach is more consistent with the clustering task of the DE 2021 competition. Thus, the simplest way to undertaking the DE2021 shared task is to use multi-language pre-trained bi-encoder model, fine-tuned for paraphrase identification or semantic similarity. The advantage of this method is the speed of work, there is no need to train the model, all it takes to make inference. The disadvantage of this approach is low quality.

The model can be improved in 4 ways: data preparation, base transformer model, pooling method, and loss function.

3.1 MLM pretrain

The improvement over a base BERT model was a pre-training model with masked language modelling (MLM). This is a common way of improving these models for specific tasks. Generally, many researches found that MLM pre-training models are more stable, train faster, and more often reach higher scores than training from the original pretrained model. You can also use other transformer model architectures, such as RoBERTa or XLM-R, but not all models have Russian language support.

As a pooling method, the authors of the SBERT model [15] suggest using averaging over the token vectors of the last layer of the model. They also suggest taking the CLS token vector and max-over-time for output vectors.

3.2 Weighted Layer Pooling

In the article [13], the authors claim that the information useful for solving the problem is contained not only in the last layer, but also in the middle layers. To leverage this information the following operation can be performed: taking mean, max, CLS pooling for the outputs of each layer (output), and then averaging the resulting vectors with the trainable weights w.

$$u = softmax(w) * output$$

3.3 Global Multihead Pooling

In the article [4], to convert token vectors to a fixed-length vector, it is proposed to use the sum of the vectors weighted with attention weights. The authors use this method to aggregate the outputs of the Bi-LSTM model. In our article, we propose to adapt this method to transformers.

$$A = softmax(W_2 ReLU(W_1 output^T + b_1)^T + b_2),$$

where output is a matrix of token vectors, the output of the last layer of the BERT model, W_1, W_2 are trainable weight matrices, b_1, b_2 are bias vectors, and A is a vector of word weights. Because of softmax function, the weights are always non-negative and sum up to 1.

The text vector can be calculated using the following formula:

$$u = A \odot output$$

Usually, the attention weights focus on a specific part of the text, so we can extend the pooling method to a multi-head way:

$$u_i = A_i \odot output$$
$$u = [u_1, u_2, \dots, u_n],$$

where n is the number of heads. Since the text vectors from each head are concatenated, the dimension of the output vector increases significantly with a larger n. So, we used $n \leq 5$.

3.4 Contrastive loss

Contrastive loss takes the output of the network for a positive example and calculates its distance to an example of the same class and contrasts that with the distance to negative examples. To put it another way, the loss is low if positive samples are encoded to similar representations and negative examples are encoded to different representations.

$$loss = y\,d^2 + (1-y)(margin - d, 0)^2$$

3.5 Online contrastive loss

Contrastive loss modification, where loss computed only for hard positive and hard negative pairs.

4 Experimental setup

4.1 Data

The initial data for the competition were presented in the format of HTML files containing articles in various languages. Each article consists of a title, text, and additional metadata.

After preprocessing, including the removal of non-news and non-Russian texts, three datasets were generated - 20,000 labeled news pairs in one day for training, 40,000 unlabeled news pairs in 2 days for testing, and about 1,200,000 raw news texts for pre-training. The average text length is about 1500 characters, including spaces.

4.2 Experiments

Multilingual transfer learning

As a base model, we used distiluse-base-multilingual-cased-v2 from the sentence transformers library.

Transformer model

We used the Deeppavlov RuBERT base cased model as the optimal model by quality and training speed [16]. We also used the sbert_large_nlu_ru model [1], based on the Russian BERT Large model, but it showed worse quality compared to RuBERT. When training, the length of the texts was limited to 250 tokens, increasing the maximum length reduced the quality of the model. We think that this is due to a decrease in the size of the batch with a longer sequence length, and therefore a decrease in the stability of the gradients. To optimize the parameters, we used AdamW with learning rate 2e-5 and 10% warmup steps.

Pooling method

To aggregate token vectors into a text vector, we tried averaging word vectors, and also used Weighted Layer pooling and Global Multihead attention approaches. The Global Multihead attention layer is best in terms of quality, as it allows to increase the weight of important words and not take into account the words of the general vocabulary.

Loss function

For our better model, we used Online contrastive loss with cosine proximity functions and margin 0.5. On the final epochs, for most batches, the loss was 0. To avoid this problem, we tried to increase the margin to 0.8, but this did not lead to an improvement in quality.

5 Results

We present the results of model evaluation in Table 1. We use the f1 metric to measure quality. We split the train dataset in 70%, 15% and 15% for train, validation and test set accordingly.

Model	val f1	test f1	public f1	private f1
Distiluse v1*	0.8955	0.8846	0.8733	0.8816
Distiluse v2*	0.8974	0.8845	0.8840	0.8763
Deeppavlov RuBERT	0.9538	0.9455	0.9331	0.9343
Deeppavlov RuBERT, Global Multihead pooling	0.9619	0.9601	0.9467	0.9438
sbert_large_nlu_ru	0.9498	0.9415	0.9108	-
Deeppavlov RuBERT pretrained	0.9556	0.9511	0.9435	0.9387
Deeppavlov RuBERT pretrained, cross-encoder	0.9668	0.9656	0.9516	0.9545
Deeppavlov RuBERT pretrained, Global Multihead pooling	**0.9631**	**0.9617**	**0.9547**	**0.9548**

Table 1: models evaluation results

* scores without fine-tuning on news dataset.
All models are trained with mean pooling, unless GlobalMultihead pooling is specified in the name. Our target approach reaches the results of cross-encoder.

6 Conclusion and future work

Sentence similarity calculation is a common task, crucial for many NLP applications. Models based on one of the most cutting-edge architectures - Transformer - show state-of-the-art results in many down-stream tasks, including paraphrase detection. Pretrained multilingual Transformer models show decent quality without any additional training. However, the best scores are achieved by pre-training on in-domain data and follow-up fine-tuning for a specific task (paraphrase detection). Another improvement suggested in this article is the use of GlobalMultihead pooling.

As for future work, we should try SBERT-WK [2] model and in particular WK Pooling. The SBERT-WK model shows a higher quality compared to the SBERT model. The SBERT-WK model uses qr matrix decomposition, which in the Pytorch implementation is very slow on the GPU at the moment. Because of this, model training takes a significant amount of time.

Acknowledgements

We thank the organizers of Shared Task Ilya Gusev and Ivan Smurov for providing the data and holding the competition. We are also grateful to members of the DeepPavlov team for their pretrained BERT models. We thank the anonymous reviewers whose valuable comments helped to improve the paper.

References

[1] BERT large model (uncased) for Sentence Embeddings in Russian language, Access mode: https://huggingface.co/sberbank-ai/sbert_large_nlu_ru
[2] Bin Wang, Jay Kuo C.-C.: SBERT-WK: A Sentence Embedding Method by Dissecting BERT-based Word Models. – 2020. – Vol. arXiv:2002.06652. – Access mode: https://arxiv.org/abs/2002.06652
[3] Cer D. et al.: Universal Sentence Encoder. – 2018. – Vol. arXiv:1803.11175. – Access mode: https://arxiv.org/abs/1803.11175
[4] Chen Q., Ling Z.H., Zhu X.: Enhancing Sentence Embedding with Generalized Pooling. – 2018. – Vol. arXiv:1806.09828. – Access mode: https://arxiv.org/abs/1806.09828
[5] Conneau A. et al.: Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. – 2017. – Vol. arXiv:1705.02364. – Access mode: https://arxiv.org/abs/1705.02364.
[6] DE2021: Russian News Clustering and Headline Selection – Clustering, Access mode: https://competitions.codalab.org/competitions/28830
[7] Devlin J., Chang M.W., Lee K., Toutanova K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. – 2019. – Vol. arXiv:1810.04805. – Access mode: https://arxiv.org/abs/1810.04805.
[8] Kiros R et al.: Skip-Thought Vectors. – 2015. – Vol. arXiv:1506.06726. – Access mode: https://arxiv.org/abs/1506.06726
[9] Le Q.V., Mikolov T.: Distributed Representations of Sentences and Documents. – 2014. – Vol. arXiv:1405.4053. – Access mode: https://arxiv.org/abs/1405.4053.
[10] Li Y. et al. (2006), Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering 18(8), 1138–1150.
[11] Matthew E. Peters et al.: Deep contextualized word representations. – 2018. – Vol. arXiv:1802.05365. – Access mode: https://arxiv.org/abs/1802.05365
[12] Mihalcea R., Corley C., Strapparava C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity. – 2006. – Access mode: https://www.aaai.org/home.html
[13] Mikhailov V., Taktasheva E., Sigdel E., Artemova E.: RuSentEval: Linguistic Source, Encoder Force! – 2021. – Vol. arXiv:2103.00573. – Access mode: https://arxiv.org/abs/2103.00573.
[14] Mikolov T.: Efficient Estimation of Word Representations in Vector Space. – 2013. – Vol. arXiv:1301.3781. – Access mode: https://arxiv.org/abs/1301.3781.
[15] Reimers N, Gurevych I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. – 2019. – Vol. arXiv:1908.10084. – Access mode: https://arxiv.org/abs/1908.10084.
[16] Shavrina T. et al.: RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. – 2020. – Access mode: https://www.aclweb.org/anthology/2020.emnlp-main.381/

Automatic Detection of Deceptive and Truthful Paralinguistic Information in Speech using Two-Level Machine Learning Model

Velichko A.N.
SPC RAS
Saint-Petersburg, Russia
alena.n.velichko@gmail.com

Karpov A.A.
SPC RAS
Saint-Petersburg, Russia
karpov@iias.spb.su

Abstract

In this work, we present a novel approach to one of computational paralinguistic tasks – automatic detection of deceptive and truthful information in human's speech. This task belongs to the aspects of destructive behaviour and was first presented at the International INTERSPEECH Computational Paralinguistics Challenge ComParE in 2016. The need of contactless method for deception detection follows from the fact that existing contact-based approaches such as polygraphs and lie detectors have multiple restrictions, which significantly limit their usage. Both for training and testing of the proposed models we used two English-language corpora (Deceptive Speech Database and Real-Life Trial Deception Detection Dataset). We extracted tree sets of acoustic features from those audio samples using openSMILE toolkit. The proposed approach includes preprocessing of the extracted acoustic features with the usage of methods for data augmentation and dimensionality reduction of feature space. We have got 1680 speech utterances and 986-dimensional informative feature vector for each utterance. The main part of the proposed approach is two-level recognition model, where the first level includes three models of gradient boosting (Catboost, XGBoost and LightGBM). The second level consists of logistic regression-based model for final prediction on truthfulness or deceptiveness that takes into account predictions from the first level. Using this approach, we have achieved the result of classification in terms of F-score = 85.6%. The proposed approach can be used both independently and as a component of multimodal systems for detection of deceptive and truthful utterances in speech, as well as in systems for detection of a destructive behaviour.

Keywords: speech technology, computational paralinguistics, detection of deceptive and truthful information in speech, machine learning, gradient boosting, multimodal systems

DOI: 10.28995/2075-7182-2021-20-698-704

Автоматическое определение ложной и истинной паралингвистической информации в речи человека с применением двухуровневой модели машинного обучения

Величко А.Н.
СПб ФИЦ РАН
Санкт-Петербург, Россия
alena.n.velichko@gmail.com

Карпов А.А.
СПб ФИЦ РАН
Санкт-Петербург, Россия
karpov@iias.spb.su

Аннотация

В работе предложен подход к решению одной из задач компьютерной паралингвистики – автоматическому определению ложной и истинной информации в речи человека. Данная задача является одним из аспектов деструктивного поведения и впервые была представлена на международных соревнованиях по компьютерной паралингвистике INTERSPEECH ComParE в 2016 году. Необходимость бесконтактного метода определения ложной информации в речи вытекает из того, что существующие контактные подходы, например, полиграфы или детекторы лжи, имеют ряд требований, которые значительно ограничивают их использование. Для обучения и тестирования предложенных моделей нами использовались два корпуса англоязычной речи (Deceptive Speech Database и Real-Life Trial Deception Detection Dataset), из аудиозаписей которых были вычислены три набора акустических признаков посредством программного инструментария openSMILE. Предложенный подход включает предобработку акустических признаков при помощи метода аугментации данных и метода уменьшения размерности признакового пространства, что в итоге позволило получить 1680 речевых высказываний, из которых были вычислен 986-размерный вектор

информативных признаков. Основой предложенного подхода является двухуровневая модель, в которой на первом уровне используются три модели градиентного бустинга (Catboost, XGBoost и LightGBM), а на втором уровне – модель на основе логистической регрессии, которая позволяет выдавать итоговое предсказание о ложности или истинности высказывания на основе предсказаний моделей первого уровня. С использованием такого подхода удалось добиться значения результата определения ложной и истинной информации по показателю F-меры, равного 85,6%. Предложенный подход может использоваться как самостоятельно, так и в качестве компонента многомодальной системы определения ложной и истинной информации в речи или системы определения деструктивного поведения.

Ключевые слова: речевые технологии, компьютерная паралингвистика, определение ложности и истинности информации в речи, машинное обучение, градиентный бустинг, многомодальные системы

1 Introduction

The task of automatic detection of deceptive and truthful information in speech belongs to the field of computational paralinguistics as well as detection of mental, emotional and physical states, detection of different diseases (including COVID-19) by voice, speech and noises etc. Moreover, lie is one of the aspects of destructive behaviour that also includes depression, aggression, etc. Nowadays, with the advancement of the Internet and social networks such destructive behaviour is more common to appear in text format [16]. On the other hand, both in virtual and real life people still use natural speech for communication and it is also a study object. In recent years many groups of researches have presented papers addressing contactless deception detection in speech. The reason is that current contact-based methods have multiple restrictions that refer both to a place and a research subject. The concept of automatic deception detection in speech is based on the hypothesis that telling lies have an impact on increasing stress level and it affects acoustic parameters. Additionally, linguistic, para- and extralinguistic factors (such as different psychophysiological states, pathologies of mentality or mental disorders, and some other diseases) have an impact on phonetic characteristics of speech and even possibility of speech production [15].

There are unimodal and multimodal approaches to detection of deceptive and truthful information in speech. Unimodal systems can be used both by itself and as a part of more complex systems for psychophysiological human states. Multimodal systems for automatic deception detection by speech can significantly improve the quality of classification because of their ability to analyze additional paralinguistic aspects. Those aspects include mimics and gestures (eyebrows movements, lips tension, gaze direction, hands movement etc.), they are informative markers for detection of deceptive or truthful speech utterances. Moreover, analysis of lexical component of speech utterance can allow different markers in speech, for example, uncertainty expressed by particular words, hesitations and interjections. Contactless systems can be used in such areas as banking (for example, loan granting), law enforcement (for example, polygraph tests, preventing of “telephone terrorism” etc.).

The first time the task of contactless detection of deceptive and truthful information was introduced was within the framework of the International Computational Paralinguistics Challenge ComParE in 2016. Organizers of the challenge also presented a speech corpus that included deceptive and truthful speech samples, Deceptive Speech Database (DSD) [1], and a set of acoustic features based on the software tool openSMILE [5]. Base system proposed by organizers achieved results in terms of unweighted average recall (UAR) = 68.3%. The winners of the challenge [12] were able to achieve the UAR = 74.9%. They used prosodic features with base acoustic feature set. Later, in 2017 another system was proposed [11]. Authors used acoustic and lexical features with the classifier based on the model of Random Forest. They achieved the result in terms of F-score = 63.9% and Precision = 76.1%. In paper [20], authors proposed methods for the task of data scarcity and imbalanced classes in data for training models. In their system authors used SMOTE method for augmentation of training data with the number of k-nearest neighbours equal 3. This system was based on Support Vector Machines (SVM) and achieved results in term of UAR = 73.5%, mean F-score = 75.0% and Precision = 77.0%. In [21] the implementation of ensemble methods and neural networks were proposed. The ensemble consists of k-Nearest Neighbours, Random Forest and Neural Network have achieved the UAR of 65.0% and 70.0%, in case of average voting and majority voting correspondingly.

Pérez-Rosas et al. [13] proposed a multimodal corpus and a multimodal system for automatic detection of deceptive and truthful information. They used classifiers (Decision Trees and Random Forest) both for verbal and non-verbal features and achieved Accuracy in range of 60.0-75.0%. The same authors later presented another multimodal corpus that included deceptive and truthful samples [18]. The paper also proposed a model of Random Forest that achieved Accuracy = 69.0%. In [23] authors used corpora Box of Lies both for training and testing models. They extracted acoustic features from openSMILE, face markers and linguistic features for training models based on Random Forest. With this system they could achieve Accuracy of classification 73.0%.

Litvinova et al. [9, 10] used lexical and syntactic markers for deception detection in texts. They created a corpus [7] that consists of 226 essays on topic "Describe one day in your life". Every participant was free to answer truthfully or to lie. Besides, the corpus contains information about participants, namely: gender, age, scale of self-esteem, information collected using different psychological tests that can reveal the correlation between language parameters of written texts and personality characteristics of their authors. The mean length of texts is 221 word, all participants (46 men and 67 women) were native Russian speakers. Authors also applied statistic modelling with the use of Linguistic Inquiry and Word Count (LIWC). They proved the hypothesis that chosen marker of lie (relationship of percent of adverbs in text and percent of personal pronouns, thus, in truthful text percent of adverbs decreases and percent of personal pronouns increases) is effective. They achieved the detection of existence/absence of deceptive information with probability of 71.0-72.0%. In [8], after statistical analysis of created corpus authors found out that rates of Accuracy in classification differed for men and women – 73.3% and 63.3% respectively. In [14], authors used three groups of markers: psycholinguistic and sentiment markers, normalized frequencies of 11 Part-of-Speech (POS) tags and bigrams of POS tags, syntactic and readability features. They applied models of Random Forest and SVM. The best result was achieved with the use of SVM and POS tags and bigrams of POS tags. They also took into account the use of conjunctions, interjections and numerals. Such system achieved Accuracy = 57.0% and F-score = 56.0%.

2 Description of methods used in the present approach for automatic deception detection in speech

In order to process audio data, a researcher has to digitize and vectorize an audio signal. Modern automatic systems for paralinguistic analysis of speech use Low Level Descriptors (LLD) that represent spaces of feature vectors with a huge size (several thousands of features). These features are usually presented as feature sets (as in software toolkit openSMILE) and include different spectral, energy and prosodic features such as: fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCC), formants or resonance frequencies of voice tract etc. In addition to those basic features, feature sets include their functionals: mean value, standard deviation, slope and shift, minimum value, relative position of minimum value, maximum value, relative position of maximum value, zero-crossing etc.

Boosting is a compositional machine learning meta-algorithm that is usually used for reduction of bias-variance tradeoff. Gradient boosting is a machine learning method that solves tasks of regression and classification using a composition of models for prediction. Boosting is a model with cascade process of training, where every following model attempts to correct the previous one. First, it creates a subset of data and the weights are equal for all training objects. Then the base model builds on this subset and makes a prediction for all set. After that it computes false predictions and updates weights (they gets bigger values). By this technique it builds other models for other subsets and makes predictions. Final model is the weighted mean of all models. Model of gradient boosting uses this technique of boosting and regression trees as a base algorithm, where every following tree builds on computed errors of previous one.

In the proposed approach we use three methods of gradient boosting: (1) Catboost, (2) XGBoost, (3) LightGBM. XGBoost (eXtreme Gradient Boosting) [19] is an implementation of the gradient boosting algorithm that has regularization function which helps to avoid the overfitting. LightGBM [6] is a faster implementation of gradient boosting and works especially good with big datasets. If compared with other implementations it builds trees in depth, not in breadth. Catboost [4] effectively copes with cat-

egorial variables and decreases time for their preprocessing, and it has a built-in detector of overfitting. Moreover, the approach uses the method of Stacking (Stacked Generalization) [3]. It is one of the ways to create an ensemble of algorithms. The idea is to combine output information of independent algorithms and use a classifier (or regressor) to make a final prediction.

3 Experimental setup

To solve the automatic detection of deceptive and truthful speech utterances task, we used two English-language corpora: (1) speech corpus Deceptive Speech Database and (2) multimodal corpus Real-Life Deception Detection Dataset (RLDDD) [13]. The first one consists of audio samples of students' speech. They played a role either a liar who stole papers from teacher's office, or an honest person. Recordings of the second corpus contain video data collected from public trial courts.

According to the other researches (for example, [22]), we decided to use both corpora simultaneously to increase the number of examples in training data, to get more generalized models and to improve the robustness of our models. Moreover, as it was found in [20], augmentation of data allows to improve results of models. In this work we use only audio features because we have only one corpus that contains video data. Overall number of audio recordings in both corpora is 1253 utterances. Small amount of data for training leads to significant restrictions in selection of machine learning methods and methods of modelling. This is also a reason of using augmentation of training data.

The proposed approach uses a software tool openSMILE to extract acoustic features. We chose three feature sets, namely: INTERSPEECH ComParE 2013 [15], ComParE 2016 (is an updated version of 2013 set) [5] and ComParE 2011 (includes acoustic features that were used in challenge for automatic detection of speaker state) [17]. Overall dimensionality of all sets was more than 12000 features. To balance classes in training data and perform an augmentation the method SMOTE (Synthetic Minority Oversampling Technique) was used. This method applies an algorithm of k-nearest neighbours that creates training objects similar to the minority class objects. Experimentally we found an optimal number of k-nearest neighbours equal 3. Due to the high dimensionality of a feature vector (more than 12000 features) there was a need to use a dimensionality reduction of feature space. To perform it the method of Principal Component Analysis (PCA) was applied. It is an implementation of programming library of Python language, Scikit-learn. As a result, we have got a set of 1680 training objects and 986 informative features. Right before the training process we shuffled the data thus each fold in 10-fold of cross-validation consists of data from both corpora and has similar distribution of classes.

To unite our models, we decided to use a two-level method of stacking, where the first level includes three models of gradient boosting and the second one contains a logistic regression. With the use of this method model on the second level makes predictions based on the results that it receives from the models on the first level. Overall scheme of such system including data preprocessing steps and two-level model for detection of deceptive and truthful information is shown on Fig.1.

For experiments we used methods of gradient boosting from three programming libraries: Catboost, XGBoost and LightGBM. Training and testing were performed with the use of 10-fold cross-validation to control and prevent overfitting. We also activated built-in overfitting detector in the Catboost. Hyper-parameters were selected empirically using grid search. As a quantitative rate F-score [2] and Unweighted Average Recall (UAR) were chosen.

4 Discussion of the results

The trained two-level model was able to achieve the quality of detection of deceptive and truthful information in speech in terms of F-score = 85.6%. For comparison, single models of Cabtoost, XGBoost and LightGBM have achieved results in terms of F-score of 84.1%, 84.6%, and 85.0% respectively. The achieved empirical results are highly competitive and comparable with the results presented by other researchers [11, 12, 17, 18, 21, 20, 23] (see Table 1). We present our results achieved with the use of two corpora and the approach described above. Moreover, the results show that the usage of several feature sets and models of gradient boosting can significantly improve the result in our task. However, as it was stated in [9] and other papers using several modalities can decrease variability in detection of

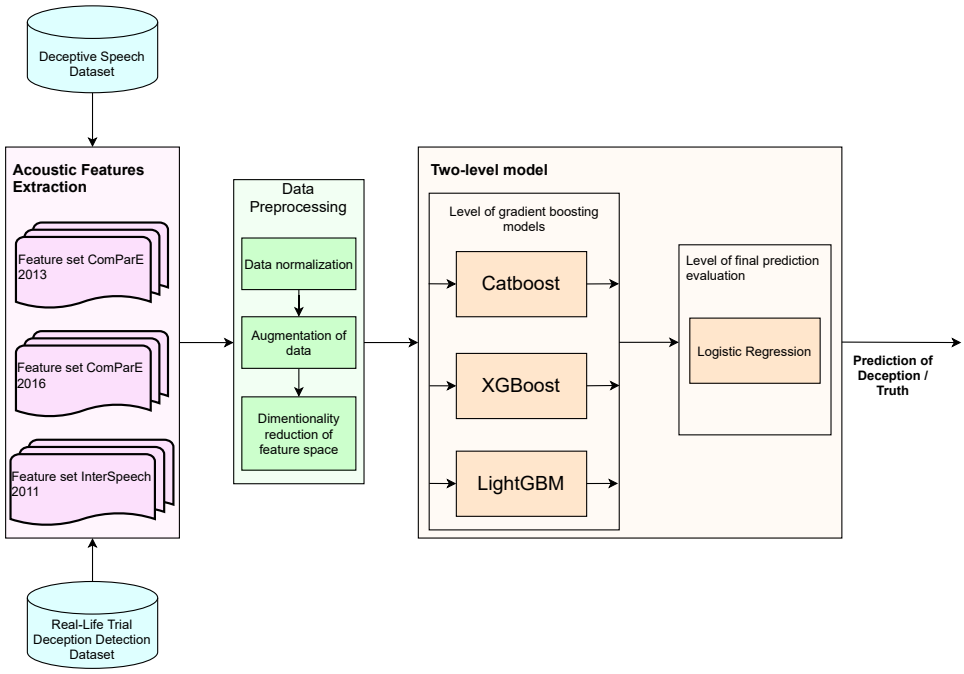


Figure 1: Architecture of the recognition system with two-level architecture for detection of deceptive and truthful information in speech.

deceptive and truthful speech utterances and improve results of the classification. We intend to check this hypothesis in future, namely we plan to analyze both lexical information and video data.

Approach	**Classification results**
Approach from [11]	F-score = 63.9%, Precision = 76.1%
Approach from [12]	UAR = 74.9%
Baseline system [17]	UAR = 68.3%
Approach from [18]	Accuracy (max) = 75.0%
Approach from [21]	UAR (max) = 70.0%
Approach from [20]	UAR = 73.5%, F-score = 75.0%, Precision = 77.0%
Approach from [23]	Accuracy = 73.0%
Catboost	F-score = 84.1%, UAR = 84.0%
XGBoost	F-score = 84.6%, UAR = 84.4%
LightGBM	F-score = 85.0%, UAR = 84.9%
Stacking (Catboost, XGBoost, LightGBM)	F-score = 85.6%, UAR = 85.5%

Table 1: Comparison of the achieved results with other known works.

5 Conclusions

The presented study is dedicated to the task of detection of deceptive and truthful information in speech that belongs to analysis of human's destructive behaviour, and this is a challenging task of computational paralinguistics. Existence of many scientific papers proves the significance of this task especially taking into account the widespread usage of the Internet and social networks. Due to restrictions in usage of contact-based methods contactless methods for deception detection in speech become more important. Since data retrieval for this task is time-consuming and difficult due to specificity of the task, the existing corpora have quite small amount of data and suffer from an imbalance in classes. To cope with these

restrictions, we have used an augmentation method (SMOTE) and a method for reducing feature space (PCA).

In the proposed approach, we have used a two-level model, where the first level includes three gradient boosting algorithms (Catboost, XGBoost, LightGBM) and the second one includes a logistic regression. The final prediction is based on predictions made on the first level. Hyper-parameters were calculated using the grid search method.

In the experiments, the proposed approach has achieved the quality of deception detection in terms of F-score = 85.6%. The proposed approach can be used to detect deceptive and truthful utterances, and gradient boosting methods can significantly improve results of the classification. The proposed approach can be applied as a component of a complex multimodal system for deception detection in speech with addition of analysis of lexical information and video data. It can also be a part of a prospective system for detection of human's psycho-physiological states and destructive behaviour.

Acknowledgements

This research was supported by the RFBR (project No. 20-37-90144), as well as by the Russian state research (No. 0073-2019-0005).

References

[1] B. Schuller. The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity & native language. // Proceedings of INTERSPEECH-2016. — 2016. — P. 2001–2005.

[2] Chinchor N. MUC-4 Evaluation Metrics. // Proceedings of the Fourth Message Understanding Conference. — 1992. — P. 22–29.

[3] D. Wolpert. Stacked generalization. // Neural networks. — Vol. 5. — 1992. — P. 241–259.

[4] Dorogush A.-V. ErshovV. Gulin A. CatBoost: gradient boosting with categorical features support. // Workshop on ML Systems at NIPS 2017. — 2017.

[5] Eyben F. et al. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. // Proceedings of the 2013 ACM Multimedia (MM). — 2013. — P. 835–838.

[6] Ke G. Meng Q. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. // Advances in Neural Information Processing Systems. — 2017. — P. 3146–3154.

[7] Litvinova O. Litvinova T. et al. Deception detection in Russian texts. // Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics. — 2017. — P. 43–52.

[8] Litvinova T. A. Seredin P. V. et al. Text Classification on Basis of "False/True" using Methods of Automatic Text Processing. (In Russ.) // Nauchnyy dialog. — T. 10 (58). — 2016. — C. 70–83.

[9] Litvinova T.A. Litvinova O.A. A Study of Linguistic Features of Deceptive Texts with the Use of the Program Linguistic Inquiry and Word Count. (In Russ.) // Bulletin of the Moscow State Region University. Linguistics. — T. 4. — 2015. — C. 71–77.

[10] Litvinova T.A. Litvinova O.A. Text-Based Deception Detection: State-of-the-Art and Perspectives. (In Russ.) // Ivzestia of the Volgograd State Pedagogical University. Pedagogical sciences. — T. 2 (267). — 2015. — C. 189–192.

[11] Mendels G. Levitan S.I. et al. Hybrid acoustic-lexical deep learning approach for deception detection. // Proceedings of INTERSPEECH-2017. — 2017. — P. 1472–1476.

[12] Montaci´e C. Caraty M.-J. Prosodic cues and answer type detection for the deception subchallenge. // Proceedings of INTERSPEECH-2016. — 2016. — P. 2016–2020.

[13] Perez-Rosas V. Abouelenien M. et al. Deception Detection using Real-life Trial Data. // Proceedings of the ACM International Conference on Multimodal Interaction (ICMI 2015). — 2015. — P. 59–66.

[14] Pisarevskaya D. Litvinova T. Litvinova O. Deception Detection for the Russian Language: Lexical and Syntactical Parameters. // Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval associated with RANLP 2017. — 2017. — P. 1–10.

[15] R.K. Potapova. Variativnost' akusticheskih parametrov zvuchashhej rechi. [Variability of acoustic parameters of sounding speech]. (In Russ.) // Vestnik Moskovskogo gosudarstvennogo lingvisticheskogo universiteta. Gumanitarnye nauki. [Bulletin of Moscow State Linguistic University. Humanitarian sciences.]. — T. 740. — 2016. — C. 137–147.

[16] S.I. Smetanin. Toxic comments detection in Russian. // Proceedings of the International Conference Dialogue 2020. — 2020.

[17] Schuller B. Batliner A. et al. The INTERSPEECH 2011 speaker state challenge. // Proceedings of INTERSPEECH-2011. — 2011. — P. 3201–3204.

[18] Soldner F. P´erez-Rosas V. Mihalcea R. Box of Lies: Multimodal Deception Detection in Dialogues. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Vol. 1. — 2019. — P. 1768–1777.

[19] Tianqi Ch. Guestrin C. XGBoost: A Scalable Tree Boosting System. // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2016. — P. 785–794.

[20] Velichko A. Karpov A. Study of Data Scarcity Problem for Automatic Detection of Deceptive Speech Utterances. // Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). — Vol. 2552. — CEUR-WS, 2020. — P. 38–46.

[21] Velichko A. Budkov V. et al. Applying Ensemble Learning Techniques and Neural Networks to Deceptive and Truthful Information Detection Task in the Flow of Speech. // Intelligent Distributed Computing XIII. IDC 2019. — Vol. 868. — Studies in Computational Intelligence, Springer, Cham, 2019. — P. 477–482.

[22] Velichko A.N. Budkov V.Yu. Karpov A.A. Study of classification methods for automatic truth and deception detection in speech [Issledovanie metodov klassifikatsii dlya avtomaticheskogo opredeleniya istinnoi ili lozhnoi informatsii v rechevykh soobshcheniyakh] // Science bulletin of the Novosibirsk state technical university [Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta]. — 2018. — T. 3 (72). — C. 21–32.

[23] Zhang J. Levitan S.I. et al. Multimodal Deception Detection Using Automatically Extracted Acoustic, Visual, and Lexical Features. // Proceedings of Interspeech 2020. — 2020. — P. 359–363.

Transformers for Headline Selection for Russian News Clusters

Pavel Voropaev
Moscow Institute
of Physics and Technology
Moscow, Russia
voropaev@phystech.edu

Olga Sopilnyak
Moscow Institute
of Physics and Technology
Moscow, Russia
olga.sopilnyak@phystech.edu

Abstract

In this paper, we explore various multilingual and Russian pre-trained transformer-based models for the Dialogue Evaluation 2021 shared task on headline selection. Our experiments show that the combined approach is superior to individual multilingual and monolingual models. We present an analysis of a number of ways to obtain sentence embeddings and learn a ranking model on top of them. We achieve the result of 87.28% and 86.60% accuracy for the public and private test sets respectively.

Keywords: headline selection, news, embeddings, transformer, bert, multilingual, russian
DOI: 10.28995/2075-7182-2021-20-705-710

Применение архитектуры Transformer для выбора заголовков к новостным сюжетам на русском языке

Воропаев П. М.
МФТИ
Москва, Россия
voropaev@phystech.edu

Сопильняк О. А.
МФТИ
Москва, Россия
olga.sopilnyak@phystech.edu

Аннотация

Исследуются применения предобученных языковых моделей в рамках соревнования по выбору заголовков Dialogue Evaluation 2021. Эксперименты показывают, что ансамбль из нескольких языковых моделей показывает лучший результат, нежели индивидуальные многоязычные и моноязычные модели. Также в статье представлен ряд способов получения векторных представлений предложений, поверх которых впоследствии обучается ранжирующая модель. Получен результат: точность 87.28% и 86.60% на публичной и приватной тестовых выборках в рамках задачи.

Ключевые слова: выбор заголовков, новости, эмбеддинги, архитектура трансформер, bert, многоязычные модели, русский язык

1 Introduction

News stories clustering task not only has a wide application area in the industry, but also helps to explore the usage boundaries of sentence embeddings obtained with different models. For example, news aggregators actively use clustering algorithms to generate news feeds from different sources and to select a single headline. The recent progress in designing multilingual models [13], trained for dozens or even hundreds of languages at once, makes it possible to use them for monolingual tasks, particularly for Russian language tasks [8]. At the same time, Russian BERT-based models are actively evolving, and their comparison with more universal multilingual ones may be of interest.

The task of generating or selecting headlines for a single news cluster has a wide range of applications along with clustering. However, even considering the current level of generative models progress, the presence of great interest to them and strong state-of-the-art models [6], it is not always possible to use this models in the industry due to the unwarranted quality of generated texts and high demand for

	left	right	draw
left	1	0	0.5
right	0	1	0.5
draw	0.5	0.5	1

Table 1: Accuracy weights for headline selection task evaluation.

computational resources. An alternative choice is to select ready-made headings among those presented in the cluster. This task can be solved as a classification or ranking problem.

In this paper, the cluster headline selection task is considered. We did not find much related work except [9] where a simple rule-based system is proposed. We use the corpora provided by the Dialogue Evaluation 2021 shared task organizers [3] and introduce the solution that has shown the best results among other participants. Our code is publicly available at `https://github.com/sopilnyak/headline-selection`.

2 Experimental evaluation

The training corpus for choosing the best headline is proposed at the Dialogue Evaluation 2021 shared task. It consists of pairs of news identifiers (URLs) each one corresponding to one of four tags: `left`, `right`, `draw`, or `bad`. The last label means that the authors of the markup have identified the pair as a clustering error. The test set is divided into two parts: for public and private leaderboard, each containing news headlines for two specific dates: May 27, 2020 and May 29, 2020. To evaluate the result, a weighted accuracy is used, while the `bad` label is omitted, and the weights for the remaining labels are shown in Table 1.

2.1 Embeddings

For each headline from the training corpus, embeddings are obtained from various Russian and multilingual models. We use pretrained BERT-based models trained both on Russian monolingual corpora (RuBERT [5], SBERT [7]) and in multiple languages (mT5 [14], XLM-RoBERTa [12]) including Russian. In addition, a multilingual version of USE [11] embeddings is used. The mentioned models show state-of-the-art results on a number of NLP benchmarks [13], including those in Russian language [8], so it was natural to test them on the task of selecting the best headline for the cluster.

To obtain a headline embedding, we take the average word embeddings from the layer 19 (of 25) for SBERT, XLM-R and mT5 and layer 8 (of 13) for RuBERT, considering the length of the headline. In the case of mT5 model, which is trained mainly for solving seq2seq tasks, the decoder is removed and the embedding is taken from layer 19 of the encoder. We use recommended pretrained tokenizers from the `transformers` library [10]. These tokenizers are based on WordPiece and SentencePiece models [4].

2.2 Classification

Then we train a classifier on top of the embeddings. We use CatBoost ranking model [2], which is a gradient boosting over decision trees algorithm. Pairs of headline embeddings are fed to the classifier as input, while the best headline in the pair is considered the "positive" element of the pair, and the other one is considered "negative". We chose `PairLogitLoss` (1) as the target loss function.

$$\texttt{PairLogitLoss}(a) = -\sum_{p,n \,\in\, \texttt{Pairs}} \log\left(\frac{1}{1+e^{-(a_p - a_n)}}\right) \tag{1}$$

An ensemble of ranking models is trained based on different features. The number of decision trees in CatBoost is set to 10^3, the best epoch is chosen based on the validation score. We obtain the final headline rank by averaging the ranks predicted by each of the models, and then each pair is assigned

Model	Validation	Public LB test	Private LB test
SBERT	85.98	84.48	83.41
RuBERT	83.97	81.38	81.64
XLM-R	87.93	84.30	84.13
mT5	88.52	84.48	82.60
USE	81.23	80.79	80.68
Blend-5	**88.77**	**87.28**	**86.60**

Table 2: We report the accuracy on custom validation set and two test sets. The best results for each set are in bold. All the results are averaged over five different training runs. The ensemble of five models (Blend-5) achieves the result of 87.28% and 86.60% accuracy for the public and private test sets respectively.

one of the `left`, `right`, or `draw` labels depending on the resulting rank difference $r_r - r_l$. More specifically, the rank difference $r_r - r_l < 0$ means the winner is `left`, $r_r - r_l > 0$ means the winner is `right`, and $|r_r - r_l| \leq 0.1$ correspond to `draw`.

To explore models and compare the results, we train classifiers on top of each type of embeddings separately. Every model was trained on an Nvidia Tesla P100 GPU provided by Google Colab. Table 2 shows the results of using various embeddings and training classifiers on top of them. The table shows that the best result is obtained by the ensemble of the five models mentioned in the paper (referred as Blend-5), but the single multilingual model mT5 shows a comparable accuracy.

During the Dialogue Evaluation 2021 competition we achieved 86.00% and 85.40% accuracy for the public and private test sets by taking the average word embeddings in the top layer. But further experiments show that middle layers perform better than top layers with the result of 87.28% and 86.60% accuracy respectively.

2.3 Analysis and results

In this section, we analyze the model and explore the impact of several aspects of out approach. We list some examples of wrongly predicted labels and report the evaluation results for different variants of the model on the private LB test set.

Error analysis. We selected 300 examples where the model confused between `left` and `right` labels and skipped examples with `draw` label. Errors can be divided into several types as listed at (2)–(5) together with several corresponding examples and their translation to English. The model is more likely to prefer titles with lack of facts and sometimes tends to choose verbose headlines. Other wrongly selected headlines include opinions and biased titles. Finally, having a pair of almost equivalent titles, the model could choose the wrong label.

(2) **Headlines containing insufficient facts**

gold: Ту-22М3М Казанского авиазавода испытали на сверхзвуке
Kazan aircraft factory's Tu-22MZM tested at supersonic mode

pred: В ОПК рассказали об испытаниях модернизированного ракетоносца Ту-22М3М
The defense industry told about the tests of the modernized Tu-22M3M missile carrier

gold: День проведения парада Победы будет нерабочим — Песков
Victory Parade day will be non-working — Peskov

pred: Песков заявил о большой вероятности объявления еще одного выходного
Peskov said about the high probability of announcing another weekend

(3) **Verbose headlines**

gold: С 27 мая москвичи могут бесплатно сдать тест на антитела
From May 27, Moscow residents can take an antibody test for free

pred: «Как и где сдать тест на антитела к коронавирусу в Москве. С 27 мая это может сделать любой желающий
How and where to take an antibody test against coronavirus in Moscow. Anyone can do it from May 27

gold: Украинский суд признал нацистской символику дивизии СС «Галичина»
The Ukrainian court found that the symbols of the SS division "Galicia" were Nazi

pred: «Победа справедливости, здравого смысла и закона»: Вятрович проиграл суд по делу о символике СС «Галичина»
"Victory of justice, common sense and law": Vyatrovich lost the court case concerning the SS "Galicia" symbols

(4) **Biased headlines**

gold: Роскомнадзор начнет блокировать в России пиратские приложения
Roskomnadzor will begin to block pirated applications in Russia

pred: Госдума приняла спорный законопроект о блокировке приложений с пиратским контентом
State Duma passed controversial bill on blocking applications with pirated content

gold: Доллар в обменниках ускорил рост
Dollar in exchange offices accelerated growth

pred: Заманивают иностранцев под покупку облигаций. Почему гривня снова падает
Luring foreigners to buy bonds. Why is the hryvnia falling again

(5) **Equivalent pairs of headlines**

gold: Россияне массово забирают валюту из банков
Russians massively withdraw currency from banks

pred: Жители страны массово снимают валюту с банковских счетов
Residents of the country are massively withdrawing currency from bank accounts

gold: Россиянам разъяснили, когда можно будет поехать в отпуск за рубеж
Russians were clarified when it will be possible to go on vacation abroad

pred: Россиянам озвучили возможные сроки возобновления поездок за границу
Russians were announced the possible date of the resumption of trips abroad

Thus, a good headline can be defined as precise, short, unbiased, and containing as many significant facts as possible. However, classifiers based on language models can rank headlines which do not meet these criteria higher than manually selected ones. We assume that adding more training data or introducing multitask learning, combining another training objectives, such as information extraction, could help achieve better results.

Sentence representations. We explore a different way to obtain sentence embeddings from language model's top layer output: using the embeddings of the first token, known as [CLS] token, which is sometimes more common than averaging the word embeddings. We compare the results for all models, except mT5, which doesn't have the [CLS] token in its dictionary. However, as shown in Table 3, averaging show slightly better results for most of the models. We assume this is because embeddings of the [CLS] token contain high-level semantic meaning [13], while for headline selection task it is more important not to lose the token-level information to meet the previously formulated criteria of a good headline. Moreover, we analyze whether sentence embeddings in the middle layers can be more suitable than in the last ones. Experiments show that layers 17 to 19 (of 25) for SBERT, XLM-R and mT5 and layers 8 to 9 (of 13) for RuBERT perform better than the top layers. The reason for this may be the same: top layers embed more high-level semantic meaning than the middle ones.

Sentence representation	SBERT	RuBERT	XLM-R	mT5	Blend-5
Top layer: average embeddings	78.62	75.50	80.01	81.60	85.20
Top layer: [CLS] token embeddings	77.77	76.53	75.48	—	84.50
Layer 23/11: average embeddings	81.23	78.65	82.59	81.83	85.64
Layer 21/9: average embeddings	83.31	81.41	83.30	82.48	86.31
Layer 19/8: average embeddings	83.41	**81.64**	**84.13**	82.60	86.60
Layer 17/8: average embeddings	**83.55**	**81.64**	84.06	**83.56**	**86.93**

Table 3: Comparison of sentence representations. The table shows accuracy on private LB test set. Sentence representations are either averaged token embeddings in middle and top layers, or [CLS] token embeddings in the top layer. Layer N/M means that for SBERT, XLM-R and mT5 we take layer N and for RuBERT we take layer M.

3 Conclusion

In this paper, we study applications of pre-trained models for headline selection and demonstrate the superiority of ensembles of modern BERT-based models. We have shown that multilingual models, such as mT5, demonstrate decent results in the task, and are superior to the single-language models in the same conditions.

Further research can be made in additional training of the top layers of multilingual models on Russian-language corpora, as well as more fine-tuning of lightweight models, such as multilingual USE or LASER [1], to reduce system requirements for headline selection.

References

[1] Artetxe Mikel, Schwenk Holger. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond // CoRR. — 2018. — Vol. abs/1812.10464. — 1812.10464.

[2] Dorogush Anna Veronika, Ershov Vasily, Gulin Andrey. CatBoost: gradient boosting with categorical features support // CoRR. — 2018. — Vol. abs/1810.11363. — 1810.11363.

[3] Gusev Ilya; Smurov Ivan. Russian News Clustering and Headline Selection Shared Task // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”. — 2021.

[4] Kudo Taku, Richardson John. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing // CoRR. — 2018. — Vol. abs/1808.06226. — 1808.06226.

[5] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // CoRR. — 2019. — Vol. abs/1905.07213. — 1905.07213.

[6] Brown Tom B., Mann Benjamin, Ryder Nick et al. Language Models are Few-Shot Learners. — 2020. — 2005.14165.

[7] Reimers Nils, Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // CoRR. — 2019. — Vol. abs/1908.10084. — 1908.10084.

[8] RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark / Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov et al. // arXiv preprint arXiv:2010.15925. — 2020.

[9] Thirunarayan Krishnaprasad, Immaneni Trivikram, Shaik Mastan. Selecting Labels for News Document Clusters. — 2007. — 01. — P. 119–130.

[10] Transformers: State-of-the-Art Natural Language Processing / Thomas Wolf, Lysandre Debut, Victor Sanh et al. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — Online : Association for Computational Linguist-

ics, 2020. — Oct. — P. 38–45. — Access mode: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[11] Universal Sentence Encoder / Daniel Cer, Yinfei Yang, Sheng-yi Kong et al. // CoRR. — 2018. — Vol. abs/1803.11175. — 1803.11175.

[12] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // CoRR. — 2019. — Vol. abs/1911.02116. — 1911.02116.

[13] XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization / Junjie Hu, Sebastian Ruder, Aditya Siddhant et al. — 2020. — 2003.11080.

[14] Xue Linting, Constant Noah, Roberts Adam et al. mT5: A massively multilingual pre-trained text-to-text transformer. — 2021. — 2010.11934.

The prosody of spoken dialogue

T. E. Yanko
Institute of Linguistics / Moscow, Russia
tanya_yanko@list.ru

Abstract

This paper is aimed at establishing the parameters of the dialogic communication expressed through Russian prosody. The linguistic and extra-linguistic constituents of dialogue are analyzed. These are: the *illocutionary meanings* that generate speech acts, characteristic of the dialogic communication; the *discourse links* that combine the successive speech acts of one interlocutor if his/her current contribution into the dialogue is not limited to a single speech act; the *prosodic characteristics of genre* typical for a concrete type of communication (a friendly talk, an exam, a press conference, a scientific presentation, or an interrogation). The proposed taxonomy is based on the analysis of the minor working corpus of spoken dialogues from the Russian National corpus (Multimodal sub-corpus Murko), the annotated database Spokencorpora.ru, video-hosting Youtube.com, films, scientific conferences, and press conferences. The computer system Praat is used to analyze the sound data. The paper is illustrated with tracings of sound records.

Key-words: spoken speech; dialogue; prosody; discourse
DOI: 10.28995/2075-7182-2021-20-711-719

Просодические параметры звучащего диалога

Т. Е. Янко
Институт языкознания РАН / Москва, Россия
tanya_yanko@list.ru

Аннотация

В работе предлагается параметрическая модель просодии звучащего диалога. Цель работы — выделение набора параметров диалогической коммуникации, имеющих просодию в качестве основного средства выражения. Рассматриваются следующие языковые и неязыковые составляющие диалогической коммуникации: 1) *иллокутивные* смыслы, которые формируют речевые акты, специфические для диалога; 2) *дискурсивные* связи, которые объединяют последовательные речевые акты одного коммуниканта, если его текущий вклад в диалог не ограничен одним речевым актом;b3) *просодические особенности жанра*, которые могут включать и иллокутивные, и дискурсивные смыслы, но которые характерны для определенного типа коммуникации (дружеской беседы, интервью, экзамена, пресс-конференции, доклада, допроса).

Ключевые слова: звучащая речь; диалог; просодия; дискурс

0 Введение

В работе предлагается параметрическая модель просодии звучащего диалога. Цель работы — выделение набора параметров диалогической коммуникации, имеющих просодию в качестве основного средства выражения. Рассматриваются следующие языковые и неязыковые составляющие диалогической коммуникации: 1) *иллокутивные* смыслы, которые формируют речевые акты, специфические для диалога; 2) *дискурсивные* связи, которые объединяют последовательные речевые акты одного коммуниканта, если его текущий вклад в диалог не ограничен одним речевым

актом; 3) *просодические особенности жанра*, которые могут включать и иллокутивные, и дискурсивные смыслы, но которые характерны для определенного типа коммуникации (дружеской беседы, интервью, экзамена, пресс-конференции, доклада, допроса).

Предлагаемая классификация разработана на основе рабочего корпуса диалогов и монологов общим звучанием около часа. Источником материала для рабочего корпуса послужил Мультимедийный подкорпус Русского Национального корпуса Мурко НКРЯ [RNC 2021], звучащий просодически аннотированный корпус [Spokencorpora 2021], видеохостинг Youtube [Youtube.com 2021], записи радио, телеканала РБК, материалы кинофильмов, пресс-конференций, научных конференций. В качестве инструментального средства использована компьютерная система анализа звучащей речи Praat [Boersma, Weenink 2021].

01 Просодическая нотация и анализ примера

В работе используется просодическая нотация, которая продолжает традицию [Kodzasov, Bonch-Osmolovskaja, Zaharov, Kobozeva, Krivnova 2005, 2006] и имеет соответствие с описанием русской просодии в духе интонационных конструкций Е.А. Брызгуновой [Bryzgunova 1982: 97-122]. Обозначения движений тона в примерах ниже расположены после ударной гласной словоформы — носителя акцента, ср. *шра\\м* в примере (1).

1. \ — падение частоты основного тона типа ИК-1 с понижением частоты на ударном слоге словоформы — носителя акцента и дальнейшим понижением или ровным низким тоном на заударных слогах, если они есть [Bryzgunova 1982: 97-122].

2. \\ — падение частоты типа ИК-2 с понижением на ударном слоге словоформы-акцентоносителя в более широком частотном диапазоне, чем в случае с ИК-1, и дальнейшим понижением на заударных слогах [Bryzgunova 1982: 97-122]. В начальной зоне артикуляции ударного слога или даже в исходе предударного слога обычно имеется компенсаторное повышение частоты, необходимое для реализации масштабного падения тона.

3. / — подъем частоты на ударном слоге акцентоносителя плюс падение на заударных слогах, если они есть. Если заударных в словоформе нет, заударное падение элиминируется (ИК-3 [Bryzgunova 1982: 97-122]).

4. \/ — падение частоты основного тона или ровный низкий тон на ударном слоге акцентоносителя плюс подъем на заударных слогах если они есть. Если заударных слогов нет, интегральное нисходяще-восходящее движение тона фиксируется на конечном или единственном слоге акцентоносителя (ИК-4 [Bryzgunova 1982: 97-122].

5. /- — подъем частоты на ударном слоге акцентоносителя плюс ровные или слабо нисходящие заударные (ИК-6 [Bryzgunova 1982: 97-122]).

Проиллюстрируем некоторые параметры просодического оформления диалогической речи на примере из интервью, данного киноактрисой журналисту:

(1) Q1. *Твой любимый шра\\м?*
A1. *Вот э\тот. Он у меня с де\тства.*
Q2. *Любимая кни\/га?*
A2. «*Гарри По\ттер*».
Q3. *Любимый худо\/жник?*
A3. *Магри\тт.*
Q4. *Любимая гру\\ппа?*
A4. *Не зна\ю. Пусть Электрик лайт* орке\стра [Youtube.com 2021].

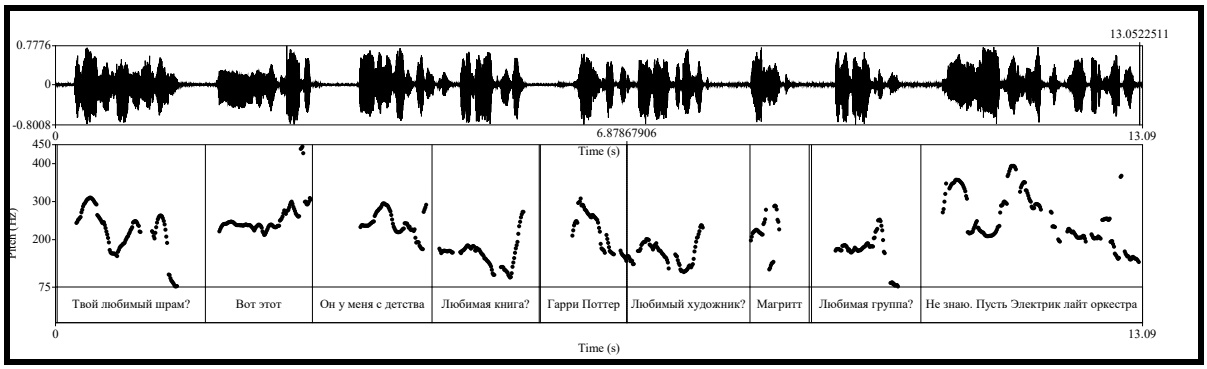


Рисунок 1: Тонограмма примера (1)

Пример (1) — это диалог из четырех вопросов и четырех ответов. Вопросы и ответы формулируются в краткой форме, что характерно для жанра блиц-интервью или опроса (в сравнении, например, со «школьными» вопросом и ответом, где требуется т. н. «полные» ответ и вопрос). Краткая форма вопроса определяется отсутствием вопросительного слова. Вопросы с отсутствующим вопросительным словом, как вопросы Q1 *Твой любимый шрам?* и Q2 *Твоя любимая книга?*, мы называем эллиптическими. Ср. с реконструируемой полной формой: *Какой твой любимый шрам?*; *Какая у тебя любимая книга?*. Для краткого ответа характерно отсутствие повторения в начале ответа конца вопроса, ср. ответ A2 «*Гарри Поттер»* с полной формой ответа: *Моя любимая книга — «Гарри Поттер».*

Тонограмма на Рис. 1 показывает, что слова — акцентоносители рем в ответных репликах несут нисходящий акцент типа ИК-1, что характерно для речевого акта сообщения. Это пример иллокутивной просодии, она маркирует определенную иллокутивную силу. Иллокутивная сила вопроса в этом диалоге маркируется отсутствующим (но реконструируемым при раскрытии эллипсиса) вопросительным словом.

Первый и четвертый вопрос Q1 и Q4 несут нисходящий акцент типа ИК-2 на акцентоносителе вопроса (*шра\\м*, *гру\\ппа*), который и характерен для одиночного (не имеющего с контекстом обозначенных дискурсивных связей) вопроса с вопросительным словом, а второй Q2 и четвертый Q4 — нисходяще-восходящий акцент типа ИК-4, который говорит о том, что задаваемые вопросы объединяются в серию. ИК-4 в вопросе Q2 говорит о том, что этот вопрос не единственный и что он связан с предыдущим сопоставительной связью: 'ты ответила на вопрос, какой твой любимый шрам, а теперь скажи, какая у тебя любимая книга'. И далее: Q3 ‘а какой тогда у тебя любимый художник?’. На этапе четвертого вопроса интервьюер возвращается к стратегии нисходящей просодии, т.е. связь между вопросами интервьюера, которая могла бы тут в принципе быть, отсутствует. Смена стратегий играет в диалогах стилистическую роль: повтор стратегии, в особенности, маркированной, как в случае с ИК-4, неизбежно приводил бы к монотонности диалога. Таким образом, пример 1) демонстрирует иллокутивную просодию сообщения и иллокутивную силу вопроса, а также иллюстрирует феномен просодического объединения вопросов в серии. Последнее, в свою очередь, демонстрирует стратегию, характерную для некоторых диалогических жанров. В данном случае это интервью, предполагающее, что текущий вопрос не единственный и что вопросы сопоставлены один другому.

Обратимся к последовательному выделению просодических стратегий, формирующих диалог, и их параметров.

1. Иллокутивные просодии

В диалоге используются все известные типы речевых актов: сообщение, вопрос, обращение, императив, восклицание. Стандартные просодии речевых актов и их компонентов, таких, как тема и рема, хорошо описаны в литературе: [Bryzgunova 1982: 99-101], [Kodzasov, Bonch-Osmolovskaja, Zaharov, Kobozeva, Krivnova 2005], [Kodzasov, Arhipov, Bonch-Osmolovskaja, Zaharov, Krivnova 2006]. В сообщении и в вопросе без вопросительного слова просодия формирует

речевой акт как имеющий определенную иллокутивную силу, а в императиве и в вопросе с вопросительным словом, а также в вопросе с частицей *ли*, просодия играет только формирующую роль, потому что иллокутивная сила здесь выражена сегментными средствами: в императиве — наклонением, а в вопросе с вопросительным словом и вопросительной частицей *ли* — вопросительным словом.

Вокативные просодии отличаются существенным разнообразием, базовый тип обращений, т. е. такой, который вносит в семантику речевого акта минимальный вклад, предполагает нисходящую просодию типа ИК-2: *Ва\\ся*!. Мы не останавливаемся здесь на просодическом формировании обращений в условиях удаленности говорящего от слушающего, близости говорящего к слушающему, в ситуации, когда говорящий не видит слушающего, и других; детали см. в [Yanko 2008: 98-107].

2 Дискурсивные просодические средства организации диалога

Из просодий, которые объединяют диалогические реплики в серии, можно выделить *катафорические* и *анафорические*. При катафорических стратегиях, или стратегиях незавершенности, используются указания на то, что текущий речевой акт не последний. Анафорические стратегии соединяют речевой акт с предыдущими речевыми актами или с известным говорящему и слушающему контекстом. Так, вопрос Q3 про любимого художника сопоставляет любимого художника и любимую книгу. Кроме того, эта стратегия может сопоставлять друг с другом вопросы из известного (подготовленного говорящим заранее) списка: *Ваше и\/мя? Во\/зраст? Факульте\/т? Ку\/рс?* [Bryzgunova 1982: 114]. Как интервью, так и опрос, а также, скажем, допрос, интересны не только сами по себе, но и как богатый источник («грибное место») примеров использования определенной дискурсивной стратегии и соответствующей просодии, разработанной русским языком. В других жанрах эта стратегия встречается статистически редко. Это серия вопросов, продуманная говорящим заранее. Поясним, что наша цель в этой работе — охватить не полноту речевых жанров, а, скорее, полноту дискурсивных стратегий, разработанных русским языком, и соответствующих системных средств просодического выражения.

2.1 Дискурсивная катафора

Выделяются 1) просодии, которые соединяют текущее *сообщение* с последующим сообщением или речевым актом иного типа (просодии незавершенного сообщения); 2) просодии незавершенности в применении к *вопросу*, за которым следует другой вопрос или речевой акт иного типа; 4) просодии незавершенности в применении к *вокативу*; 3) просодии незавершенности в применении к *императиву*.

Обратимся к примеру объединения в единое дискурсивное целое двух сообщений.
(2) *Потом мы с папой искали эту соба/чку. Потом нашли\ эту собачку* [Spokencorpora 2021].

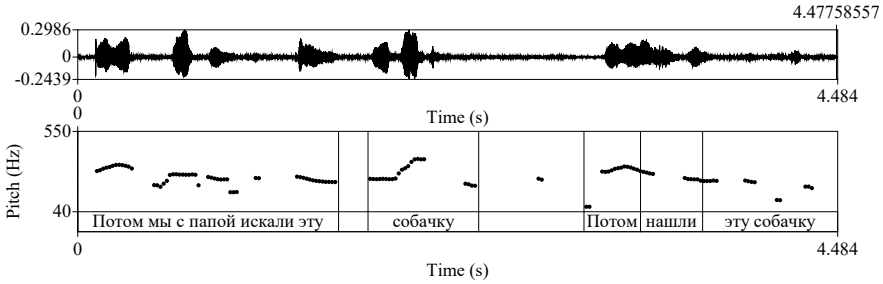


Рисунок 2: Тонограмма примера (2).

Первое из двух сообщений содержит подъем частоты на ударном слоге с последующим нисходящим тоном в заударной зоне типа ИК-3 на акцентоносителе незавершенности словоформе *соба/чку*. Второе сообщение завершает дискурсивный фрагмент, о чем говорит падение частоты типа ИК-1 на словоформе *нашли*. Реализация этого небольшого текста в отсутствие межфразо-

вой связи, которую обеспечивает в нем просодия незавершенности на первой встречаемости словоформы *собачку*, также была бы вполне возможной. Иначе говоря, в примере (2) реализована одна из двух возможных стратегий. Это говорит о том, что связь незавершенности между предложениями в данном случае факультативна: при артикуляции этого текста говорящий осуществляет свободный выбор между стратегией с указанием на незавершенность и без таковой. Между тем дискурс может содержать и такие средства указания на незавершенность, которые можно считать обязательными. Так, в примере (3), который представляет собой сложносочиненное предложение с союзом *а*, имеется показатель дискурсивной незавершенности и несколько показателей т.н. внутренней незавершенности (например, перехода от темы к реме или перехода от одного члена сочинительной конструкции к другому). Внутренние показатели незавершенности обязательны.

(3) *Напомню, это очень сильный окисли/тель, а не/дра Земли, ма/нтия, а уж тем более ядро/ являются очень сильными восстанови\телями* [RNC 2021].

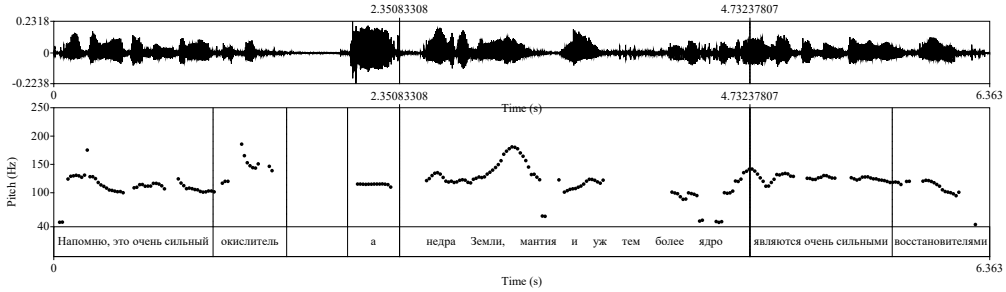


Рисунок 3: Тонограмма примера (3).

Мощный языковой сигнал к реализации незавершенности — это сочинительный союз *а*, который соединяет два предложения *… это очень сильный окисли/тель* и *недра Земли…являются очень сильными восстанови\телями*. Фактически союз управляет подъемом типа ИК-3 на акцентоносителе незавершенности *окисли\тель* в первом из сочиненных предложений. Остальные подъемы в этом предложении — это показатели внутренней незавершенности (*не/дра, ма/нтия, ядро/*). Второе предложение в составе сложносочиненного имеет падение на словоформе *восстанови\телями.* Этот речевой акт концептуализуется как не связанный с последующим речевым контекстом. В принципе, ситуации, в которых говорящий заканчивает свою речь показателем незавершенности, также вполне возможны, например, если говорящего перебили «на полуслове», или при сомнениях (хезитации) говорящего по поводу того, продолжать ли ему речь или остановиться.

Подъемы дискурсивной незавершенности весьма характерны для обращения (ср. подъем незавершенности типа ИК-3 на *ребя/т* в примере *Слушайте, ребя/т, ну хватит, ну шесть крупнейших госбанков…*[RNC 2021]), императива и вопроса с вопросительным словом, ср. в примере *Ну, Валерий,в двух словах мне, пожалуйста, расскажи/-те, что случи/-лось. Почему вы здесь нахо/-дитесь, что такого вы де/-лали…* [Youtube.com 2021] подъемы незавершенности типа ИК-6 на словоформах *расскажи/-те, случи/-лось, нахо/-дитесь* и *де/-лали*.

Между тем, что касается *да-нет*-вопроса (вопроса без вопросительного слова), проблема способности его встраивания в дискурсивный ряд незавершенности долгое время оставалась нерешенной. Причиной неразработанности соответствующей языковой техники представлялось совпадение маркера *да-нет*-вопроса и базового маркера незавершенности. Это восходящий акцент типа ИК-3. Анализ материала, представленного в Мультимедийном подкорпусе НКРЯ [RNC 2021], задачи «Сочетается ли контекст *да-нет*-вопроса с дискурсивной незавершенностью?» не решал. «Заполнить пустую клетку» удалось лишь после обращения к материалу пресс-конференций: когда журналисту дают слово, он, как правило, задает сразу несколько вопросов, потому что второй раз слово ему уже не дадут. В таком контексте и были обнаружены *да-нет*-вопросы с искомым компонентом катафорической связи. Обратимся к примеру (4).

(4) *Мо/жете дать свои оценки на семнадцатый го/д по прибыли банковского сектора, <потому что в общем-то банки в шестнадцатом году показывают неплохую при/быль, и э-э-э из чего будут формироваться их доходы в семнадцатом году/?>* [Youtube.com 2021].

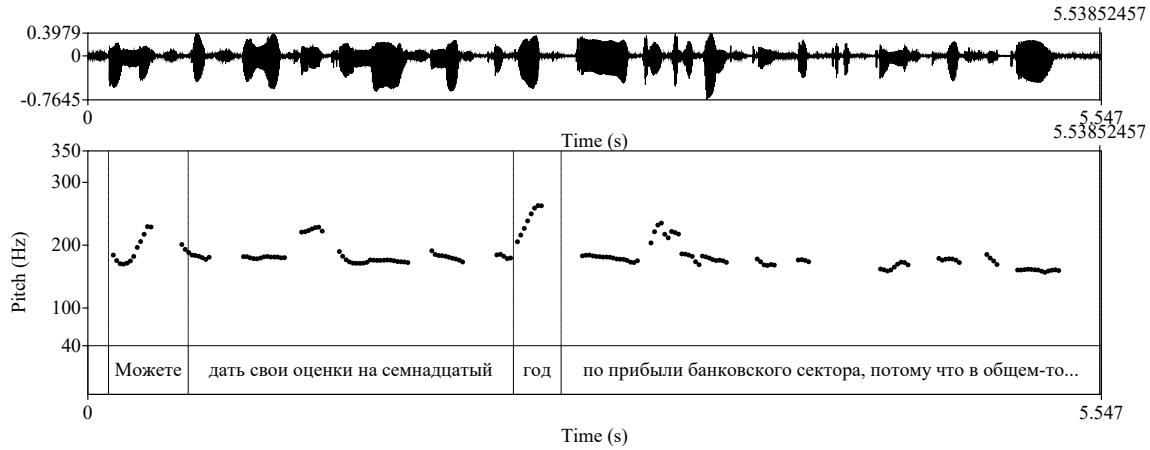


Рисунок 4: Тонограмма примера (4)

Журналистка задает председателю Банка России вопрос, состоящий из начального *да-нет*-вопроса *Мо/жете дать свои оценки на семнадцатый го/д…?*, за которым следует объяснение, почему задается такой вопрос (*потому что в общем-то банки в шестнадцатом году показывают неплохую прибыль*). Можно видеть, что показателем *да-нет*-вопроса служит подъем на начальной словоформе *можете*, а показателем незавершенности в рамках синтаксической структуры того же вопроса становится подъем на словоформе *год*. На этом сложное предложение не кончается; говорящий встраивает в его структуру еще один вопрос: это вопрос с вопросительной группой *из чего*. Об объединении речевых актов различного типа в одно сложное предложение см. [Kobozeva 1999]. Ср. также примеры с *да-нет*-вопросами в контексте незавершенности, аналогичные сложному предложению (4): *Мо/жете прокомментировать понижение прогноза по нефти в следующем году/? Вы уже упомянули, что одна из причин долларовой волатильности последних месяцев…*; *Пра/вильно я понимаю, что вы не так сильно верите в силу этого соглаше/ния и что… все равно цена на нефть будет снижаться? Совет директоров ЦБ … как-то учи/тывает действующие прогнозы анали/тиков…?* [Youtube.com 2021].

По материалу интервью, вопросов на научных конференциях и, в особенности, по данным пресс-конференций можно заключить, что русская речь сформировала определенный жанр серий вопросов, а также вопросов, за которыми следует обоснование того, почему задается вопрос. Такие вопросы функционируют в условиях жесткого дефицита времени: журналист задает все вопросы, которые он заготовил, одной обоймой. А этих условиях регулярно возникают вопросы, в том числе *да-нет*-вопросы, в которых содержится просодическое указание на то, что текущий вопрос не последний в ряду артикулируемых речевых актов. Таким образом, пресс-конференция для лингвиста — это источник примеров на сочетаемость в одном речевом акте иллокутивной силы *да-нет*-вопроса и дискурсивной незавершенности. При этом и *да-нет*-вопрос, и незавершенность имеют одинаковую просодию (ИК-3), но различные словоформы — акцентоносители такой просодии.

Просодия вопросительных предложений с вопросительными словами и частицей *ли* вариативна и заслуживает отдельного обсуждения, которое мы здесь оставляем здесь без детализации, см. [Bryzgunova 1982: 99-111], [Kodzasov 2009: 175-198]. На функционирование вопросов с вопросительными словами и частицей *ли* в контексте дискурсивной незавершенности никаких ограничений мы не видим.

2.2 Дискурсивная анафора

Связь не только с последующим, но и с предшествующим, а также просто с известным контекстом в диалогической речи можно наблюдать на примере вопросительных предложений, где акцентоноситель маркирован акцентом типа ИК-4. Выбор анафорической стратегии может быть продиктован говорящему логической или прагматической связью между вопросами. Так, для ситуации, выраженной акцентом ИК-4 на словоформе *кинотеатре* в примере (5), мы предлагаем

следующее толкование: ‘если речь зашла о походе в кино, то в какой конкретно кинотеатр ты ходила?’:

(5) Q5. *Так, после шестнадцати часов где\ была?*
A5. *В кино ходи\\ла.*
Q6. *М-м-м. Какой фи\льм?*
A6. Худо\\жественный.
Q7. *В каком кинотеа\/ре?*
A7. *В кинотеа/тре имени Пу\шкина, товарищ следователь* [Youtube.com 2021].

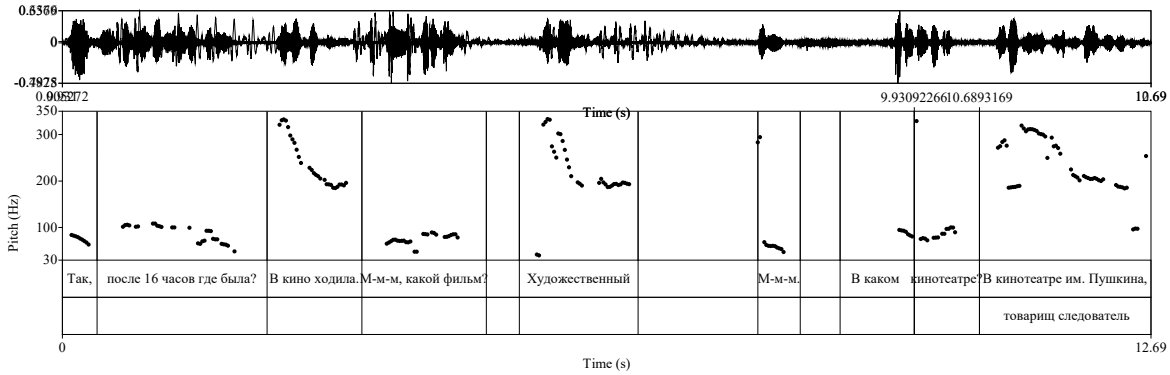


Рисунок 5: Тонограмма примера (5)

Тонограмма демонстрирует акцент типа ИК-4 на словоформе *кинотеатре*. Аналогичная стратегия наблюдается и в вопросах Q2 и Q3 из интервью (1), где вопросы связаны с предшествующим контекстом сопоставительной связью.

Стратегии указания на дискурсивную связь, как правило, чередуются с отсутствием указания на предшествующий или последующий контекст: *— Имя-о\/тчество? — Фролов Олег Ива\нович. — Дата рожде\/ния? — Семьдесят седьмого ноль пятого тридца\того. — Ме\\сто рождения? — Город Электросталь* [Youtube.com 2021]. В последнем примере из полицейского допроса мы наблюдаем два вопроса с акцентом ИК-4, который характеризует т.н. «демографическую» зону опроса, и вопрос с ИК-2, относящийся к той же зоне, но содержащий контраст между датой рождения и местом рождения. (ИК-2 — это, в частности, акцент контраста.) Соответственно, при переходе от даты рождения к месту рождения вторая встречаемость словоформы *рождения* произносится безакцентно, а словоформа *место* получает контрастный нисходящий акцент.

Анализ говорит о том, что задающие такие вопросы всегда варьируют стратегии артикуляции вопросов: в имеющемся рабочем массиве серии вопросов, которые произносились бы с жестким соблюдением какой-либо единой стратегии, отсутствуют. Даже функционально однообразной речи свойственно семантическое и просодическое разнообразие.

Детали вклада акцентов (ИК-3, ИК-4 и ИК-6) в семантику незавершенности, а также проблемы выбора акцентоносителя незавершенности, мы здесь не приводим, т.к. это тема для отдельного исследования.

2.3 Передача слова. Средства связи в вопросо-ответной паре

Аналогично повествовательному предложению (сообщению), где различаются коммуникативные компоненты «тема» и «рема», в вопросительном предложении мы выделяем невопросительный компонент и вопросительный компонент. Так, в вопросе *Какие звуки издаёт фагот?* [RNC 2021] *какие звуки* — это вопросительный компонент, а *издаёт фагот* — невопросительный. В вопросе *Вот этот букет сто рубле/й?* [RNC 2021] *сто рублей* — это вопросительный компонент, а *вот этот букет* — невопросительный.

Наблюдается следующая закономерность. В т.н. полном ответе на вопрос с порядком следования коммуникативных компонентов Тема-Рема вопросительный компонент вопроса преобразуется в тему ответа с соответствующей восходящей просодией темы на акцентоносителе темы. При этом собственно ответ на вопрос становится ремой ответа. Рассмотрим диалог (6):

(6) — *Завтра в десять она придет и принесет все деньги*.

Q6. — *А если не придет*?

A6. — *Если не приде/т, в десять ноль пять я положу вам рапорт об увольне\нии* [Youtube.com 2021].

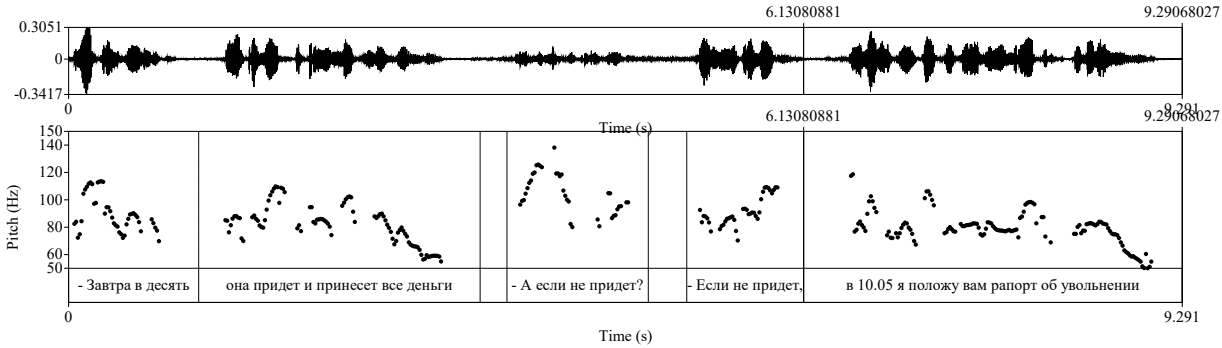


Рисунок 6: Тонограмма примера (6).

Фрагмент *если не придет* переносится из вопроса в «начало» (тему) ответа. Закономерность, проиллюстрированная здесь примером (6) и устанавливающая фактическое дублирование «конца» вопроса в «начале» ответа с соответствующей просодией темы, позволяет упростить алгоритм порождения ответной реплики в машинных диалоговых системах по сравнению с порождением ответной реплики без использования материала вопроса.

3 Просодические особенности жанра

Коммуникативные и просодические стратегии, рассмотренные выше, позволяют установить некоторые закономерности, которые определяются функциональными особенностями жанра речи.

3.1 Пресс-конференция, лекция

Протокол массовых мероприятий предусматривает процедуру ответов ведущего (лектора, докладчика, ответственного лица) на вопросы слушающих: журналистов, коллег, студентов. Наиболее напряженная атмосфера создается на пресс-конференциях, когда журналисту приходится задавать вопросы и давать объяснения, почему такие вопросы возникают, в кратчайшее время.

Обстановка пресс-конференции заставляет говорящих вырабатывать специальные коммуникативно-просодические стратегии, связывающие вопрос с контекстом, прежде всего, с последующим: говорящий просодически сигнализирует слушающему, что текущий вопрос не последний. Видеоряд пресс-конференций председателя Банка России Э. Набиуллиной подтверждает, что в тот момент, когда задающий вопрос журналист артикулирует сигнал незавершенного вопроса (акценты типа ИК-3 или ИК-6), Набиуллина берется за авторучку, чтобы начать записывать: по просодии она понимает, что от говорящего ожидается несколько вопросов. В более непринужденной обстановке, где вопрос «иллокутивно вынуждает» непосредственно следующий за ним ответ, вопросы не объединяются в последовательности; об иллокутивном вынуждении см. [Krejdlin, Baranov 1992]. На пресс-конференциях в последовательности группируются даже *да-нет*-вопросы. Можно предположить, что на лекциях и научных конференциях это тоже возможно, однако в подготовленном рабочем массиве такие примеры пока отсутствуют.

3.2 Интервью и допрос

Для жанров с подготовленными вопросами характерно просодическое соотнесение текущего вопроса с другими вопросами из заготовленного ряда. Маркер сопоставления вопроса с контекстом — акцент ИК-4. ИК-4 как показатель сопоставительной связи функционирует в контексте вопросов с вопросительным словом, в том числе, эллиптических. Если слушающий был в кино (или говорит, что был) естественно возникает вопрос, какой фильм он смотрел и в каком кинотеатре:

следователь готов начать проверку показаний; см. пример (5) выше. Жанр интервью с его особенностями формулировки вопросов, также соотносимых с контекстом, был рассмотрен во Введении, см. пример (1).

***

Предложена классификация параметров звучащего диалога, которые в качестве средства выражения используют просодические показатели. Это набор следующих значений. 1) *Иллокутивные цели* участников диалога. Говорящие обмениваются сообщениями, задают друг другу вопросы, дают ответы, обращаются друг к другу по имени, прозвищу, титулу, или с использованием слов, которые выражают чувства (*котик!; предатель!*). Иллокутивные смыслы оформляются прежде всего просодически, но также и другими языковыми средствами, а также комбинациями просодических и сегментных средств. 2) *Просодические средства, формирующие связный дискурс* в речи одного персонажа и при передаче слова, прежде всего, в вопросо-ответной паре. Соотнесение реплик друг с другом в рамках диалога также по преимуществу осуществляется через просодию. 3) *Жанровые просодические характеристики диалога* рассмотрены на примере специфических просодий, проявляющих себя на пресс-конференциях, в полицейских допросах и журналистских интервью.

References

[1] Boersma P., Weenink D. Praat: Doing phonetics by computer. Version 6.1.39. 2021. Online: http://www.praat.org/.

[2] Kobozeva I. M. An essay in characterizing lexical-syntactic, semantic, and pragmatic properties of interrogative dialogical turns in terms of features [Opyt razrabotki priznakovoj bazy dlja harakteristiki leksiko-sintaksicheskih, semanticheskih i pragmaticheskih svojstv voprositel'nyh replik] // Proceedings of the International Conference “Dialog 2005” [Trudy mezhdunarodnoj konferencii «Dialog 2005»]. 2005. — P. 238–244.

[3] Kodzasov S. V. Investigations in Russian Prosody [Issledovanija v oblasti russkoj prosodii]. M.: Jazyki slavjanskih kul'tur, 2009. — 496 c.

[4] Kodzasov S. V., Arhipov A. V., Bonch-Osmolovskaja A. A., Zaharov L. M., Krivnova O. F. Data Base ‘Intonation of Russian Dialogue: Interrogative Phrases’ [Baza dannyh «Intonacija russkogo dialoga»: voprositel'nye repliki] // Proceedings of the International Conference «Dialog 2006». [Trudy mezhdunarodnoj konferencii «Dialog 2005»]. 2005. — P. 245– 249.

[5] Kodzasov S. V., Arhipov A. V., Bonch-Osmolovskaja A. A., Zaharov L. M., Krivnova O. F. Data Base ‘Intonation of Russian Dialogue: Commanding Propositions’ [Baza dannyh «Intonacija russkogo dialoga»: pobuditel'nye repliki] // Proceedings of the International Conference «Dialog 2006». [Trudy mezhdunarodnoj konferencii «Dialog 2006»]. 2006. — P. 236–242.

[6] Krejdlin G.E., Baranov A.N. On illocutionary imposing in the structure of dialogue [Illokutivnoe vynuzhdenie v strukture dialoga] // Problems of linguistics [Voprosy jazykoznanija]. № 2. 1992. — P. 84-99.

[7] Yanko T. Intonational strategies of the Russian speech from a contrastive perspective [Intonatsionnye strategii russkoj rechi v sopostaviteljnom aspekte]. 2008. — Moscow: Jazyki slavjanskih kul’tur.

Web sources

[1] Russian National Corpus (RNC 2021): <www.ruscorpora.ru> (14.02.2021).

[2] Prosodically Annotated Corpus of Spoken Russian (Spokencorpora 2021). Online: http://spokencorpora.ru.

[3] Youtube.com (2021) Consulted online in February-March 2021

Russian discourse markers *vidimo* and *po-vidimomu* ('apparently'): synchronic and diachronical semantics

Anna A.Zalizniak
Institute of Linguistics of the Russian Academy of Sciences /
1, B. Kislovskij Ln., Moscow, 125009, Russia
Institute of Informatics Problems of the Russian Academy of Sciences /
44-2 Vavilova St., Moscow, 119333, Russia
anna.zalizniak@gmail.com

Abstract

The article analyzes the meaning of Russian discursive words *vidimo* and *po-vidimomu* ('apparently'), and reconstructs the ways of their semantic evolution over the past two centuries. It is shown that the meaning of an inference made by the speaker on the basis of some data, which is the only one for both words in modern language, arose in different ways. The semantic evolution of both words includes the the replacement of the meaning of visual perception with the meaning of epistemic evaluation and the acquisition of egocentric semantics. The word *vidimo* initially served as a marker of a true visual impression; the word *po-vidimomu* which initially included an interpretative component, acquired the meaning of a potentially false judgment, which was subsequently lost. The research is based on texts included in the Russian National Corpus (www.ruscorpora.ru).

Key words: Russian language, semantics, semantic evolution, discourse markers, visual perception, epistemic evaluation, inference

DOI: 10.28995/2075-7182-2021-20-720-728

Дискурсивные слова *видимо* и *по-видимому*: актуальная и диахроническая семантика[1]

Анна А. Зализняк
Институт языкознания РАН /125009 Москва, Б. Кисловский пер. 1;
ФИЦ ИУ РАН /119333 Москва, ул. Вавилова 44-2
anna.zalizniak@gmail.com

Аннотация

В статье анализируется значение русских дискурсивных слов *видимо* и *по-видимому* и восстанавливаются пути их семантической эволюции на протяжении двух последних веков. Демонстрируется, что значение умозаключения, сделанного говорящим на основании каких-то данных, являющееся в современном языке единственным для обоих слов, возникло разными путями. Семантическая эволюция обоих слов включает замену значения зрительного восприятия на значение эпистемической оценки и приобретение эгоцентрической семантики. При этом слово *видимо* исходно служило маркером истинности зрительного впечатления; слово *по-видимому*, изначально включавшее компонент интерпретации зрительных данных, приобрело значение потенциально ложного предположения, которое впоследствии было утрачено. Исследование проведено на материале текстов Национального корпуса русского языка (www.ruscorpora.ru).

Ключевые слова: русский язык, семантика, семантическая эволюция, дискурсивные слова, зрительное восприятие, эпистемическая оценка, инференция

[1] Работа выполнена при частичной поддержке РФФИ, грант № 19-012-00505.

1 Вводные замечания

Дискурсивные слова *видимо* и *по-видимому* в современном русском языке практически синонимичны: оба обозначают умозаключение (предположение), сделанное говорящим на основании каких-то данных[2]. С функционально-стилистической точки зрения они различаются по двум параметрам: *видимо* преобладает в художественной и устной речи, а *по-видимому* – в научно-деловой и письменной[3]. Однако в течение всего XIX и почти до конца XX века оба эти слова имели более широкий круг значений, среди которых были существенно различные.

Исследователями неоднократно отмечался тот факт, что если что-то «видно», то это может означать что это так и есть и что это, наоборот, «только кажется» (ср., напр. [10: 167, 172]; [4: 95]). Ю.Д.Апресян [1: 40] отмечает, что значение «истинного» и «ложного» перцептивного впечатления может совмещаться в пределах одной лексемы, иллюстрируя этот тезис глаголами *выглядеть*, *показаться*, *послышаться*. Значение «ложного впечатления» имеют некоторые слова, производные от основы *вид-*, ср. *привидеться*, *видимость*, а также выражения *с виду* (*он хорош, да зелен*), *сделать вид*; это значение, помеченное как «уходящее», выделяется также в [2: 123] у прилагательного *видимый* (ср. *видимое благополучие*).

Эта двойственность возникает из-за того, что зрительное восприятие по своей природе включает элемент интерпретации – которая может оказаться правильной или неправильной. Интерпретативный компонент в разных условиях обнаруживает себя в большей или меньшей степени. Так, сам глагол *видеть* в конструкции с изъяснительным придаточным с *что* (за исключением контекста синхронного репортажа) вводит умозаключение – в отличие от конструкции с *как*, вводящей описание самой воспринимаемой зрительно ситуации[4]. Это умозаключение представляет собой результат интерпретации данных – как перцептивных, так и любых других, тем самым в конструкции с придаточным с *что* глагол *видеть* может утрачивать компонент зрительного восприятия, ср.:

(1) ***Видя, что*** *толку не доищешься*, Фекла поспешно подобрала подол и продолжала путь. [Д. В. Григорович. Бобыль (1847)]
(2) Урок был большой и трудный, я ничего не знал и ***видел, что*** *уже никак не успею хоть что-нибудь запомнить из него* <…>. [Л. Н. Толстой. Отрочество (1854)]
(3) Сейчас я ***вижу, что*** *в случившемся было много иррационального, но был, между прочим, момент и вполне рациональный*. [Вера Белоусова. Второй выстрел (2000)]

Это умозаключение принимается субъектом за истинное; значения ложного впечатления глагол *видеть* выражать не может.

Обратимся теперь к значению производных от глагола *видеть* дискурсивных слов *видимо* и *по-видимому*.

2 *Видимо*

В словаре [6: I, 174] у *видимо* выделяется два значения: «*устар. нареч*. заметно, явно», иллюстрируемое примером (4), и «*в знач. вводн. сл*. вероятно, по-видимому», пример (5).

[2] В этот ряд входит также близкое по значению, но более разговорное (и обладающее рядом других особенностей) слово *видно*, а также еще более специфическое *видать*, которые в данной статье рассматриваться не будут. О словах *видимо, по-видимому, видно* см. также [12], [8], [13].
[3] По данным НКРЯ слово *видимо* употребляется в художественных текстах в 1,25 раза чаще, чем в официально-деловых, публицистических и научно-популярных (по классификатору НКРЯ; соответственно, 121,2 ipm и 97 ipm), слово *по-видимому*, наоборот, в 1,35 раза реже (соответственно, 52,3ipm и 70,6 ipm). Причем эти обобщенные подсчеты существенно нивелируют реальную картину: некоторые авторы (напр. Бунин, Чехов, Платонов, Шолохов) в своих художественных текстах слова *по-видимому* вообще практически не употребляют. С другой стороны (по подсчетам М.Ю.Михеева, устное сообщение), в художественной прозе Тынянова слово *по-видимому* имеет частоту 82 ipm, а в научных текстах – 698 ipm, т.е. в научных текстах встречается в 8,5 раз чаще. В письменной речи слово *видимо* употребляется в 1,3 раза реже, чем в устной (соответственно, 108 ipm и 142 ipm), а слово *по-видимому*, наоборот, в письменной речи встречается в 1,6 раза чаще, чем в устной (соответственно, 62 ipm и 38,5 ipm).
[4] Об этом различии конструкций с *что* и с *как* см. [9: 87], [14: 451].

(4) Казалось, готовилась гроза; черные тучи росли и ползли по небу, ***видимо*** *меняя свои дымные очертания*. (Тургенев. Первая любовь)
(5) Он взял Подгорина под руку и все уводил его вперед, ***видимо*** *собираясь поговорить с ним*. (Чехов. У знакомых)

В первом значении слово *видимо* является наречием образа действия (т.е. ‘таким образом, что это заметно’[5]) и указывает на реальное зрительное восприятие названой в предложении ситуации (ср. пример (4), где *тучи* меняют свои очертания именно *видимо*, т. е. явно). Это наречие, в свою очередь, производно от пассивного причастия в составе глагольной формы, ср:

(6) Как император, так и императрица были очень довольны своим коронованием, и *это* довольство ***было видимо*** для всех. [П. И. Ковалевский. Император Павел I (1900-1910)]

В значении наречия образа действия слово *видимо* несет на себе фразовое ударение и является маркером истинности описываемого впечатления. Это значение, диахронически исходное, оставалось актуально в течение XIX в., ср.:

(7) Она подала мне свою руку, холодную, слабую и дрожащую; *грудь ее* ***видимо*** *подымалась и опускалась* [Н. М. Карамзин. Письма русского путешественника (1793)]
(8) Провидение сохранило Россию. Можно сказать, что *Оно* ***видимо*** *хранит и начинающееся царствование*. Какой день был для нас 14-го числа! В этот день всё было на краю погибели: минута, и всё бы разрушилось. [В. А. Жуковский. Письмо к А.И.Тургеневу (1825)]
(9) Его некрасивое, но выразительное лицо, с глазами, блестящими умом и энергией, было ***видимо*** *взволновано*. [Н. Э. Гейнце. Коронованный рыцарь (1898)]

На протяжении XX в. его употребительность постепенно сокращается, и для языка XXI в. оно является безусловно устаревшим, хотя и продолжает изредка встречаться, ср.:

(10) Когда ***видимо*** *стало распадаться на части* живое тело Державы, казавшееся могучим, <…> он не приметил начавшегося разрушенья — и саморазрушенья тем более. [В. Г. Галактионова. 5/4 накануне тишины // «Москва», № 11, 12. 2004]

Во втором значении *видимо* является вводным словом, указывающим на ментальный акт предположения[6]; это предположение может касаться реконструируемой говорящим причины некоторого известного говорящему положения дел (ср. в примере (11): замолчали – потому что устали), в том числе – цели субъекта описываемого действия, ср. (описание ситуации, относительно причины/цели которой делается предположение, выделено подчеркиванием):

(11) Они еще немного о чем-то поговорили и, ***видимо*** *устав от разговоров, еды и вина,* замолчали. [Сергей Шикера. Египетское метро // «Волга», 2016]
(12) Отец не стал заканчивать фразу, ***видимо****, предлагая теперь высказаться сыну*. [Ю. И. Лунин. Три века русской поэзии // «Волга», 2016]
(13) Она вообще всегда старалась держать меня в стороне от своей банкирской жизни, ***видимо***, *рассчитывая, что мы с Петькой будем только стричь купоны*, занимаясь при этом своими делами. [Вера Белоусова. Второй выстрел (2000)]
(14) В результате мне всё-таки дали роль, ***видимо***, *для очистки совести*. [Сати Спивакова. Не всё (2002)]
(15) Однако он снова, — ***видимо****, для того, чтобы себя испытать*, — взбивал свою тощую подушку, ложился на усыпанную колючими крошками койку, открывал книгу [Ю. И. Лунин. Три века русской поэзии // «Волга», 2016]

Слово *видимо* в этом значении является безударным (и характеризуется определенным интонационным контуром, характерным для вводных слов); на письме вводные слова обычно выделятся запятыми. Компонент зрительного восприятия оказывается в нем утрачен: здесь

[5] Ср. значение 2 прилагательного *видимый* в [2: 123]: ‘такой, наличие которого заметно’: *Они поссорились без видимых причин.*
[6] Ср. замечание Н.Д.Арутюновой [3: 816], касающееся, в том числе, слов *видимо* и *по-видимому*: «[…] попадая в позицию вводного слова, почти все предикаты истинности, необходимости, восприятия и знания получают значение предположительности».

происходит тот же семантический сдвиг, что и в самом глаголе *видеть* в конструкции с придаточным с *что* (ср. выше): *видимо* вводит умозаключение, сделанное на основании каких-то данных, в общем случае не связанных с реальным зрительным наблюдением, ср. примеры (11)-(17).

(16) В молодости он, ***видимо***, *был страстно в неё влюблен*. [Сати Спивакова. Не всё (2002)]
(17) Что ж, страдающим от пыльцы аллергикам отныне, ***видимо***, *следует взять за привычку ежедневно читать о ботанике в Интернете*. [Дмитрий Анохин. Березовая каша для аллергиков // «Вечерняя Москва», 2002.05.16]

Итак, слово *видимо* имеет два отчетливо противопоставленных значения: значение маркера истинности зрительного впечатления и маркера предположения, сделанного на основании аргументов любого рода. Эти два значения, однако, в определенных условиях могут оказаться трудноразличимы – и именно такого рода случаи обеспечивают соответствующий семантический переход. Наиболее показательным в этом отношении является контекст описания чужого внутреннего состояния, обнаруживаемого по каким-то внешним признакам.

Так, в примере (18) из воспоминаний брата Ф.М.Достоевского ситуация 'его лицо пылало счастьем' не может быть результатом умозаключения: это именно непосредственно наблюдаемая ситуация, и тем самым *видимо* здесь имеет перцептивное значение[7].

(18) Раз как-то помню еще в начале моего пребывания в Москве, приезжает утром к Карепиным Александр Павлович Иванов; все лицо его было радостное и, ***видимо***, *пылало счастием*. [А. М. Достоевский. Воспоминания (1896)]

В примере (19) *видимо* выражает, наоборот, предположение:

(19) Дискретный Билль, — который, ***видимо***, *гордился* тем, что мог творить чудеса одной рукой, — принес открытые им жестянки пива. [В. В. Набоков. Лолита [автоперевод с английского] (1967)]

Однако относительно примеров (20)-(23) трудно сказать, вводит ли слово *видимо* описание непосредственно наблюдаемой ситуации или предположение, сделанной на основании ее наблюдения.

(20) <…> молодой человек лет двадцати, подслеповатый и белокурый, с ног до головы одетый в чёрную одежду, ***видимо*** *робел*, но язвительно улыбался… [И. С. Тургенев. Гамлет Щигровского уезда (1849)]
(21) — Еще бы отказаться, — пробасил Ракитин, ***видимо*** *сконфузившись*, но молодцевато прикрывая стыд, — это нам вельми на руку будет, дураки и существуют в профит умному человеку. [Ф. М. Достоевский. Братья Карамазовы (1880)]
(22) В конце июля 1860 года я по старой памяти отправился из Новоселок на Неручь на охоту, избрав главным центром сельцо Ивановское, имение моего зятя А. Н. Шеншина, женатого на родной сестре моей Любиньке. Они, ***видимо***, *обрадовались* моему приезду и старались по возможности устроить меня поудобнее [А. А. Фет. Мои воспоминания / Часть I (1862-1889)]
(23) — Примите меры, доктор, умоляю, — истерически крикнула девица. На лестницу выбежал секретарь филиала и, ***видимо***, *сгорая от стыда и смущения*, заговорил, заикаясь: [М. А. Булгаков. Мастер и Маргарита, часть 1 (1929-1940)]

3. *По-видимому*

Слово *по-видимому* в современном русском языке имеет единственное значение, а именно, оно касается положения дел, относительно которого говорящей не располагает знанием и на

[7] Выделение его запятыми здесь очевидно является ошибочным.

основании каких-то имеющихся в его распоряжении данных, формирует мнение, что дела обстоят именно так[8]. Ср.:

(24) Старушка, ***по-видимому***, *привыкшая к таким странным выходкам*, смотрела на него без удивления. [М. Ю. Лермонтов. Княгиня Лиговская (1836-1837)]

(25) Над тем, как функционирует бесписьменный язык, *ему*, ***по-видимому***, *вообще не приходилось задумываться*. [А. А. Зализняк. Лингвистика по А. Т. Фоменко // «Вопросы языкознания», 2000]

Заметим, что это значение появляется у слова *по-видимому* уже в XVIII в., ср.:

(26) Противъ самаго входу на горѣ находятся остатки низменнаго вала, который ***по видимому*** *для защищенія сего удобнаго входу былъ сдѣланъ*. [И. И. Лепехин (1770)]

(27) Ему не мудрено было меня помнить; и онъ, ***по видимому***, *не забывая ни стараго ни новаго состоянія своего*, предавался поперемѣнно то робости, то спѣси. [А. С. Шишков. Записки (1780-1814)]

Для значения слова *по-видимому* (как и для актуального на сегодня значения слова *видимо*) существенны следующие три обстоятельства. Во-первых, основанием для данного мнения могут служить данные любого рода, в том числе – не перцептивные (т.е. то, что нельзя «видеть» в прямом смысле). Так, фраза *По-видимому она обиделась* может быть произнесена скорее на основании фактов типа того, что она не отвечает на звонки или отказалась принять приглашение, чем на основании наблюдения ее выражения лица (в последнем случае уместнее *кажется* или *похоже*)[9]. Во-вторых, высказываемое при помощи *по-видимому* мнение принадлежит говорящему (или, в нарративе – повествователю): оно не может передаваться другому лицу (т.е. *по-видимому* является «жестким эгоцентриком» по Е.В.Падучевой [11]). В-третьих, хотя высказываемое при помощи *по-видимому* мнение может быть разной степени уверенности, говорящий все же его придерживается, т.е. в некотором смысле принимает за истинное: по крайней мере, он не может продолжить высказывание с *по-видимому* словами *…а может быть / а на самом деле / но оказалось, что это не так* – в отличие от таких выражений как *с виду*, *на вид*, *на первый взгляд*, такое продолжение легко допускающих.

Все эти обстоятельства представляют собой результат семантической эволюции слова *по-видимому* – от соответствующего его внутренней форме исходного значения ‘на основании того, что X видел, он сделал вывод, что вероятно P’ (ср. пример (28) ниже), т.е. указания на вывод, сформировавшийся на основании некоторого зрительного впечатления, изначально не обладающего перечисленными выше ограничениями (назовем его значением «мягкой инференциальной оценкой»), к современному значению «жесткой инференциальной оценки» ‘на основании каких-то данных я считаю, что скорее всего P’.

Однако на протяжении всего XIX в. сохраняется возможность употребления *по-видимому* в исходном значении «мягкой инференциальной оценки», в котором оно вводит мнение, обладающее следующими признаками: (i) оно представляет собой интерпретацию **зрительного** впечатления; (ii) субъектом этого мнения является **лицо, отличное от говорящего** (повествователя); (iii) это мнение не охарактеризовано относительно истинности, т.е. является **потенциально ложным**.

Рассмотрим следующие примеры.

(28) Но Эльфрида падаетъ къ ногамъ его и требуетъ помилованія. *Эдгаръ* ***по видимому*** *смягчается*, *уничтожаетъ приговоръ свой*, говоритъ о другомъ средствѣ къ примиренію, не сказывая онаго, и велитъ Ательвольду слѣдовать за нимъ въ лѣсъ. Эльфрида думаетъ, что супругъ ея прощенъ; ***но*** скоро радость ея перемѣняется въ отчаяніе, когда приходятъ ей сказать, что Король закололъ Ательвольда, вызвавъ его на поединокъ. [Н. М. Карамзин. Московской театръ // Московской журналъ. Часть VI, 1792]

[8] В [6: III, 159] у слова *по-видимому* отмечается единственное значение «вероятно, должно быть, весьма возможно». В [8] у этого слова различается несколько «контекстных конфигураций» значения в зависимости от его места в дискурсивной последовательности. В рамках нашего исследования эти различия не учитываются.

[9] Ср. обсуждение аналогичных примеров в [7: 92], [5: 303].

(29) Все любили молодого учителя — Кирила Петрович за его смелое проворство на охоте, Марья Кириловна за неограниченное усердие и робкую внимательность, Саша — за снисходительность к его шалостям, домашние за доброту и за *щедрость* ***повидимому*** *несовместную с его состоянием*. [А. С. Пушкин. Дубровский (1833)]

В примере (28) вывод, что Эдгар смягчился и отменил свой приговор, принадлежит Эльфриде (признак (ii)), он сделан на основании наблюдения поведения Эдгара (признак (i)) и впоследствии он оказывается ложным (признак (iii)). В примере (29) мнение, что щедрость молодого учителя несовместима с его состоянием – это предположение, принадлежащее домашним Кирилы Петровича (признак (ii)), сделанное ими на основании имеющихся в их распоряжении сведений (как мы знаем, ошибочных, признак (iii)).

Употребления, где *по-видимому* вводит потенциально ложный вывод, который может быть кем-то сделан на основании зрительного впечатления или каких-то других данных, а говорящий считает (знает), что это не так, широко представлены в текстах XIX и начала XX в. Ложность этого вывода может быть выражена разными способами (*но*, *а на самом деле* и др.), ср. (опровержение вводимого словом *по-видимому* впечатления выделено подчеркиванием):

(30) В дверях, когда он уже оделся, показалась фигура Шацкого, — который, ***по-видимому***, *небрежно смотрел на публику*, а на самом деле внимательно следил за Карташевым и не верил, что он действительно уйдет. [Н. Г. Гарин-Михайловский. Студенты (1895)]

(31) Вот он сидит, *по-**видимому**, тихоня,* а на самом деле ужасная язва. [А. М. Федоров. Его глаза (1913)]

(32) Работа с самого его детства шла, ***по-видимому***, *самая разнообразная*, но, в сущности, все одна и та же, состоящая в том, чтобы во всех делах, представлявшихся ему на пути, достигать совершенства и успеха, вызывающего похвалы и удивление людей. [Л. Н. Толстой. Отец Сергий (1890)]

(33) Самый теперешний социализм французский, — ***по-видимому***, *горячий и роковой протест против идеи католической* всех измученных и задушенных ею людей и наций <…> — самый этот протест <…> есть не что иное, как лишь вернейшее и неуклонное продолжение католической идеи. [Ф. М. Достоевский. Дневник писателя. 1877. Год II-й (1877)]

(34) Но они обречены на застой и на смерть, несмотря на всю, ***по-видимому***, *энергию их и тоску сердца их*. [Ф. М. Достоевский. Из записных тетрадей (1877)]

В более поздних текстах это значение представлено единичными примерами, ср.:

(35) Св[ятой] Исаак сказал: «Ничего нет каждому полезнее, как совет свой». А совет чуждый, хотя — ***по-видимому*** *состоящий из благих и разумных слов*, приносит душе лишь мучение, расстройство [монахиня Игнатия (Петровская). Святитель Игнатий — богоносец российский (1980-1990)]

(36) Моль заводится при сохранении сырья в сырых местах и хотя, ***по-видимому**, она подбивает одну шерсть,* но на самом деле она этим не ограничивается и портит лицо кожи. [Краткая энциклопедия скорняка, 1999]

(37) И вдруг засветил Руслану в глаз. — Ого, готов чувачок, — Глухой уже вызвал скорую. Он был, ***по-видимому***, *спокоен*, но какое уж тут спокойствие. [Н. Б. Черных. Слабые, сильные. Часть вторая // «Волга», 2015]

С точки зрения современной нормы такое употребление является устаревшим: для выражения этого значения используются дискурсивные слова *с виду*, *на первый взгляд*, *казалось бы* и т.п.

4 Эффект ближней семантической эволюции

Как мы видим, на протяжении последних двух веков значение слов *видимо* и *по-видимому* довольно существенно изменилось. А именно, для слова *видимо* значение 'так, что это было видно' стало устаревшим, а слово *по-видимому* утратило значение потенциально ложного умозаключения, сделанного кем-то помимо говорящего на основании зрительного впечатления. При этом, наряду с контекстами, в которых эти слова в современном языке употреблены быть не могут (ср. *грудь ее видимо подымалась* в (7) или *по-видимому, горячий и роковой протест* в (33)),

имеются и такие, в которых эти слова могут быть употреблены в современном значении. Это обстоятельство порождает «эффект ближней семантической эволюции» (см. [15]), т.е. ошибочное понимание этих слов в новом значении. Рассмотрим следующий пример:

(38) Мы условились драться за скирдами, что находились подле крепости, и явиться туда на другой день в седьмом часу утра. *Мы разговаривали,* ***повидимому***[10]*, так дружелюбно*, что Иван Игнатьич от радости проболтался. «Давно бы так» — сказал он мне с довольным видом; — «худой мир лучше доброй ссоры, а и нечестен, так здоров». [А. С. Пушкин. Капитанская дочка (1836)]

Современный читатель прочитывает это предложение, как если бы здесь было сказано «По-видимому, мы разговаривали так дружелюбно, что…», где *по-видимому* вводит предложение говорящего (повествователя), не обращая внимание на неуместность такого понимания. На самом деле слово *по-видимому* в (38) употреблено в утраченном сегодня значении потенциально ложного вывода на основании зрительного впечатления (значение, выражаемое в современном языке словом *с виду*); эта фраза выражает не предположение Гринева, а его утверждение, что его разговор со Швабриным имел видимость дружелюбного.

Об ошибочном понимании современным читателем значения слова *видимо* как вводного в примере (39) свидетельствует выделение его запятыми, в первоначальном издании отсутствующее[11]. Здесь слово *видимо* – это наречие, означающее ‘[Акулина привыкала к лучшему складу вещей], и это было видно’; ср. аналогичное значение слова *приметно* во второй части фразы.

(39) *Акулина,* ***видимо****, привыкала к лучшему складу речей,* и ум её приметно развивался и образовывался. [А. С. Пушкин. Барышня-крестьянка (1830)]

Существенно также, что описываемое впечатление принадлежит персонажу (Алексею), и оно является ложным: крестьянка Акулина – это на самом деле барышня Лиза, чей склад речей был и так достаточно хорош.

Эффект ближней семантической эволюции демонстрирует также пример (18).

5 Заключение

Слова *по-видимому* и *видимо* в современном русском языке являются довольно точными синонимами и представляют собой дискурсивные единицы – вводные слова, выражающие предположение говорящего, основанное на обработке каких-то данных (как перцептивных, так и любых других) и принимаемое им − за неимением лучшего − за истинное.

В XIX и частично в XX в. оба эти слова имели более широкий спектр значений: слово *видимо* употреблялось также в производной от пассивного причастия функции наречия со значением подтверждения истинности зрительного впечатления, а слово *по-видимому* (также в соответствии с его внутренней формой ‘согласно тому, что видно’) имело, в том числе, значение впечатления, которое, наоборот, может оказаться ошибочным. На протяжении последних двух веков оба эти значения были практически утрачены; оба слова имеют в современном языке единственное значение «уверенного предположения». Ход семантической эволюции единиц *видимо* и *по-видимому* обеспечивается тем обстоятельством, что обе границы – между непосредственной фиксацией увиденного и результатом вывода, сделанного на основании перцептивных данных, а также между истинным и ложным выводом на основании зрительного впечатления – являются проницаемыми.

Я благодарю Н.В.Перцова за содержательные замечания к первоначальному варианту статьи и любезно предоставленные мне ссылки на прижизненные издания пушкинских текстов, а также анонимных рецензентов Диалога за замеченные ими недостатки, которые были по возможности исправлены в окончательной версии статьи.

[10] В прижизненном издании («Современник», 1836, том IV, с. 85) написание раздельное: *по видимому*.

[11] См. Повѣсти покойнаго Ивана Петровича Бѣлкина, изданныя А. П. СПб. Печатано въ Типографіи Плюшара. 1831, с. 179. В целом, однако, наличие или отсутствие выделения запятыми в текстах XIX в. не может служить критерием выбора значения обсуждаемых слов, поскольку оно проводилось непоследовательно.

Литература

[1] Апресян Ю.Д. (2008), От истины до лжи по пространству языка. // Логический анализ языка. Между ложью и фантазией. М.: Индрик. С. 23–45.

[2] Апресян Ю.Д. (ред.) (2014), Активный словарь русского языка. Т. 2. М., 2014.

[3] Арутюнова Н.Д. (1998), Язык и мир человека. М.: Языки русской культуры.

[4] Арутюнова Н.Д. (2008), Ви́денье и виде́нье // Логический анализ языка. Между ложью и фантазией. М.: Индрик. С. 92–105.

[5] Булыгина Т.В., Шмелев А.Д. (1997), Языковая концептуализация мира (на материале русской грамматики). М.: «Языки русской культуры», 1997.

[6] Евгеньева А.П. (ред.) (1999), Словарь русского языка. Под ред. А.П.Евгеньевой. В 4-х тт. 4-е изд., М., 1999.

[7] Иоанесян Е.Р. (1993), Классификация ментальных предикатов по типу вводимых ими суждений // Логический анализ языка: Ментальные действия. Москва, 1993. С. 89-95.

[8] Киселева К., Пайар Д. (2003), Механизмы семантического варьирования на примере группы единиц с корнем *вид-*: ВИДИМО, ПО-ВИДИМОМУ, ВИДНО // Д.Пайар (ред.) Дискурсивные слова русского языка: контекстное варьирование и семантическое единство. М., 2003. С.50-79.

[9] Кобозева И.М. (1988), Отрицание в предложениях с предикатами восприятия, мнения и знания // Логический анализ языка. Знание и мнение. Москва. С.82-97.

[10] Кустова Г.И. (2004), Вид, видимость, сущность // Сокровенные смыслы. Слово. Текст. Культура. Сборник статей в честь Н.Д.Арутюновой. М., 2004. С. 155-175.

[11] Падучева Е.В. (2018), Эгоцентрические единицы языка. М.: Языки славянских культур, 2018.

[12] Разлогова Е.Э. (1996), Модальные слова и оценка степени достоверности высказывания // Русистка сегодня, 1996/3. С. 21-47.

[13] 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. М., 2017. Т.2. С. 370–380.

[14] Зализняк Анна А. (2006), Многозначность в языке и способы ее представления. М.: «Языки славянских культур», 2006.

[15] Зализняк Анна А. (2012), Об эффекте ближней семантической эволюции. // PHILOLOGICA. Под ред. А. С. Белоусовой, И. А. Пильщикова. 2012, том 9, № 21/23, с. 11-22.

References

[1] Apresyan Yu. D. (2008), From truth to falsehood in the space of language [Ot istiny do lzhi po prostranstvu yazyka], Logical analysis of language. Between lie and fantasy [Logicheskij analiz yazyka. Mezhdu lozh'yu i fantaziej] Moskva: Indrik. pp. 23–45.

[2] Apresyan Ju.D. (ed.) (2014), Active dictionary of Russian language [Aktivnyj slovar' russkogo yazyka]. Vol. 2.

[3] Arutyunova N.D. (1998), The language and the human world [Yazyk i mir cheloveka]. Moskva: Yazyki russkoj kul'tury, 1998.

[4] Arutyunova N.D. (2008), Vision and phantom [Vídenie i vidénie] Logical analysis of language. Between lie and fantasy [Logicheskij analiz yazyka. Mezhdu lozh'yu i fantaziej] Moskva: Indrik. pp. 92–105.

[5] Bulygina T.V., Shmelev A.D. (1997), Linguistic conceptualization of the world [Yazykovaya konceptualizaciya mira (na materiale russkoj grammatiki)]. Moskva: Yazyki russkoj kul'tury.

[6] Evgen'eva A.P. (ed.) (1999), Dictionary of Russian language [Slovar' russkogo yazyka]. In 4 volumes. 4th edition.

[7] Ioanesyan E.R. (1993), Classification of mental predicates by the type of judgments they intruduce [Klassifikaciya mental'nyh predikatov po tipu vvodimyh imi suzhdenij], Logical

analysis of language: Mental actions [Logicheskij analiz yazyka: Mental'nye dejstviya.] Moskva, pp. 89-95.

[8] Kiseleva K., Pajar D. (2003), Mechanisms of semantic variation on the example of a group of units with a root *vid-*: VIDIMO, PO-VIDIMOMU, VIDNO [Mekhanizmy semanticheskogo var'irovaniya na primere gruppy edinic s kornem *vid-*: VIDIMO, PO-VIDIMOMU, VIDNO], D.Paillard (ed.) Discursive words of the Russian language: contextual variation and semantic unity [Diskursivnye slova russkogo yazyka: kontekstnoe var'irovanie i semanticheskoe edinstvo]. Moskva, pp. 50-79.

[9] Kobozeva I.M. (1988), Negation in sentences with predicates of perception, opinion and knowledge [Otricanie v predlozheniyakh s predikatami vospriyatiya, mneniya i znaniya], Logical analysis of language. Knowledge and opinion [Logicheskij analiz yazyka. Znanie i mnenie]. Moskva, pp. C.82-97.

[10] Kustova G.I. (2004), View, visibility, essence [Vid, vidimost', sushchnost'], Secret meanings. Word. Text. Culture. Collection of articles in honor of N.D. Arutyunova [Sokrovennye smysly. Slovo. Tekst. Kul'tura. Sbornik statej v chest' N.D.Arutyunovoj]. Moskva, pp. 155-175.

[11] Paducheva E.V. (2018), Egocentric units of language [Egocentricheskie edinicy yazyka]. Moskva: Yazyki slavjanskikh kul'tur.

[12] Razlogova E.E. (1996), Modal words and the degree of reliability of the statement [Modal'nye slova i ocenka stepeni dostovernosti vyskazyvaniya], Rusistka segodnya [Russistics today], №3, pp. 21-47.

[13] Shmelev A.D., Zaliznyak Anna A. (2017), Reverse translation as a tool for linguistic analysis of discourse words [Reversivnyj perevod kak instrument lingvisticheskogo analiza diskursivnyh slov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialog 2017 [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Trudy Mezhdunarodnoj Konferencii Dialog 2017]. Moskva, Vol 2, pp. 370–380.

[14] Zaliznyak Anna A. (2006), Polysemy in language and ways of its representation [Mnogoznachnost' v yazyke i sposoby ee predstavleniya]. Moskva: Yazyki slavyanskikh kul'tur.

[15] Zaliznyak Anna A. (2012), On the effect of short-range semantic evolution [Ob effekte blizhnej semanticheskoj evoljucii], A.S.Belousova, I.A. Pil'shchikov (eds.) PHILOLOGICA. Vol 9, № 21/23, pp. 11-22.

Linking the Ancient Greek WordNet to the Homeric Dependency Lexicon

Chiara Zanchi
University of Pavia
Corso Strada Nuova 65
I-27100 Pavia PV, Italy
chiara.zanchi01@unipv.it

Silvia Luraghi
University of Pavia
Corso Strada Nuova 65
I-27100 Pavia PV, Italy
silvia.luraghi@unipv.it

Erica Biagetti
University of Pavia / Bergamo
Corso Strada Nuova 65
I-27100 Pavia PV, Italy
erica.biagetti01@universitadipavia.it

Abstract

The Ancient Greek WordNet is a new resource that is being developed at the Universities of Pavia and Exeter, based on the Princeton WordNet. The Princeton WordNet provides sentence frames for verb senses, but this type of information is lacking in most WordNets of other languages. In fact, exporting sentence frames from English to other languages is not a trivial task, as sentence frames depend on the syntax of individual languages. In addition, the information provided by the Princeton WordNet is not corpus-based but relies on native speakers' knowledge. This type of information is not available for dead languages, which are by definition corpus languages. In this paper, we show how sentence frames can be extracted from morpho-syntactically parsed corpora by linking an existing dependency lexicon of Homeric verbs (HoDeL) to verbs in the Ancient Greek WordNet. Given its features, HoDeL allows automatically extracting all subcategorization frames available for each verb along with information concerning their frequency as well as semantic information regarding the possible arguments occurring in specific frames. In the paper, we show our method to automatically link the two resources and compare some of the resulting sentence frames with the English sentence frames in the Princeton WordNet.

Keywords: WordNet; valency lexica; Ancient Greek; HoDeL; sentence frames
DOI: 10.28995/2075-7182-2021-20-729-737

Извлечение информации для древнегреческого WordNet из Словаря синтаксических зависимостей гомеровского греческого

Кьяра Дзанки
Университет Павии
Corso Strada Nuova 65
I-27100 Pavia PV, Italy
silvia.luraghi@unipv.it

Сильвия Лураги
Университет Павии
Corso Strada Nuova 65
I-27100 Pavia PV, Italy
chiara.zanchi01@unipv.it

Эрика Бияджетти
Университет Павии / Бергамо
Corso Strada Nuova 65
I-27100 Pavia PV, Italy
erica.biagetti01@universitadipavia.it

1 Introduction

WordNets (henceforth WNs) are lexical databases in which meaning is stored in a relational way [4]. They comprise nodes for lemmas to which meanings are associated in the form of synsets i.e. sets of cognitive synonyms accompanied by brief definitions. Lemmas are linked to each other by lexical relations, whereas semantic relations interlink synsets, resulting in a network of meaningfully related words and concepts.

In this paper we show how syntactic information can be extracted from a corpus-based dependency lexicon and automatically linked to a WN. Our pilot study[1] links the Ancient Greek WordNet (AGWN) to the Homeric Dependency Lexicon (HoDeL). In Sec. 2 we illustrate the properties of the AGWN. Sec. 3 contains a brief description of HoDeL and shows how we linked the data. Sec. 4 reviews how syntactic information is provided in some other WNs. In Sec. 5 we survey the sentence frames for two Ancient Greek (AG) verbs. In Sec. 6 we discuss our findings.

2 The family of WordNets for Ancient Indo-European languages

The AGWN belongs to a family of WNs for ancient Indo-European (IE) languages, an ongoing project jointly developed at University of Pavia, University of Exeter, and the Center for Hellenic Studies at Harvard University [1]. Besides AG, it currently comprises WNs for Sanskrit and Latin.

The architecture of our WNs aims to facilitate their integration with other resources for ancient languages, adopting a standardized set of lemma-based URIs to guarantee identification (e.g. in the case of ambiguous word forms) and allow easily tying together information from disparate databases (Sec. 2).

The Sanskrit, AG and Latin WNs were designed to be interoperable with each other and to enable the cross-linguistic comparison of linguistic structures. To enhance compatibility, we maximize usage of the synsets deriving from the Princeton English WordNet [4], and adopt a principled view of polysemy, which entails assuming that all literal and non-literal senses of a lemma can be organized in a structured semantic network. Literal senses are detected based on their early attestation, concreteness, and predominance in the network [7], while non-literal senses are divided into metaphorical and metonymical ones. E.g. synsets associated to the AG verb *pléō* ‘sail’ are classified as follows:

(1) **Literal senses ‘sail’, ‘float’**

a. v#01260993 | travel by boat

b. v#01299504 | float on the surface of water

(2) **Metonymic sense ‘depart’**

v#01262245 | move away from a place into another direction

(3) **Metaphoric sense ‘behave in a certain way’**

v#00007023 | behave in a certain manner

For each synset, we provide information on period(s), literary genre(s), author(s), and work(s) in which they are attested. Thus, the senses in (1)a-c feature the metadata in Tab. 1.

[1] The Ancient Greek WordNet does not include information about sentence frames, and in this paper we show our first attempts at linking the WordNet semantic information with syntactic information provided by HoDeL starting with two verbs. These verbs do not exhaust the amount of lemmas already annotated in the Ancient Greek WordNet.

Sense	Period	Genre	Loci
(1)a	Archaic (8th-6th BCE)	poetry epic historiography theater …	Il.3.444, Pi.P.4.69 X.Cyr.6.16 A.Ag.691
(2)	Archaic (8th-6th BCE)	epic	Od.12.5
(3)	Classical (5th c.-323 BCE)	theater philosophy oratory	S.Ant.190 D.19.250 Pl.Lg.813d

Table 1: Diachronic and stylistic metadata associated with the senses of *pléō* in (1)-(3)

Diachronic and stylistic metadata are meant to enable studies on semantic change over time and across literary genres and authors.

To allow comparison between the languages of our WNs, we added language-specific features. At the lexical level, we provide each lemma with morphological information, as summarized in Tab. 2.

Field	Subfield	Value
Etymology		PIE. **pleu-* 'float'
Lemma		*pléō*
POS		Verb
Morpho		v1spia--1e
Uri		π07077
Morphology	Principal Parts	*pleúsomai épleusa pépleuka epléusthēn pleusthḗsomai pèpleusmai*
	Prosody	*pléō*
Form Tokens	Form	v1sfia--1e
	Token	*plesoūmai* (FUT.IND.1SG)
	Alternative	✓ (late form of *pleúsomai*)

Table 2: Lemma annotation for *pléō*

We employ an extended set of lexical relations comprising derivation, lexical antonymy with privative *a-*, parasynthesis, composition, inclusion in multi-word expressions, lexicalized participles, exemplified in Tab. 3.

Relation	Label and Example	Inverse
Derivation	*makró-tēs* 'length' IS DERIVED FROM *makrós* 'long'	IS RELATED TO
Antonymy	*anomía* 'lawlessness' IS PRIVATIVE OF *nomía* 'lawfulness'	HAS PRIVATIVE
Parasynthesis	*ánoos* 'without understanding' IS PARASYNTHETIC OF *nóos* 'mind'	HAS PARASYNTHETON
Composition	*makropoiéō* 'enlarge' IS COMPOSED OF *makrós* 'big'; *makropoiéō* 'enlarge' IS COMPOSED OF *poiéō 'make'*	COMPOSES
Inclusion	*thalássia érga* 'navigation' INCLUDES *thalássios* 'related to the sea'; *thalássia érga* 'navigation' INCLUDES *érgon* 'work'	IS INCLUDED IN
Participle	*eikṓs* 'seeming like' IS PARTICIPLE OF *éoika* 'to be like'	HAS PARTICIPLE

Table 3: Family-specific Lexical Relations

3 Enriching the Ancient Greek WordNet with sentence frames: a pilot study

As a further step, we aim to integrate sentence frames in the metadata associated to each verbal entry of the AGWN. To do so, we propose a semi-automatic two-step workflow and present a pilot study of Homeric verbs. Ultimately, we aim to enhance the AGWN with sentence frames extracted from the whole Ancient Greek and Latin Dependency Treebank (AGLDT; https://github.com/PerseusDL/treebank_data). Working with the Homeric verbal lexicon has a number of advantages: (i) it allows

exploiting the user-friendly HoDeL (https://hodel.unipv.it/hodel-res) to support the annotators' work; (ii) focusing on a limited data-set allows evaluating the efficiency of the automatic step of our workflow.

HoDeL is a corpus-based lexicon of Homeric verbs induced from the Homeric poems treebanked at the AGLDT 2.0. As documented in [8] and [9], HoDeL was obtained with a series of SQL queries extracting from the analytical/syntactic layer of AGLDT 2.0 all finite and nonfinite verbal forms along with some of their child nodes (the ratio of data extraction, its expected and actual output are documented in [8] and [9]). Extracted data was stored in a backend relational database, which lies behind HoDeL online query interface.

HoDeL features most relevant for the purposes of the present paper are the following:

- it allows obtaining corpus-based data concerning sentence frames (in the form of syntactic subtrees), their frequency, and instantiations;
- it gives information about the syntactic (e.g. case marking) and semantic features (e.g. animacy) of verbal dependents;
- it provides aligned translation of Homeric passages.

Thus, HoDeL can be used by annotators with limited expertise on treebanks, .xml, SQL and PML-tree queries. Moreover, it eases the annotators' task to manually pair sentence frames with WN synsets, as it gives translations.

The first step to link HoDeL with the AGWN consisted in automatically mapping the Homeric verbs in HoDeL with the corresponding verbal entries in AGWN. We extracted, for each word in HoDeL backend database, the features 'lemma' and 'postag', and selected only words for which the first value of the postag is 'v' i.e. verbs. This extraction yielded 2,482 results. Then, we extracted verbal lemmas from the .csv files of AGWN in the relative Github repository (https://github.com/greekwordnet), selecting the value 'v' in the pos fields. This extraction yielded 28,405 results, which represent all verbal lexicon of AG. Finally, we automatically compared the outputs of the two extractions, summarized in Tab. 4:

Type of correspondence (HoDeL-to-Ancient Greek WordNet)	**Frequency**
one-to-one	2,256 (90.94%)
one-to-many (two)	41 (1,65%)
one-to-zero	185 (7.41%)
TOTAL	2,482 (100%)

Table 4: Results of the comparison of HoDeL and AGWN verbal lexicon

This pairing attempt is satisfactory, given that more than 90% of the Homeric verbs found a unique correspondent in the WN: in the vast majority of cases we were able to univocally associate URIs of the AGWN verbs to HoDeL verbs.

Cases of one-to-many correspondences are due to (i) homophonous verbs, represented by ambiguous characters strings, which need to be manually disambiguated: e.g. the string "δέω" (BetaCode: de/w) represents verbs assigned URIs "δ01019" and "δ01020", corresponding to the meanings 'tie' and 'lack'; (ii) verbs contained twice in the AGWN lexicon due to importation errors from the electronic version of the reference dictionary of AGWN, the Liddell-Scott-Jones: e.g. "τρέφω" being assigned the URIs τ02408 and τ02407 (this error is no longer in AGWN, and thus this study also offered the chance to improve the quality of AGWN base data).

One-to-zero correspondences are due to manyfold factors, which can be clustered as follows:

- different lemmatizations (due to voice (a), dialectal variants (b), pos assignments to defective participles used as nouns or adjectives (c));
- different notations, due to diacritics: hiatus (d), subscript iota (e);
- errors in AGWN: missing verbs (f) and/or incorrect pos assignment (g).

Mismatch type	HoDeL	AGWN
(a) different voice	μετατρέπομαι	μετατρέπω
(b) variants	διαπέταμαι	διαπέτομαι
(c) participles	θυμηγερέων (v)	θυμηγερέων (n)
(d) hiatus	διικνέομαι	διϊκνέομαι
(e) subscript iota	θνήσκω	θνῄσκω
(f) missing verbs	καταβρόχω	καταβρόξειε (n)
(g) incorrect pos	μέμονα	μέμονα (n)

Table 5: Summary of mismatch types

As the number of non-one-to-one correspondences is manually manageable, such mismatches can be fixed. Crucially, this pilot study also helped us identify notations and lemmatizations that require standardization before mapping the entire AG verbal lexicon from AGLDT 2.0 to AGWN.

Once AGWN URIs are assigned to all HoDeL verbs, sentence frames can be imported from the HoDeL backend database into AGWN metadata. Next, we will ask annotators to assign the extracted sentence frames to the correct synsets. Thus, sense disambiguation will be done manually. This task would be much more time-consuming and effortful for the annotators, had HoDeL not provided aligned English translations of the AG passages. Given that sentence frames need to be manually sorted to the correct synsets, it is more convenient to assign them to each lexicon entry while AGWN is still under construction. When the pairing of sentence frames and synsets will be complete, one will be able to determine the frequency distribution of sentence frames, the most frequent filling words for each frame-slot, and their associative connections with synsets, as these data can easily be extracted from HoDeL (synsets make it possible to generalize these associative connections across verbal lexemes).

4 Syntactic information about verbs' construction in WordNet

The Princeton WN provides limited information concerning the types of construction in which verbs can occur. It provides 35 sentence frames, which indicate "the number of noun arguments that the verb subcategorizes for" [4]. Verb entries in WN typically contain example sentences for each synset, and may also contain the relevant sentence frames.

E.g. the verb *announce* has four synsets. The first sense is 'denote' and has four sentence frames, see Fig. 1.

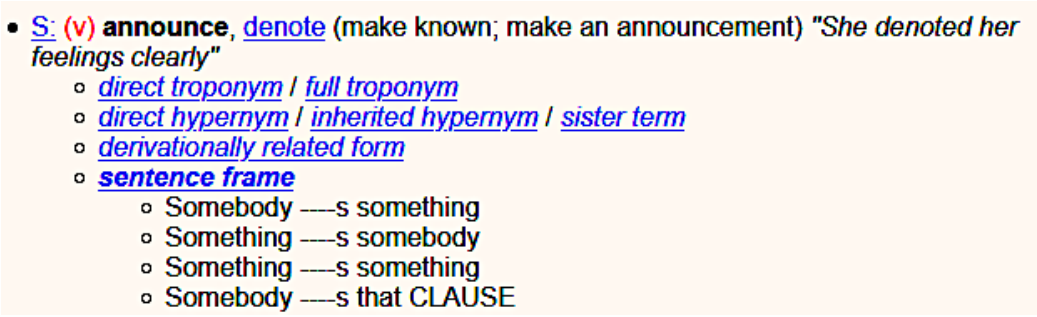


Figure 1: Sentence frames of *announce*

As highlighted by Fellbaum [4] "the noun slots in these syntactic frames are not at present linked to either thematic roles or semantic categories." Other resources have been used in interaction with WN to provide more information, notably FrameNet and VerbNet (see e.g. [6]), but a direct linking among these resources is not implemented in WN. In addition, sentence frames are constructed by the WN annotators based on their knowledge of the language, and the examples are not extracted from corpora.

As verbs' constructions are largely language specific, sentence frames cannot be exported from the Princeton WN when creating a new WN for a different language. For this reason, WNs available for other languages vary as to the type of syntactic information provided for verbs. Here we focus on the

German GermaNet (https://weblicht.sfs.uni-tuebingen.de/rover/), which provides information concerning the verbs' "actual language use by giving at least one example sentence per lexical unit. In order to link lexical semantic and syntactic information, details on syntactic subcategorisation are indicated for each lexical item." The information provided is partly corpus based, as the examples are "based on the Complementation Codes provided with the German Version 2.5 of Release 2 of the CELEX Lexical database [2]." For the verb *verkünden* 'announce' one finds three synsets. The first synset 'öffentlich sagen' ('announce') contains three verb frames with examples:

- NN.Dn.DS – *Sie verkündete, dass sie heiraten ihn würde*. "She announced that she was going to get married."
- NN.Dn.AN – *Der Sektenführer verkündete seinen Anhängern die frohe Botschaft*. "The leader of the sect announced his followers the good news."
- NN.AN – *Er verkündete die Fusion der beiden Unternehmen*. "He announced the merger of the two companies."

The abbreviations in the verb frames indicate the following:

- NN: grammatical subject that is realized as a noun phrase in the nominative case
- AN: obligatory Accusative Complement
- Dn: optional Dative Complement
- DS: complement clause (*dass*-phrase)

The Princeton WN and GermaNet differ as to the type of language specific information they provide. Thus, GermaNet provides information about case marking, as German nouns feature morphological case. On the other hand, information about verbal aspect is available in the Princeton WN indicated by ---s / is ---ing, while it is not available in GermaNet, as verbal aspect is not grammaticalized in German.

As for language independent information, GermaNet distinguishes between complements, which can be obligatory or optional as the accusative and the dative complement in the verb frames of *verkünden* mentioned above, and adverbials, typically prepositional phrases, which can also be obligatory or optional. E.g. the locative PP with *wohnen* 'live (in a place)' annotated as BL is obligatory, while the directional PP with *segeln* 'sail' annotated Bd is optional. On the other hand, the fact that nominals are simply indicated by N rather than by 'someone' or 'something' as in the Princeton WN has the effect that no information is provided about the animacy feature of the fillers.

5 Verb frames in HoDeL

In this section, we discuss the verb frames that we have extracted from HoDeL for two verbs, *pléō* 'sail, float' (28 occurrences; already validated) and *angéllō* 'announce' (27 occurrences; not yet validated). Both verbs only feature active forms in Homer. We tentatively adapt the annotation of GermaNet adding information about animacy. Note that AG is a null subject language, but we still indicate subject NPs as NN (nominative NP) as the subject triggers verbal agreement (there are no impersonal forms among the occurrences analyzed here). Given the limited extent of this pilot study, for the time being we refrain from giving a complete list of tags, but only use those that reflect actual occurrences of the two verbs.

Verbal aspect is grammaticalized in AG and interacts with tense [3]. Verbs may feature three aspectual stems, the present or imperfective, the aorist or perfective, and the perfect or resultative. In addition, the future tense is not sensitive to aspect. Hence, aspectual features need to be reflected in the annotation, as well as other morphological information that turns out to be relevant for the possible occurrence of a given sentence frame. In particular, aspectual information is essential as different aspectual stems of the same verb can show different values for voice (e.g. *gígnomai*:PRS.M/P 'become' vs. *gégona*.PF.ACT). We discuss this issue in the next sections.

5.1 *Pléō*

The verb *pléō* comprises three synsest 'float', 'sail', 'depart' (Sec. 1; the metaphoric sense is not relevant for Homeric Greek). All occurrences feature the present stem, hence they are imperfective, here

annotated as ...impf, except for one occurrence of the future tense. For all senses we found sentence frames 1 and 2.

1 NN(-a/+a) ...impf
a) FLOAT

Od. 5.240	*tá*	*hoi*	*plṓoien*	*elaphrôs*
	DEM.NOM.PL	3SG.DAT	float.OPT.PRS.3SG	lightly

'They (=the trees) would float lightly for him.'

b) SAIL

Od. 10.80	*ennêmar*	*mèn*	*homôs*	*pléomen*
	for_nine_days	PTC	alike	sail.IMPF.1PL

'We sailed for nine days alike (night and day).'

c) DEPART

For this sense we found the relevant sentence pattern with a future tense:
1[i] NN(-a/+a) ...fut

Od. 12.5	*háma*	*d'*	*ēoî*	*phainoménēphi*	*pleúsesth'*
	together	PTC	dawn.DAT	appearing.DAT	sail.FUT.MID.2PL

'At dawn you will depart.'

2 NN(-a/+a) ...impf PP
b) SAIL
2[i] Perlative PPs or adverbs (*epì oínopa pónton* 'over the wine-dark sea', Il. 7.88)
2[ii] Direction PPs
c) DEPART
Source PPs or adverbs (*apò Krḗtēs* 'from Krete' Od. 14.253; *póthen* 'where form?' Od. 3.71)

5.2 *Aggéllō*

The verb *aggéllō* is not yet fully annotated in the AGWN. Here we consider the synset #00659537 | make known.

Syntactically, it shows a more complex set of constructions than *pléō*, as it can also take subordinate clauses; in addition, it features both imperfective and perfective forms (the latter annotated as aor), and several future participles. We found the following four sentence frames.

1 NN(+a) ...ptcp.fut/aor Nd(+a)

Od. 16.150	*sú*	*g'*	*aggeílas*	*opísō*	*kíe*
	2SG.NOM	PTC	announce.PTCP.AOR.NOM	back	come.IMP.PRS.2SG

'After announcing, come back.'

Often with motion verbs: *ôrto ... aggeléousa* '(she) hastened announcing (aor.)'; *bê d' ímen aggeléōn* '(he) went announcing (fut.)'
A variant of this pattern is instantiated by a finite form of the verb in a single occurrence (Il. 9.617).
1[i] NN(+a) ...fut

2 NN(+a) ...impf/aor Na(-a) Nd(+a)

Il. 15.159	*Poseidáōni*	*ánakti*	*pánta*	*tád'*	*aggeîlai*
	P.DAT	king.DAT	all.ACC.PL	DEM.ACC.PL	announce.INF.AOR

'(Go) to announce all these things to kingly Poseidon.'

Variants of this frame include special values of the animacy feature.
2[i] NN(+a) NA (+a) 'bring information about someone' Od. 14.122-123
2[ii] NN(-a) NA(-a) Od. 13.93-13.94 with an inanimate subject *astḕr* 'star' and an inanimate object *pháos Ēoûs* 'the light of dawn'

Pattern 3 is instantiated by a single occurrence; hence the information about verbal aspect cannot be considered significant, in the light of the fact that other sentence frames allow both the imperfective and the perfective aspect.

3 NN(+a) ...aor ND(+a) INFINITIVE

Od. 16.350	*keínois*	*aggeílōsi*	*thoôs*	*oîkónde*
	DEM.DAT.PL	announce.SBJV.AOR.3PL	quickly	home
	néesthai			
	travel.INF.PRS.MID			

'(That they) announce them to quickly travel home.'

4 NN(+a) ...impf/aor COMPL CLAUSE

This sentence frame contains two subtypes that feature different types of complement clause:

4[i] NN(+a) AcI

Il. 8.517-8.519	*kḗrukes ...*	*aggellóntōn*	*paîdas ...*	*gérontas*
	herald.NOM.PL	announce.IMP.PRS.3PL	boy.ACC.PL	old_man.ACC.PL
	léxasthai			
	gather.INF.AOR.MID			

'Let the heralds announce that boys (and) old men gather.'

4[ii] NN(+a) ND *hótti* clause

Il. 22.438-439	*ou*	*gár*	*hoí ...*	*ḗggeil'*	*hóttí*	*rhá*
	NEG	PTC	3SG.DAT	announce.AOR.3SG	that	PTC
	hoi		*pósis*	*éktothi mímne*		*puláōn*
	3DAT.SG		spouse.NOM	outside abide.IMPF.3SG		gate.GEN.PL

'No (messenger coming) had announced her that her husband abode outside the gates.'

6 Discussion

As remarked in Sec. 5, in the description of the sentence frames we adopted the annotation from GermaNet. This implies also indicating whether NPs are obligatory or not. Clearly, we can gauge obligatoriness only based on the occurrences in the corpus on which HoDeL relies, the Homeric poems. Hence, when we indicate that the sentence frame 2[ii] NN(-a) NA(-a) with *aggéllō* has an obligatory accusative object, this claim is based on a single occurrence of this pattern. Given the limited number of occurrences, we did not find cases in which an omitted direct object is definite and referential [5]. Annotating a larger corpus would also imply deciding whether such types of omitted objects must still be considered obligatory (note that we considered subjects obligatory, even when they are omitted, as anticipated in Sec. 5).

The formulaic language of the Homeric poems accounts for the comparatively high frequency of certain patterns. Indeed, some expressions occur in the same form several times in the Homeric poems, and must be considered formulas, such as occurrences of *pléō* with certain PPs or adverbs (e.g. *énthen dè protérō pléomen* 'hence we sailed forth' occurs five times in the *Odyssey*). Sentence frame 1 of *aggéllō* NN(+a) PTCP Nd(+a) shows a typical usage of the participle, which often occurs in formulas in the Homeric poems, but is not limited to formulaic language and remains widespread throughout the history of AG.

Concerning verbal aspect/tense, in spite of the limited number of occurrences some considerations can be made. The two verbs show a different aspectual profile, as *pléō* only features imperfective forms, while *aggéllō* can be used both in the imperfective and in the perfective aspect. Moreover, sentence frame 1 of *aggéllō* is almost only limited to participles, with a preference for the future tense (7 occurrences out of 9, plus one occurrence of a finite form).

Acknowledgements

This research was carried out in the framework of the project *Dipartimenti di Eccellenza* 2018-2022 (Ministry of University and Research). This chapter results from joint work of the authors. For academic purposes, Erica Biagetti is responsible of Section 2, Chiara Zanchi of Section 3 and Silvia Luraghi of Sections 4 and 5. Sections 1 and 6 were written jointly by the authors.

References

[1] Biagetti E., Zanchi C., Short W. M. (2021), Toward the creation of WordNets for ancient Indo-European languages. Proceedings of the 11th Global WordNet Conference, S. Bosch, C. Fellbaum, M. Griesel, A. Rademaker, P. Vossen (eds), EACL/GWC, Global WordNet Association, pp.258-266.

[2] Baayen R H., Piepenbrock R., Gulikers L. (1995), CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium.

[3] Duhoux Y. (1992), The verb in Ancient Greek [Le verbe en grec ancien]. Peeters, Louvain-La-Neuve.

[4] Fellbaum C. (ed.) (1998), WordNet: An electronic lexical database. MIT Press, Cambridge, MA.

[5] Luraghi S. (2003), Definite referential null objects in Ancient Greek. Indogermanische Forschungen 108, pp. 169-196.

[6] 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, pp. 100-111. Springer, Berlin.

[7] Tyler A., Evans V. (2003), The Semantics of English Prepositions: Spatial Scenes, Embodied Meanings and Cognition. CUP, Cambridge, UK.

[8] Zanchi C., Luraghi S. (2020), Presenting Hodel—A New Resource for Research On Homeric Greek Verbs. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2020" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2020"], Moscow, Supplementary volume, pp.1188-1201.

[9] Zanchi,C. (forthc.). The Homeric Dependency Lexicon: what it is and how to use it. Journal of Greek Linguistics.

Russian predicatives and the ontology of states

Anton Zimmerling
Institute of linguistics, Russian Academy of Science/
Puskhin state Russian language Institute, Moscow, Russia
fagraey64@hotmail.com

Abstract

Basing on the frequency dictionary of Russian predicatives, I measure the volume of the lexical class of non-agreeing predicatives licensing the productive dative-predicative sentence pattern, where the predicative assigns dative case to its animate subject. The tested vocabulary includes 422 elements. Their frequency rates are derived from the main corpus of RNC using an approximation — the number of hits in the context "predicative + dative subject in 1Sg" in the window {-1; 1}. I argue that the Russian dative-predicative construction has an invariant meaning of internal state, i.e. spaciotemporal stative situation with a priority argument. However, most predicatives licensing dative-predicative structures in Russian also express external states, i.e. spaciotemporal stative situations without a priority argument, if used without overt referential dative subject. This can be proved both for words denoting physical sensations, cf. *X-y kholodno* 'X is cold' vs *kholodno* 'It is cold' and for some words denoting affections, cf. *tosklivo* 'dreary', 'sad', *X-y tosklivo* 'X feels sad' vs *zdes' tosklivo* 'It's dreary here'. The shift from internal state to external state is licensed in Russian. If a lexical item has regular uses in the dative-predicative structure, it generally can express the meaning of external state outside this structure. The reverse if false: if a lexical item has regular uses as an external state, cf. *vetreno* 'windy', *pyl'no* 'dusty', it only can have infrequent side uses with a dative subject. This asymmetry is confirmed by the corpus data. I check an additional list of words with the meaning of external state, measure their frequency rate in the context "predicative + dative subject in 1Sg" in the window {-1; 1} and compare them to standard dative predicatives.

Keywords: ontology, internal states, external states, predicatives, corpus grammar, Russian language
DOI: 10.28995/2075-7182-2021-20-738-748

Русские предикативы и онтология состояний

Циммерлинг А. В.
Институт языкознания РАН/Государственный институт
русского языка имени А. С. Пушкина, Москва, Российская федерация
fagraey@hotmail.com

Аннотация

В статье определяются предикатные значения внутреннего и внешнего состояния, строится онтология русской конструкции с предикативами дативно-предикативной структуры и проверяется гипотеза о том, что данная конструкция имеет инвариантное значение внутреннего состояния. Переход от внешних состояний к внутренним в русском языке затруднен, в то время как переход от внутренних состояний к внешним регулярен. Предикативы дативно-предикативной структуры при устранении позиции семантического субъекта могут выражать значение внешнего состояния. В этом случае предикатив может описывать внешне наблюдаемую референтную ситуацию, оцениваемую одинаково любым наблюдателем. Глагольные предложения с качественным наречием, коррелятивным предикативу ДПС, проецируют структуру события, содержащую отсылку либо к внутреннему, либо к внешнему состоянию.

Ключевые слова: онтология, внутренние состояния, внешние состояния, предикативы, корпусная грамматика, русский язык

1 Варьирование и лингвистические модели

1.1 Варьирование в лексике и грамматике

Модели классической лингвистики опираются на понятие инварианта. Варьирование изучают специальные дисциплины — диалектология, которая исследует территориальное распределение единиц и конструкций в идиомах, признаваемых вариантами некоторого языка [Henry 2005], социолингвистика и корпусная лингвистика. Варьирование в области словаря касается корреляций между корнем и значением. Варьирование в области грамматики касается ограничительных условий, регулирующих воспроизводство правильно построенных выражений. При параметризации грамматик языков мира можно с помощью одного и того же набора параметров описать как внутриязыковое, так и межъязыковое варьирование.

1.2 Конструкции

Описательная лингвистика называет конструкциями структурные схемы, ограничивающие отбор лексического материала[1]. Такое понимание совместимо с изучением внутриязыкового варьирования. Предположим, что конструкция C строится при помощи лексем из множества {a, **b**, c, **d**, e, f, g, h}, но в идиоме L_1 используются только {a, **b**, c, **d**}, в идиоме L_2 — только {**b**, **d**, e, f}, а в идиоме L_3 — только {**b**, **d**, g, h} Тогда элементы множества {b, d}, представленные в каждом идиоме, можно признать ядром C, а прочие элементы — его расширением в идиомах $L_{1\text{-}3}$. Гипотеза о том, что в разных идиомах реализуется одна и та же конструкция, держится на допущении о том, что у C есть инвариантное значение, которое сохраняется при лексических заменах типа a, c $\rightarrow$ e, f. Кроме того, гипотеза о том, что C поддерживается в языке L лексемами из множества {a, **b**, c, **d**, e, f, g, h}, опирается на допущение о возможности вывести усредненную форму языка L на основе пересечения или объединения грамматик и словаря идиомов L_1, L_2, L_3. Границы варьирования проверяются на основе распределения конструкций $C_1..C_n$ в разных идиомах L, или в корпусе текста, где можно установить частотность элементов словаря конструкций. При комбинировании корпусных и социолингвистических методов в благоприятном случае можно установить корреляцию между частотностью элемента или комбинации элементов в корпусе и степенью их одобрения носителями языка, см. описание двойного эксперимента в [Zimmerling 2017].

Инвариантное значение конструкции, если жестких запретов на пополнение ее словаря нет, проверяется двумя способами. Одним из них является уточнение набора денотативных ситуаций, покрываемых ее употреблением, т.е. построение *онтологии конструкции*. Другим способом является уточнение *таксономической семантики*, т.е. ниши, занимаемой данной конструкцией языка L в общей таксономии предикатных типов, приложимой к разным языкам.

2 Предикативы дативно-предикативной структуры в русском языке

2.1 Схема и семантика

Предикативы дативно-предикативной структуры (далее — *ДПС*) есть несогласуемые неглагольные слова, реализующимися в схеме дат. п. лица — связка — предикатив. Н.С.Поспелов указал три характеристики конструкции ДПС: 1) наличие у несогласуемого предикатива беспредложной валентности на одушевленный актант в дат. п. 2) связь между схемой ДПС и предикатным значением «состояния»; 3) отсутствие у предикатива ДПС канонического подлежащего в имен. п. [Pospelov 1955]. Мы принимаем анализ Поспелова и разграничиваем ДПС и дативно-номинативные структуры (ДНС). Частотность неадъективных предикативов ДПС типа *не к лицу* подсчитывается далее только для контекстов *X-у не к лицу делать Z,* но не для контекстов *X-у не к лицу Z*.[2]

Позиция актанта в дат. п., за вычетом случаев персонификации, замещается одушевленным существительным/местоимением, о примерах типа *пирогу надо остыть* см. [Zimmerling 2020a].

[1] В т.н. грамматике конструкций [Goldberg 2006] принимаются постулаты о том, что конструкцией является сложное выражение с идиоматичным значением, и о том, что выделение конструкций связано с парсингом «сверху вниз» (top down).

[2] Теоретические проблемы, связанные с гипотезой Поспелова, обсуждаются в [Zimmerling 2018ab], где обосновывается ее непротиворечивость.

В позиции связки, помимо нулевых и ненулевых форм *быть*, возможны связки *стать* и *делаться,* в некоторых случаях в архаических текстах используются также глаголы *приходиться* и *приходить*. Структуры с полузнаменательными связками могут анализироваться как биклаузальные, но при обработке примеров предложения типа *Мне стало холодно* допустимо считать вариантами предложений типа *Мне было холодно*. Меньшая часть говорящих, по данным корпусного исследования [Zimmerling 2018c], допускает предложения ДПС со связкой *казаться* и несогласуемым предикативом, ср. *?Мне казалось холодно*, *?Мне показалось удивительно*, но они не соответствуют употреблению большинства. В целом, отделение стандартных употреблений от маргинальных для первых двух позиций ДПС является технической проблемой.

Иначе обстоит дело с заполнением третьей позиции. Класс предикативов ДПС пополняем, и установить, есть ли у русского предикатива валентность на дат. п., без анализа варьирования нельзя. Авторы текстов порождают примеры **мне солнечно*, *пыльно*, но частотных употреблений с дат. п. лица у слов *солнечно* и *пыльно* нет, что подтверждается негативной реакцией большинства информантов на соответствующие примеры [Zimmerling 2017]. В то же время, у *глупо* есть регулярная валентность на дат. п. лица, но неясно, соответствуют ли предложения типа *Мне глупо отказываться от этого* значению «состояния», постулированному для конструкции ДПС. В таких предложениях дат. п. *мне*, на первый взгляд, указывает не на переживаемое Х-м состояние, а на оценку, ср. парафразу *Со стороны Х-а глупо отказываться*. Чтобы проверить гипотезу о том, что валентность на дат. п. лица диагностирует класс русских предикативов, нужно указать критерии их отбора.

2.2 Критерии отбора предикативов

В русском языке нет предикативов, которые реализуются исключительно в схеме *Х-у Z-во* и не допускают реализаций типа *Было очень Z-во*, *Z-во, что P*, *делать P — Z-во* и т. п. Импликация «если русский предикатив имеет реализацию с дат. п. лица, он также имеет реализацию без дат. п.», верна.

(i) *Х-у Z-во* → *Z-во*.

Действует ли в русском языке обратная импликация (ii), утверждающая, что для любого если предикатива может быть построен контекст с дат. п. лица, неясно.

(ii) *Z-во* → *Х-у Z-во*.

Проверка (ii) зависит от избранной модели. Базовых моделей две. Согласно первой, валентность на дат. п. задана на уровне словаря. Согласно второй, валентность на дат. п. не специфична для какого-либо класса слов и задается правилами грамматики. В [Švedova 1982: 151] ИГ в дат. п. лица трактуется как субъектный детерминант, т.е. универсальный расширитель разных схем предложения, не зависящий от конкретной лексической вершины. Другими примерами субъектных детерминантов признаются предложные группы типа *для Х-а, у Х-а*, и обороты типа *на душе (у Х-а)*. Субъектные детерминанты можно отождествить с семантической валентностью на одушевленного субъекта, которая необязательно соответствует стандартной синтаксической валентности.

Объем активного словаря конструкции ДПС у большинства информантов, по данным социолингвистического эксперимента [Zimmerling 2017], не превышает 245 единиц, при том, что тестовый словник включал 422 стимула. Это означает, что носители языка, каждый — в своем идиолекте, отбирают предикативы ДПС на основе их семантики. Если описание грамматики ориентировано на употребление большинства, гипотеза о том, что импликация (ii) ложна, и существуют предикативы, которые не допускают сочетаний с дат. п. лица, сохраняет силу. Остается задать порог, при котором число корпусных употреблений с дат. п. лица признается незначимым.

2.3 Ранжирование предикативов

Корпус, где проверяется конструкция ДПС, должен быть достаточно велик для ранжирования 300–500 единиц. Диагностика ДПС связана с отсевом омонимичных структур в контексте шире элементарного предложения. В [Zimmerling 2017] была предложена аппроксимативная мера. Число употреблений предикатива с дат. п. лица оценивалось для контекста предикатив + субъектное местоимение 1 л. ед. ч. (*мне*) в окне <-1; 1>. Коэффициент m указывает число диагностированных клауз ДПС в этом контексте. Для ранжирования словника по *м*-мере основного корпуса НКРЯ[3] достаточно. При обращении к основному корпусу НКРЯ *м*-мера дает положительные значения для 275 единиц из 422, высокочастотными можно считать 145 предикативов ДПС с m ≥ 10, все предикативы с m ≥ 10 имеют высокий рейтинг одобрения. Емкость словаря конструкции ДПС в среднестатистическом идиолекте русского языка по данным эксперимента составила 245 единиц. Этой цифре соответствуют предикативы с m ≥ 2, но надежнее положить в виде нижней границы m = 3 (199 предикативов ДПС). Наличие 3 и более клауз ДПС в выборке по *м*-мере при данных размерах корпуса подтверждает, что предикатив имеет регулярную валентность на дат. п. Обратное неверно: некоторые предикативы с m ≤ 2 имеют высокий рейтинг, что объясняется спецификой социолингвистических и корпусных методов. Информантам не составляет труда построить контекст, где уместны предикативы ДПС типа *X-у в ручье по шею, по щиколотку* и т. п., но такие денотативные ситуации, особенно в режиме актуального настоящего в 1л., встречаются в корпусе редко.

2.4 *М*-мера и автореферентность

Большинство русских предикативов ДПС автореферентны, т.е. ориентированы на употребление в 1 л. в режиме актуального настоящего времени: состояние X-а фиксируется самим X-м. В русском языке свойство автореферентности связано с падежом субъекта [Zimmerling 2020b]. Предикативы ДПС, требующие дат. п., как правило, автореферентны, отражением чего является статистическое преобладание предложений в 1 л. ед. ч. с субъектным местоимением *мне*. Таких предложений в корпусной выгрузке обычно больше, чем предложений, где субъект выражен личным местоимением дат. п. в 2-3 л. Напротив, предикативы, требующие подлежащего в имен. п., ср. *X навеселе, X без чувств*, *X не в духе* ориентированы на описание чужих состояний, отражением чего является обратная пропорция 1 и 3 л. в выгрузке.

Так как для большинства предикативов ДПС употребление в 1 л. является приоритетным, выборка по *м*-мере дает репрезентативную картину частоты предикативов ДПС. Не все примеры контактной позиции предикатива и его субъекта (*мне Z-во ~ Z-во мне*) связаны со значением актуального настоящего[4], но значительная часть выгрузки относится именно к этому случаю. Синтаксическое ожидание состоит в том, что при таком линейном порядке связка будет нулевой, и, соответственно, в предложении будет выражаться настоящее время. Так, у предикатива ДПС *безразлично* (m = 101) только 2 предложения (1, 98%) в выгрузке по *м*-мере имеют ненулевую связку. У предикатива ДПС *холодно* (m = 211) в выгрузке по *м*-мере ненулевая связка обнаруживается в 23 предложениях (1,09%).

Предикативов ДПС, ориентированных на 3 л. или 2 л., немного, ср., *жирно, слабó, неладно, пусто, так и надо, свойственно, присуще*. Если выгрузка по 2 л. и 3 л. дает большой прирост по сравнению с выгрузкой по *м*-мере, частотность предикатива ДПС в НКРЯ можно измерять, заменив *мне* на субъектное местоимение другого лица: так, у <*Чтоб X-у было*> *пусто*$_2$ (m = 1) замена субъектной формы *мне* на форму 3 л. ед. ч. *ему* увеличивает выгрузку в 46 раз (m=46). Для подавляющего большинства предикативов коррекция выгрузки по *м*-мере не требуется.

[3] 23 803 881 предложение на время проведения подсчетов, дата обращения 05.03.2017. В марте 2021 г. список предикативов ДПС был сверен.

[4] Контактный порядок *…мне + Z-во…* допускает наличие ненулевой связки прошедшего или будущего времени или показателя сослагательного наклонения в левом или правом контексте.

3 Онтология конструкции ДПС и значение внутреннего состояния

3.1 Словник ДПС по данным эксперимента

В [Zimmerling 2017] словарь конструкции ДПС был разбит на 15 тематических классов, в совокупности образующих *онтологию* конструкцию ДПС: 1) физические состояния (27 стимулов основного списка/10 стимулов дополнительного списка); 2) модальности (44/5); 3) эмоциональные состояния (57/24); 4) моральные оценки (16/0); 5) удобство исполнения (8/0); 6) уместность/неуместность (13/2), 7) внутренняя потребность (7/2); 8) (не) соответствие задаче (11/0); 9) трудность выполнения (10/4); 10) (не) желание выполнять (9/1); 11) общая оценка (41/7); 12) (не) релевантность (16/4); 13) (не) эффективность (6/1); 14) сенсорные и интеллектуальные реакции (25/12); 15) параметрический признак (52/8). Тематические группы предикатной лексики соответствуют типам денотативной ситуации. Большинство классов пополняемо, в каждом классе есть как высокочастотные и высокорейтинговые, т.е. одобренные значительным большинством информантов, предикативы, так и единицы, употребляемые лишь частью опрошенных. Классы включают не леммы, а предикатные элементы в связи с выражаемым ими значением. Так единицы <*X-у*> *плохо*$_1$ «X испытывает ощущение дурноты» (физическое состояние, класс 1) и <*X-у*> *плохо*$_2$ «X считает свое положение плохим» (общая оценка, класс 11) трактовались как омонимы и проверялись отдельно. Точно так же, омографы *чу́дно* «X-у очень хорошо» (класс 11), и *чудно́* «X-у странно» (класс 3) трактуются как разные элементы. Омонимия в пределах тематического класса не признается: если значения *неудобно*$_1$ «X испытывает физические неудобства при выполнении действия» (класс 5) и *неудобно*$_2$ «X испытывает моральные затруднения» (класс 4) разводятся, они должны быть отнесены к разным классам. При корпусном анализе учитывались только ДПС типа *X-у было* ***не по душе****, что Y купил книги*, а предложения типа *Книги были X-у* ***не по душе*** игнорировались.

Для 442 стимулов получено два показателя: коэффициент m и рейтинг одобрения информантами. Около 145 единиц, с m ≥ 10 по НКРЯ, можно считать словарным ядром конструкции ДПС. Так как средняя емкость словаря ДПС в идиолектах опрошенных составляет 245 единиц, очевидно, что носители языка достраивают словарь ДПС самостоятельно, выбирая разные единицы, имеющие среднюю и низкую частотность в НКРЯ. В [Ivanova, Zimmerling 2019] эта онтология проверялась на материале двух языков — русского и болгарского. Болгарская конструкция ДПС, словарь которой меньше, покрывает больше денотативных ситуаций, в силу чего для нее нужны дополнительные классы онтологии. В то же время, нет никаких данных, указывающих на то, что для описания русской конструкции ДПС, нужны иные типы ситуаций: если бы это было так, им соответствовали бы тематические классы, в состав которых входили бы частотные предикативы ДПС.

3.2 Внутренние и внешние состояния

Состояния принято определять как разновидность локализованных во времени и пространстве событий (spatiotemporal events), занимающих отрезок на временной оси и не связанных с изменением мира от начальной до конечной точки отрезка [Davidson 1980; Bulygina 1982; Seliverstova 1982]. Л. В. Щерба и его последователи трактуют несогласуемые предикативы *(X-у) весело, X навеселе, X не в духе*, и даже *X в сюртуке* как слова <категории> состояния [Ščerba 1928; Vinogradov 1947; Lekant 2015]. Проведенное Л. В. Щербой на материале первичных предикатов различение «состояний», рус. *мне весело*, и «качеств» (свойств), рус. *я веселый*, эквивалентно выявленному Г. Карлсоном контрасту между предикатами актуализованного признака (stage-level predicates, SLP), которые обозначают временные характеристики референтных ситуаций, ср. англ. There are firemen available «пожарные (в данный момент) доступны» и предикатами индивидного признака (individual-level predicates, ILP), которые обозначают свойства, отвлеченные от конкретных ситуаций, ср. невозможность англ. *There are firemen altruistic, подр. зн. «В данный момент пожарные альтруистичны» [Carlson 1977]. Контраст между SLP и ILP преимущественно иллюстрируется на материале вторичных предикатов [Kratzer 1995; Kosta 2014; 2020], что мешает осознанию эквивалентности дэвидсоновских/щербовских состояний и SLP-предикатов в традиции Г. Карлсона.

Для описания специфики предикативов ДПС требуется дополнительное различение внешних и внутренних состояний. Термин «внутренние состояния» был введен в [Zaliznjak 1995] для группы глагольных значений и использован для классификации именных предикатов в [Zimmerling 2018a]. Внутренние состояния обозначают ситуации с приоритетным одушевленным аргументом, в то время как у внешних состояний нет приоритетного аргумента, независимо от одушевленности и местности предиката (ср. *Здесь пыльно*, *X дружит с Y-м*). Релевантное отличие состоит в том, что внутренние состояния нельзя квантифицировать экстенсионально: в одном и том же локусе в одно и то же время X-у может быть холодно и скучно, а Y-у — нет. Кроме того, внутренние состояния, в отличие от внешних, не характеризуют визуально и аудиально воспринимаемые ситуации. Напротив, предикаты внешнего состояния типа *<Здесь> пыльно*; *<сегодня> пасмурно* обозначают ситуации, где любой наблюдатель подтвердит, что высказывание истинно.

3.3 Смена семантического типа

3.3.1. *От внешнего состояния к внутреннему*
Переход $S_{EXT} \rightarrow S_{INT}$ в истории русского языка подтверждается тем, что число основ, от которых могут быть образованы предикативы ДПС, за последние 500 лет увеличилось в несколько раз [Zimmerling 2018b]. Вместе с тем, в каждый момент времени в большинстве идиолектов сохраняются основы типа *пыльн-*, *пасмурн-*, от которых предикативы ДПС не образуются.

3.3.2. *От внутреннего состояния к внешнему*
Переход $S_{INT} \rightarrow S_{EXT}$ может реализоваться за счет устранения внешне выраженного субъекта. В русском языке нет конструкций с обязательным местоименным экспериенцером, ср. [Ivanova 2016]. Один и тот же предикатив может реализоваться и в ДПС в значении S_{INT}, и в схеме без семантического субъекта в значении S_{EXT}. Сдвиг $S_{INT} \rightarrow S_{EXT}$ при опущении внешне выраженного субъекта автоматически, но контексты, где значения «*X-у Z-во*» vs «*Z-во*» разделены, существуют. В примере (1) говорящий противопоставляет значения «смешно с точки зрения любого человека» (*∅ смешно*) и «смешно конкретному человеку» (*смешно мне*):

(1) *И какая все это беспорядица, неурядица, глупость и пошлость, и я всему причиною. А впрочем, иногда бывает ∅ смешно* (S_{EXT}) **— *мне*** (S_{INT}) *по крайней мере.* [Ф. М. Достоевский. Игрок (1866)].

4 Предикаты состояния в НКРЯ

Возможность образования вторичных внутренних состояний от внешних ($S_{EXT} \rightarrow S_{INT}$) проверяется путем выявления порога регулярности такого перехода в НКРЯ. В разделе **4.1.** приводятся данные для 50 предикатов внешних состояний. Проверка обратного перехода $S_{INT} \rightarrow S_{EXT}$ требует семантической аннотации. В п. **4.2.–4.4.** доказывается, что предикатив *тоскливо* и коррелятивное ему наречие могут передавать значение S_{EXT} при устранении субъекта из схемы предложения.

4.1 Пополнение словаря ДПС

На материале НКРЯ в 2021 г. проверялись ограничения на употребление 50 предикативов внешнего состояния. 5 единиц входило в дополнительный список стимулов, для которых в [Zimmerling 2017] получены низкие оценки приемлемости употреблений с дат. п. лица: Гипотеза состояла в том, что если переход $S_{EXT} \rightarrow S_{INT}$ для предикатива затруднен, число клауз ДПС в выборке по *м*-мере должно быть $0 < m < 3$, т.е. ниже порога, подтверждающего включение элемента в словарь ДПС. Прогноз оправдался для 49 предикативов из 50. Для *сухо* $m = 3$, при этом в контекстах *Здесь сухо* (S_{EXT}) и *Мне сухо* (S_{INT}) реализуются разные лексические значения: в двух примерах из трех предложение ДПС значит ‘мне нужно выпить’, ср. (2):

(2) *И не крикнет покойник, встав со смертного ложа: други,* ***сухо мне****!* [М. Шагинян. Перемена (1959)].

Контексты типа (2) были незнакомы части опрошенных, чем объясняется низкий рейтинг одобрения. Для сравнения, у предикатива *X-у мокро*, включенного в основной список эксперимента 2017 г. и одобренного 12 информантами из 18, m = 1, при 21 предложениях ДПС в НКРЯ. Тем самым, ни *X-у сухо,* ни *X-у мокр*о не относятся к словарному ядру конструкции. Чтобы оценить употребления ДПС для 50 проверявшихся предикативов, нужно сопоставить три показателя — общее число вхождений в корпус, число клауз ДПС и коэффициент m.

	Предикативы	Число вхождений в НКРЯ	Число клауз ДПС	m
m > 1	*сухо*	235	6 (2,55%)	3
m = 1	*пыльно, свежо, скользко, пустынно, морозно, тряско, бездомно*	1900	34 (1,79%)	1
m = 0	*солнечно, знойно, облачно, пасмурно, дождливо, болотисто, туманно, мглисто, ветрено, промозгло, влажно, сыро, хмуро, склизко, липко, безлюдно, снежно, бесснежно, безветрено, людно, лунно, звездно, беззвездно, сумрачно, грязно, росно, росисто, топко, рыхло, вязко, гористо, лесисто, каменисто, терпко, вонько, вонюче, зловонно, парко, тинисто, дымно, смрадно, чадно, вьюжно, тенисто, слякотно, угарно*	4572	35 (0,76%)	0

Таблица 1. Предикаты внешнего состояния в контексте S_{INT}

Анализ подтвердил, что переход $S_{EXT} \rightarrow S_{INT}$ в большинстве идиолектов русского языка затруднен. Построение вторичного предикатива ДПС типа ??*X-у пыльно* не запрещено, но эксперименты носителей языка с помещением предикатов внешнего состояния в контекст, характерный для S_{INT}, обычно не востребованы другими носителями и относятся к зоне индивидуального варьирования.

4.2 Переход $S_{INT} \rightarrow S_{EXT}$

Переход $S_{INT} \rightarrow S_{EXT}$ демонстрируется на примере предикатива *тоскливо*. Концепт *тоски* описан в [Wierzbicka 1992: 169-174; Šmelev 2002: 90-92; Zaliznjak 2006: 207; Zaliznjak, Levontina, Šmelev 2012: 41; Zaliznjak 2013: 41, 376].

4.2.1. *Тоскливо$_2$ как предикат внутреннего состояния*

НКРЯ, на март 2021 г., указывает 3133 вхождений *тоскливо* в качестве предикатива, согласуемого прилагательного ср. р. и наречия: их следует вводить как разные леммы. Для обработки была доступна выборка из 1632 клауз, самые ранние примеры относятся к 1909 г. В этой выборке m = 23, что позволяет отнести *тоскливо$_2$* к ядру ДПС. По [Zimmerling 2017], 83,3% информантов (15 из 18) активно используют *тоскливо$_2$* как S_{NT}. Распределение синтаксических структур указано в таб. 2. Обращает на себя внимание ничтожное (0,06%) число примеров с прилагательным *тоскливо$_1$* и нетривиальное соответствие между дат.п. лица и наличием оборота *на сердце/на душе (у X-а)*, указывающим на субъект состояния. В 307 клаузах ДПС такой оборот встретился 1 раз (0,32%), в то время как в клаузах без дат. п., доля примеров с ним составила 14,47%. Выгрузка наречия *тоскливо$_3$* разделена сообразно двум позициям — а) *тоскливо$_3$* в позиции модификатора прилагательного или другого наречия, ср. *тоскливо-щемящий*; б) *тоскливо$_3$* в позиции элемента глагольной группы (vP, VP) или включающей ее группы функциональной категории предложения.

I.Краткое прил. ср.р. *тоскливо*$_1$	II. Предикатив *тоскливо*$_2$				III. Наречие *тоскливо*$_3$	
	ДПС		без дат. п.			
	+ на душе	без расширителя	*+ на душе/на сердце*	без расширителя	ADV1 (ADJ/ADV$_2$)	vP, VP
1	1 (0,32%)	306	59 (14,47%)	343	38	885
	307		402		923	

Таблица 2. Прилагательное, предикатив и наречие *тоскливо* в НКРЯ

Предложения с *тоскливо*$_2$, где есть один из маркеров субъекта состояния — группа со значением дат.п. лица или *оборот на душе/на сердце* — либо оба, можно отождествить с S_{INT}, их доля составляет 50,63% от выгрузки *тоскливо*$_2$. Вместе с тем, отсутствие данных маркеров еще не свидетельствует, что *тоскливо*$_2$ передает значение S_{EXT}.

4.2.2. *Семантические и синтаксические актанты*

Оборот *на душе/на сердце* при *тоскливо*$_2$ указывает на одушевленного участника, но в отличие от беспредложного дат. п. лица, не выступает в роли приоритетного синтаксического актанта. Оборот *на душе/на сердце* находится в дополнительной дистрибуции с сентенциальной валентностью. Допускаются структуры типа *(Мне) было тоскливо*$_2$ *(на душе)* и типа *(мне) было тоскливо*$_2$ *выходить из дому*, но не структуры типа **было тоскливо*$_2$ *на душе выходить из дому*.

Предикативы ДПС (307)			Предикативы без дат.п. лица (402)		
+ на душе (1)	без расширителя (306)		*+ на душе/на сердце* (59)	без расширителя (343)	
	что Р	инфинитив		что Р	инфинитив
0	1 (0,32%)	13 (4,24%)	0	2 (0, 49%)	57 (16,61%)

Таблица 3. Сентенциальные валентности предикатива *тоскливо*$_2$, по НКРЯ

2.3. *Ориентация на 1-е лицо*

В 78,5% случаев позиция дат.п. лица при *тоскливо*$_2$ замещается личным местоимением (245 примеров из 307). При контактной позиции местоимения доля высказываний в 1 л. составляет 48% (24 из 50), при дистантной — 61,78% (118 из 191). Эти данные согласуются с гипотезой о том, что *тоскливо*$_2$, как и большинство русских предикативов ДПС — автореферентное слово, ориентированное на описание состояния говорящего.

4.3 *Тоскливо*$_2$ как внешнее состояние

Не менее 52 примеров выборки строится по схеме *Z-е тоскливо*$_2$ и выражает значение S_{EXT}. Некоторые примеры можно рассматривать как рефлексию над тем фактом, что *тоскливо*$_2$ параллельно употребляется как в значении S_{INT}, так и в значении S_{EXT}.

(3) *Тоскливо все и внутри Феликса, и вокруг него*. [А. Ростовский. По законам волчьей стаи (2000)].

(4) *Выло в трубе, и становилось тоскливо на душе*. [К. С. Станиславский. Работа актера над собой (1938)].

В (3)–(4) неясно, оценивается ли референтная ситуация с точки зрения конкретного вовлеченного в нее человека — Феликса в (3), рассказчика в (4), или же любого наблюдателя. Если одушевленного участника нет, вторая интерпретация является единственной.

4.4 Наречие *тоскливо*$_3$ и структура события

Употребление *тоскливо*$_2$ служит базовым контекстом для толкования наречия *тоскливо*$_3$. Ситуации, описываемые *тоскливо*$_3$ и *тоскливо*$_2$, подразумевают два типа одушевленных участников. Одним из них является субъект состояния (X), который может быть не выражен синтаксически, но реконструироваться однозначно, например, при имени аффекта *воспоминание* в (5):

(5) ***Яркое воспоминание*** (→ X) ***тоскливо*** *стеснило грудь* (→ X), *но* Гирин (X) *отбросил его*. [И. А. Ефремов. Лезвие бритвы (1959-1963)].

Вторым типом одушевленного участника, обычно остающимся имплицитным, является наблюдатель, он же субъект оценки (Y). Пример (6) допускает только одну интерпретацию, когда оценка присваивается наблюдателем, в роли которого может выступить любой человек:

(6) ***Негромкий звук выстрела*** *(→ Y) одиноко и* ***тоскливо*** *прокатился эхом по лесу и иссяк где-то вдали.* [В. Кондратьев. Сашка (1979)].

При наличии одушевленного участника предложение становится амбивалентным. Так, пример (7) описывает ситуацию, где на улице мерзнут некие милиционеры (X). Эта ситуация доступна внешнему наблюдению и может отражать точку зрения стороннего наблюдателя, стоящего на противоположной стороне улицы (Y) = «Было *тоскливо*$_2$ (S_{EXT}) смотреть, как на улице мерзли милиционеры». Но она может также отражать перспективу самого X-а, в этом случае толкование будет содержать отсылку к внутреннему состоянию — «На улице мерзли милиционеры, и им было *тоскливо*$_2$ (S_{INT})».

(7) *На противоположной стороне улицы (→ Y)* ***тоскливо*** *мерзли три милиционера (→ X)*. [Е. Строителева. «Иисус, как и Ленин, добра людям хотел» (2002) // «Известия», 2002.11.08].

5 Заключение

На материале НКРЯ была проверена модель варьирования грамматической конструкции, опирающуюся на онтологию данной конструкции. Ограничение поиска выделенным контекстом — контактной позицией предикатива и субъектного местоимения 1 л. ед.ч. — позволяет установить порог, при котором употребления предикатива с дат.п. лица могут считаться значимыми. Регулярные семантические переходы работают только в одну сторону — от внутренних состояний к внешним, но не в обратном направлении. При устранении дат.п. лица предикатив может описывать внешне наблюдаемую ситуацию, оцениваемую одинаково любым наблюдателем. Глагольные предложения с качественным наречием, коррелятивным предикативу, проецируют структуру события, содержащую отсылку либо к внутреннему, либо к внешнему состоянию.

References

[1] Bulygina Tatiana V. (1982). Towards Predicate Typology in Russian [K Postroeniyu Tipologii Predikatov v Russkom Jazyke], O. N. Seliverstova (ed.), Semantic types of the predicates [Semanticheskie Tipy Predikatov], Moscow, Nauka, 1982, pp. 7–85.

[2] Carlson Greg. Reference to Kinds in English. PhD dissertation. — MIT, 1977.

[3] Davidson Donald. The Individuation of Events // D. Davidson (ed.) Essays on Actions and Events. — Oxford, Clarendon Press, 1980. — P. 163–180.

[4] Goldberg Adele. Constructions at Work: the Nature of Generalization in Language. — Oxford, OUP, 2006.

[5] Henry Alison. Non-standard Dialects and Linguistic Data // Lingua, vol. 115, 2005. — P. 1599–1617.

[6] Ivanova Elena Yu. (2016). Impersonal Sentences with an Obligatory Pronominal Experiencer in Bulgarian [Bezlichnye Predlozheniya s Obyazatel'nym Mestoimennym Vyrazheniem Eksperiencera v Bolgarskom Jazyke], A.V.Zimmerling, E.A.Lyutikova (eds.), Clause Architecture in the Parametric Models. Syntax, Information Structure, Word Order, [Arxitektura Klauzy v Parametricheskix Model'ax: Sintaksis, Informacionnaya Struktura, Por'adok Slov], Moscow, LRC, pp. 332–68.

[7] Ivanova Elena Yu., Zimmerling Anton V. Shared by All Speakers? Dative Predicatives in Bulgarian and Russian // Bulgarski Language and Literature, № 4, 2019. — P. 353–363.
[8] Kosta Peter. Adjectival and Adjectival and Argumental Small Clauses vs. Free Adverbial Adjuncts. A Phase-Based Approach within the Radical Minimalism with Special Criticism of the Agree, Case and Valuation Notions. — 2014. — Access mode: https://www.semanticscholar.org/paper/Adjectival-and-argumental-Small-Clauses-vs.-free-%E2%80%93-Kosta/1c5b2340f433d01c17eb689ae579cfab8313b61f
[9] Kosta Peter. The Syntax of Meaning and the Meaning of Syntax: Minimal Computations and Maximal Derivations in a Label-/Phase-Driven Generative Grammar of Radical Minimalism. Potsdam Linguistic Investigation 31. — Berlin: Peter Lang, 2020.
[10] Kratzer Angelica. Stage Level and Individual Level Predicates // Carlson G. & Pelletier F.J. (eds.): The Generic Book. — Chicago: The University of Chicago Press, 1995. — P. 125–175.
[11] Lekant Pavel A. (2015). Analytical Forms and Analytical Constructions in Modern Russian [Analiticheskie Formy i Analiticheskie Konstrukcii v Sovremennom Russkom Jazyke], Moscow, Inform.
[12] Pospelov Nikolai S. (1955). In Defence of the State Category [V Zasch'itu kategorii Sostojaniya], Issues in Linguistics [Voprosy Jazykoznaniya], № 2, pp. 55–65.
[13] Ščerba Lev V. (1928). On Parts of Speech in Russian [O Chastyax Rechi v Russkom jazyke], Russian Speech. New Series, II [Russkaja Rech'. Novaja Seriya], — Leningrad, Academia, — pp. 5 —27.
[14] Šmelev Alexei. D. (2002). Russian Natural Language Metaphysics. Preliminaries for a Dictionary [Russkaja Jazykovaya Model' Mira. Materialy k Slovar'yu], Moscow, LRC.
[15] Seliverstova Olga N. (2002). An Alternative Variant of the Predicate Taxonomy and Some Predicate Types in Russian [Vtoroj Variant Klassifikacionnoj Setki i Opisanie Nekotoryx Predikatnyx Tipov Russkogo Jazyka], O.N.Seliverstova (ed.), Semantic Types of the Predicates [Semanticheskie Tipy Predikatov], Moscow, Nauka, pp. 86–157.
[16] Švedova Nina Yu. (1982). Russian Grammar. In 2 vols. Vol 2 [Russkaja Grammatika, N.Yu.Shvedova (ed.)], Moscow, Nauka.
[17] Vinogradov Victor V. (1947). Russian Language. The Grammatical Theory of a Word [Russkij Jazyk. Grammaticheskoe Uchenie o Slove], Moscow- Leningrad, Vysšaya škola.
[18] Wierzbicka Anna. Semantics, Culture and Cognition: Universal Human Concepts in Culture-specific Configuration. — N.Y., 1992.
[19] Zaliznjak Anna A. (1995). Studies in the Semantics of the Predicates of the Inner States [Issledovaniya po Semantike Predikatov Vnutrennego Sostoyaniya], München, Otto Sagner.
[20] Zaliznjak Anna A. (2006). Polysemy in language and its modeling [Mnogoznachnost' v Jazyke i Sposoby Eje Predstavleniya], Moscow, LRC.
[21] Zaliznjak Anna A. (2013). Russian Semantics from a Typological Perspective [Russkaja Semantika v Tipologicheskoj Perspektive], Moscow, LRC.
[22] Zalizjak Anna A., Levontina Irina B., Šmelev Alexei D. (2012). Constants and Variables in the Russian Natural Language Metaphysics [Konstanty i Peremennye Russkoj Jazykovoj Kartiny Mira], Moscow, LRC.
[23] Zimmerling Anton V. (2017). Russian Predicatives from the Perspective of an Experiment and Corpus Grammar [Russkie Predikativy v Zerkale Eksperimenta i Korpusnoj Grammatiki], Computational Linguistics and Intellectural Technologies, Issue 16 (23) [Komp'yuternaja Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialogue 2017"], Moscow, pp. 466 — 482.
[24] Zimmerling Anton V. (2018a), Predicatives and Predicates of State in Russian [Predikativy i Predikaty Sostojanija v Russkom Jazyke], Slavic Review [Slavistična Revija], № 1, pp. 45–64.
[25] Zimmerling Anton V. (2018b), Impersonal Constructions and Dative-predicative Structures in Russian [Impersonal'nye Konstrukcii i Dativno-predikativnye struktury v Russkom Jazyke] // Issues in Linguistics [Voprosy Jazykoznaniya], № 5, pp. 7–33.
[26] Zimmerling Anton V. (2018c). Two dialects of Russian. Corpus grammar and a formal model [Dva Dialekta Russkoy Grammatiki: Korpusnye Dannye i Model'], Computational Linguistics and Intellectural Technologies, Issue 17 (24) [Komp'yuternaja Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialogue 2018"], Moscow, pp. 818 — 833.
[27] Zimmerling Anton V. (2020a). Animacy. Russian Language [Odushevlennost'. Russkij Jazyk], Proceedings of the Vinogradov Russian Language Institute [Trudy Instituta Russkogo Jazyka Imeni V.V.Vinogradova RAN], issue 24, pp. 43–56.
[28] Zimmerling Anton V. (2020b). Autoreferentiality and Predicate Classes [Aftoreferentnost' i Klassy Predikativnyx Slov], M.D.Voejkova, V.V.Kazakovskaya (eds.), Issues in Functional Hrammar. Reference to the Speaker in Grammatical Seemantics, [Problemy Funkcional'noy Grammatiki: Otnoshenie k Govor'jaschemu v Semantike Grammaticheskix Katerogorij, Moscow, LRC, pp. 23–58.

Литература

[1] Carlson Greg. Reference to Kinds in English. PhD dissertation. — MIT, 1977.
[2] Davidson Donald. The individuation of events // D. Davidson (ed), Essays on Actions and Events. — Oxford, Clarendon Press, 1980. — P. 163–80.
[3] Goldberg Adele. Constructions at Work: the Nature of Generalization in Language. — Oxford, OUP, 2006.
[4] Henry Alison. Non-standard Dialects and Linguistic Data // Lingua, vol. 115, 2005. — P 1599–1617.
[5] Ivanova Elena Yu., Zimmerling Anton V. Shared by All Speakers? Dative Predicatives in Bulgarian and Russian // Bulgarski Language and Literature, № 4, 2019. — P. 353–363.
[6] Kosta Peter. Adjectival and Adjectival and argumental Small Clauses vs. free adverbial Adjuncts. A phase-based approach within the Radical Minimalism with special criticism of the Agree, Case and Valuation notions. — 2014. Access mode: https://www.semanticscholar.org/paper/Adjectival-and-argumental-Small-Clauses-vs.-free-%E2%80%93-Kosta/1c5b2340f433d01c17eb689ae579cfab8313b61f
[7] Kosta Peter. The Syntax of Meaning and the Meaning of Syntax: Minimal Computations and Maximal Derivations in a Label-/Phase-Driven Generative Grammar of Radical Minimalism. Potsdam Linguistic Investigation 31. — Berlin: Peter Lang, 2020.
[8] Kratzer Angelica. Stage Level and Individual Level Predicates // Carlson G. & Pelletier F.J. (eds.): The Generic Book. — Chicago: The University of Chicago Press, 1995. — P. 125–175.
[9] Wierzbicka Anna. Semantics, Culture and Cognition: Universal Human Concepts in Culture-specific Configuration. — N.Y., 1992.
[10] Булыгина Т.В. К построению типологии предикатов в русском языке // О.Н.Селиверстова (ред.). Семантические типы предикатов. — Москва: Наука, 1982. — С. 7–85.
[11] Виноградов В.В. Русский язык. Грамматическое учение о слове. — Москва - Ленинград: Высшая школа, 1947.
[12] Иванова Е.Ю. Безличные предложения с обязательным местоименным выражением экспериенцера в болгарском языке // А.В.Циммерлинг, Е.А.Лютикова (ред.), Архитектура клаузы в параметрических моделях: синтаксис, информационная структура, порядок слов. — Москва: Языки славянской культуры, 2016. — С. 332–368.
[13] Зализняк Анна А. Исследования по семантике предикатов внутреннего состояния. — München, Otto Sagner, 1995.
[14] Зализняк Анна А. Многозначность в языке и способы ее представления. Москва: Языки славянской культуры, 2006.
[15] Зализняк Анна А. Русская семантике в типологической перспектива. — Москва: Языки славянской культуры, 2013.
[16] Зализняк Анна А., Левонтина И.Б., Шмелев А.Д. Константы и переменные русской языковой картины мира. — Москва: Языки славянской культуры, 2012.
[17] Лекант П.А. Аналитические формы и конструкции в современном русском языке. —Москва: Информ, 2015.
[18] Поспелов Н.С. В защиту категории состояния // Вопросы языкознания, 1955, № 2. — С. 55–65.
[19] Селиверстова О.Н. Второй вариант классификационной сетки и описание некоторых предикатных типов русского языка // О.Н.Селиверстова (ред.), Семантические типы предикатов. — Москва: Наука, 1982. — С. 86–157.
[20] Шведова Н.Ю. (ред.). Русская грамматика. В 2 т. — Т. 1. — Москва: Наука, 1982.
[21] Шмелев А.Д. Русская языковая модель мира. Материалы к словарю. — Москва: Языки славянской культуры, 2002.
[22] Щерба Л.В. О частях речи в русском языке // Русская речь. Новая серия. — Ленинград: Academia, 1928. — С. 5–27.
[23] Циммерлинг А.В. Русские предикативы в зеркале эксперимента и корпусной грамматике // Компьютерная лингвистика и интеллектуальные технологии. — Вып. 16 (23). Труды международной конференции «Диалог 2017». — Москва, 2017. — С. 466–482.
[24] Циммерлинг А.В. (2018a). Предикативы и предикаты состояния в русском языке // Slavistična Revija, № 1, 2018. — P. 45 — 64.
[25] Циммерлинг А.В. (2018b), Имперсонвальные конструкции и дативно-предикативные структуры в русском языке // Вопросы языкознания, № 5, 2018. — С. 7–33.
[26] Циммерлинг А.В. (2018c). Два диалекта русской грамматики: корпусные данные и модель // Компьютерная лингвистика и интеллектуальные технологии. — Вып. 17 (24). Труды международной конференции «Диалог 2018». — Москва, 2018. — С. 818–833.
[27] Циммерлинг А.В. (2020a). Одушевленность. Русский язык // Труды Института русского языка имени В.В.Виноградова РАН. — Вып. 24, 2020. — С. 43–56.
[28] Циммерлинг А.В. (2020b). Автореферентность и классы предикативных слов // М.Д.Воейкова, В.В.Казаковская (ред.). Проблемы функциональной грамматики. Отношение к говорящему в семантике грамматических категорий. — Москва: Языки славянской культуры, 2020b. — Москва: Языки славянской культуры, 2020. — С. 23–58.

Abstracts

MATCHING SEMANTIC SKETCHES TO PREDICATES IN CONTEXT USING THE BERT MODEL

Aleksandrova P., **Mokhova A.**, **Nikolaenkova M.**, NRU HSE, Moscow, Russia

Modern language models have extensive information about the compatibility and meanings of various words. One of the ways to represent such lexical information, which is presented in the present study, is the construction of semantic sketches.

This paper presents a solution to the task of predicting a predicate from its most frequent actants and sirconstants using the application of the BERT neural network, which showed the best quality metrics in the Dialogue Evaluation SemSketches competition. The study analyzed several solutions approaching this task and ways to improve them based on the peculiarities of the architecture and the nature of data in terms of linguistics.

The results of testing the selected methods showed that the most successful tool for determining the semantic sketch of a predicate is the Conversational RuBERT model combined with the search for synonyms of the verbs sought in the training data.

Other promising ways to improve the quality of mapping the predicate to its semantic sketch include the use of contextualized embeddings to be able to take context into account, as well as fine-tuning of the models used.

ANNOTATED SPAN NORMALIZATION AS A SEQUENCE LABELLING TASK

Anastasyev D. G., Yandex, Moscow, Russia

In this paper, we describe a way to perform span normalization as a sequence labelling task. Our model predicts the modifications that should be applied to the span tokens to normalize them. This prediction is performed via sequence labelling, which means that each token is normalized independently. Despite the simplicity of the approach, we show that it can lead to the stateoftheart results. We compare different pretraining schemas in application to this task. We show that the best quality can be achieved when the normalizer is trained on top of a BERTbased morphosyntactic parser's representations. Moreover, we propose some additional features useful in the task and prove that auxiliary morphosyntactic losses can help the model. Furthermore, we show that the model compares favourably with other contestant models of the RuNormAS competition.

DEEPMISTAKE: WHICH SENSES ARE HARD TO DISTINGUISH FOR A WORDINCONTEXT MODEL

Arefyev N.†‡§, **Fedoseev M.**†, **Protasov V.**ʃ, **Homskiy D.**†, **Davletov A.**†◊, **Panchenko A.**ʃ, †Lomonosov Moscow State University; ‡Samsung Research Center Russia; §HSE University; ʃSkolkovo Institute of Science and Technology; ◊RANEPA, Moscow, Russia

In this paper, we describe our solution of the Lexical Semantic Change Detection (LSCD) problem. It is based on a WordinContext (WiC) model detecting whether two occurrences of a particular word carry the same meaning. We propose and compare several WiC architectures and training schemes, and also different ways to convert WiC predictions into final word scores estimating the degree of semantic change.

We participated in the RuShiftEval LSCD competition for the Russian language, where our model achieved 2nd best result during the competition. During postevaluation experiments we improved the WiC model and managed to outperform the best system. An important part of this paper is detailed error analysis where we study the discrepancies between WiC predictions and human annotations and their effect on the LSCD results.

AN INTERPRETABLE APPROACH TO LEXICAL SEMANTIC CHANGE DETECTION WITH LEXICAL SUBSTITUTION

Arefyev N. V., Lomonosov Moscow State University, Samsung R&D Institute Russia, HSE University, Moscow, Russia;
Bykov D. A., Lomonosov Moscow State University, Moscow, Russia

In this paper we propose a new Word Sense Induction (WSI) method and apply it to construct a solution for the RuShiftEval shared task on Lexical Semantic Change Detection (LSCD) for the Russian language. Our WSI algorithm based on lexical substitution achieves stateoftheart performance for the Russian language on the RUSSE2018 dataset. However, our LSCD system based on it has shown poor performance in the shared task. We have studied mathematical properties of the COMPARE score employed in the task for measuring the degree of semantic change, as well as the discrepancies between this score and our WSI predictions. We have found that our method can detect those aspects of semantic change, which the COMPARE metric is not sensitive to, such as appearance or disappearance of a rare word sense. An important property of our method is its interpretability, which we exploit to perform the detailed error analysis.

NEAR-DUPLICATE HANDWRITTEN DOCUMENT DETECTION WITHOUT TEXT RECOGNITION

Bakhteev O.†‡, **Kuznetsova R.**§, **Khazov A.**†, **Ogaltsov A.**†, **Safin K.**§, **Gorlenko T.**†, **Suvorova M.**†, **Ivahnenko A.**†, Botov.†, **Chekhovich Y.**†‡, **Mottl V.**‡,
† Antiplagiat, Moscow, Russia; ‡ Dorodnicyn CC FRS CSC RAS, Moscow, Russia; § MIPT, Moscow, Russia

The paper presents a novel method for near-duplicate detection in handwritten document collections of school essays. A large amount of online resources with available academic essays currently makes it possible to cheat and reuse them during high school final exams. Despite the importance of the problem, at the moment there is no automatic method for near-duplicate detection for handwritten documents, such as school essays. The school essay is represented as a sequence of scanned images of handwritten essay text. Despite advances in recognition of handwritten printed text, the use of these methods for the current task is a challenge. The proposed method of near-duplicate detection does not require detailed markup text, which makes it possible to use it in a large number of tasks related to the information extraction in zero-shot regime, i.e. without any specific resources written in the processed language. The paper presents a method based on series analysis. The image is segmented into words. The text is characterized by a sequence of features, which are invariant to the author's writing style: normalized lengths of the segmented words. These features can be used for both handwritten and machine-readable texts. The computational experiment is conducted on IAM dataset of English handwritten texts and the dataset of real images of handwritten school essays.

IDIOMATICITY OF A TEXT AS A MATTER OF THE INDIVIDUAL STYLE: A QUANTITATIVE APPROACH

Baranov A. N., Russian Language Institute of RAS; **Dobrovol'skij D. O.**, Russian Language Institute of RAS, Institute of Linguistics, Moscow, Russia; Stockholm University, Stockholm, Sweden

The paper suggests one of the ways to formally define the degree of idiomaticity of a given text. Text idiomaticity is understood as the density of the use of idioms per text unit. The assessment of the degree of idiomaticity is carried out in the proposed approach as the ratio of the total number of idioms to the volume of the text in which they met. The conducted corpus experiment allows us to conclude that the degree of idiomaticity of the most important repre-sentatives of the prose of the second half of the 19th century varies significantly. Thus, the degree of idiomaticity of the text turns out to be an essential factor of the individual style.

THE ORDER OF OBJECTS IN RUSSIAN: A CORPUS STUDY

Bazhukov M. O., **Chubarova L. I.**, **Slioussar N. A.**, **Toldova S. Yu.**, NRU HSE, Moscow, Russia

The paper presents the results of a corpus study of the order of direct and indirect objects in ditransitive constructions in Russian (like Petya dal Mashe yabloko 'Petya gave Masha an apple' or Petya dal yabloko Mashe 'Petya gave an apple to Masha'). This topic has been widely discussed in the literature, but previous hypotheses have been based on individual examples and have never been tested on corpus data. Based on earlier research, we have selected parameters that affect the order of the objects, such as the length, depth, animacy and role of individual verbs and statistically tested their real effect on two subsamples: with a dative indirect object and with a prepositional one.

CORPUS REGIONAL LEXICOGRAPHY: PRINCIPLES, METHODS, AND PRELIMINARY RESULTS

Belikov V. I., MIPT, ABBYY Lab, Moscow, Russia; **Dubyaga A. O.**, RSUH, Moscow, Russia; **Rvanova L. Y.**, MIPT, ABBYY Lab, Moscow, Russia; **Selegey V. P.**, ABBYY, Moscow, Russia

The article summarizes the results of the long-term project "Languages of Russian Cities" (LoRC) of the regional vocabulary collecting and researching, which, unfortunately, was not depicted in any academic publications for a number of reasons. About 4 thousand pieces of regional materials were collected, systematized, and became the basis of the typology of regional differences consideration and the concept of a regional norm discussion. Reliability issues and methods of computer-based regional corpus research, including automatic text classification and author profiling, are paid attention to. Along with this article, the "reincarnation" of the LoRC project is also returning to the fund of open lexicographic resources basing on the joint portal for distinctive sociolinguistic research, which includes the General Web-corpus of Russian Language and the interactive dictionary "Languages of Cities and People" (LoC&P).

INFLUENCE OF SPEECH BREATHING AFTER PHYSICAL ACTIVITY ON INTONATIONAL-PAUSAL SEGMENTATION OF SPEECH

Belkova L., Lomonosov Moscow State University

This study raises the problem of the difference between normal and forced (deep) speech breathing. The aim of this work was to study the intonational-pausal segmentation of speech in normal and forced breathing after physical activity. The results of the study show that in the process of reading, the structure of the text determines the organiza-tion of breathing, and the breathing rate and respiration depth have an impact on the intonational-pausal segmentation of speech, as well as on the duration and quantity of intonation pauses.

EXAMINING THE ROLE OF LINGUISTIC CONTEXT IN ASPECTUAL COMPETITION: A STATISTICAL STUDY

Bernasconi B., Catholic University of the Sacred Heart Milan, Italy; **Noseda V.**, Roma Tre University, Sapienza University, Rome, Italy

This paper aims to show the results of a quantitative study on verbal aspect in modern Russian. Adopting a corpus-based approach, we investigate the phenomenon known as 'aspectual competition', which can take place when the imperfective aspect (ipf) is used instead of perfective to designate a single and complete event in the past. In particular, we investigate the interaction between the choice of aspect and co-textual factors in overlapping situations. In this study the attention is focused on one aspectual pair, namely pokupat'ipf—kupit'pf, 'to buy'. The work consists of two parts: in Phase 1 data were collected from the spoken subcorpus of the Russian National Corpus and the webcorpus RuTenTen11, annotated for several morpho-syntactic factors, and then examined. In Phase 2 a questionnaire was submitted to native speakers in order to collect more empirical evidence on aspect choice and verify the results obtained from the corpus study. In both phases, statistical methods were used to analyse the data. Results show that the aspect of the target verb mainly interacts with two factors: the presence of a contiguous verbs in the linguistic context and the presence of an object modifier.

PRAGMATIC MARKERS OF RUSSIAN EVERYDAY SPEECH: QUANTITATIVE DATA

Bogdanova-**Beglarian N. V.**†, **Blinova O. V.**†‡, **Sherstinova T. Ju.**‡†, **Troshchenkova E. V.**†, **Gorbunova D. A.**†, **Zajdes K. D.**†, **Popova T. I.**†, **Sulimova T. S.**†,
†Saint Petersburg State University Saint Petersburg, Russia; ‡HSE University, Saint Petersburg Russia

The article summarizes the results of a large research project dedicated to investigation of pragmatic markers (PM) in Russian everyday speech. Pragmatic markers are essential in spontaneous spoken discourse; thus, the quan-titative data on their usage are necessary for solving both theoretical and practical issues related to the study of spo-ken communication. New results were obtained on the data of two speech corpora: "One Day of Speech" (ORD; mostly dialogues; the annotated subcorpus contains 321,504 tokens) and "Balanced Annotated Text Library" (SAT; monologues; the annotated subcorpus includes 50,128 tokens). Statistical data were calculated for PM in dialogic and monologic speech, pragmatic markers common in both types of speech (e. g., hesitative markers like vot, tam, tak) are identified, as well as PM that are the most typical for monologues (e. g., boundary markers like znachit, nu, vot, vs'o) or dialogue (e. g., 'xeno'-markers such as takoi, grit and metacommunicative markers like vidish', (ja) ne znaju). Special attention is given to the pragmatic markers usage in different communicative situations.

SEMANTIC REPRESENTATIONS IN COMPUTATIONAL AND THEORETICAL LINGUISTICS: THE POTENTIAL FOR MUTUAL ENRICHMENT

Boguslavsky I. M.†, **Dikonov V. G.**, **Inshakova E. S.**, **Iomdin L. L.**, **Lazursky A. V.**, **Rygaev I. P.**, **Timoshenko S. P.**, **Frolova T. I.**, A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia; †Universidad Politécnica de Madrid, Madrid, Spain

Research in semantics is actively conducted both in theoretical and computational linguistics, but the formulation of tasks, objectives and results of semantic research in the two communities are usually largely different. As a step towards reducing this gap and increasing the awareness of theoretical linguists about what computational linguists are doing, we examine meaning representation approaches in computational linguistics and contrast them with how this is done within one of the best-known theoretical approaches — the Meaning ⇔ Text Theory.

SEMANTIC FEATURES AND VALENCY PROPERTIES OF THE RUSSIAN VERB 'PODOŽDAT' *WAIT*

Boguslavsky I. M., A. A.Kharkevich Institute for Information Transmission Problems, Moscow, Russia; Universidad Politécnica de Madrid, Spain; **Iomdin L. L.**, A. A.Kharkevich Institute for Information Transmission Problems, Moscow, Russia

The paper presents a detailed account of the semantics of the Russian perfective verb 'подождать' (≈ 'wait some time'), which belongs to the family of words focused around the verb 'ждать' 'wait'. The verb, much like the whole family, has a set of unique and non-trivial semantic properties that have not been so far adequately represented either in traditional and computer dictionaries of the Russian language or in scientific descriptions. The main features of this verb include its peculiar morphological and semantic relationship with the dominant word of the family, the verb ждать, as well as a ramified valence frame, characterized by rarely occurred means of implementing semantic valencies and unusual conditions of cooccurrence.

BUILDING DATASET AND MORPHEME SEGMENTATION MODEL FOR RUSSIAN WORD FORMS

Bolshakova E. I., Lomonosov Moscow State University, HSE, Moscow, Russia; **Sapin A. S.**, Lomonosov Moscow State University, Moscow, Russia

The paper describes a way to generate a dataset of Russian word forms, which is needed to build an appropriate neural model for morpheme segmentation of word forms. The developed generation procedure produces word forms segmented into morphs that are classified by morpheme types, based on existing dataset of segmented lemmas and additional dictionary data, as well as fine-grained classification of Russian inflectional paradigms, which makes it possible to correctly process word forms with alternating consonants and fluent vowels in endings. The built repre-sentative dataset (more than 1,6 million word forms) was used to develop a neural model for morpheme segmenta-tion of word forms with classification of segmented morphs. The experiments have shown that in detecting morphs boundaries the model has comparable quality with the best segmentation models for lemmas (98% of F-measure), slightly outperforming them in word-level classification accuracy (with score 91%).

ON (NON-)COMPATIBILITY OF GENITIVE PARTITIVE AND IMPERFECTIVE IN RUSSIAN: A CORPUS STUDY

Chuikova O. Iu., Herzen State Pedagogical University of Russia

The paper provides the results of the study of the use of the genitive case with partitive semantics as the means of direct object marking within imperfective verbs in Russian. The genitive partitive is traditionally claimed to be compatible with perfective verbs and as an exception with imperfective verbs used as the substitution for perfective verbs in neutralization contexts. The analysis of the data from the Russian National Corpus and the Russian-language Internet shows that the use of the genitive partitive within imperfective verbs is neither rare nor marginal. The compatibility level of the genitive and imperfective aspectual correlates of prefixed perfective verbs is dependent on the imperfectivability level and frequency. The use of the genitive partitive is sensitive to the semantics of the imperfective, however, it means the coverage of a broader range of phenomena than it is traditionally assumed. Although the use of the genitive partitive is mostly restricted to neutralization contexts such as iterativity and historical present, a number of gradual achievement imperfective verbs with progressive semantics as well as verbs that refer to constant situations are compatible with the genitive partitive.

METHODS FOR DETOXIFICATION OF TEXTS FOR THE RUSSIAN LANGUAGE

Dementieva D.‡, **Moskovskiy D.**‡, **Logacheva V.**‡, **Dale D.**‡, **Kozlova O.**†, **Semenov N.**†, **Panchenko A.**‡
‡Skolkovo Institute of Science and Technology, Moscow, Russia; †Mobile TeleSystems (MTS), Moscow, Russia

We introduce the first study of automatic detoxification of Russian texts to combat offensive language. Such a kind of textual style transfer can be used, for instance, for processing toxic content in social media. While much work has been done for thae English language in this field, it has never been solved for the Russian language yet. We test two types of models — unsupervised approach based on BERT architecture that performs local corrections and supervised approach based on pretrained language GPT-2 model — and compare them with several baselines. In addition, we describe evaluation setup providing training datasets and metrics for automatic evaluation. The results show that the tested approaches can be successfully used for detoxification, although there is room for improvement.

A QUANTITATIVE STUDY OF SIMPLIFICATION STRATEGIES IN ADAPTED TEXTS FOR L2 LEARNERS OF RUSSIAN

Dmitrieva A., University of Helsinki, Finland; HSE University, Moscow, Russia; Pushkin State Russian Language Institute, Moscow, Russia; **Laposhina A.**, **Lebedeva M.**, Pushkin State Russian Language Institute, Moscow, Russia

Nowadays there has been a growing interest in the topic of Russian text adaptation, both in theoretical aspects of intralingual translation into Simple and Plain Russian, and in practical tasks like automatic text simplification. Therefore, it is important to study the characteristics that make an adapted text more accessible. In this paper, we aim to investigate the strategies that human experts employ when simplifying texts, particularly when the texts are being adapted for learners of Russian as a foreign language. The main data source for this research is the RuAdapt parallel corpus, which consists of Russian literature texts adapted for the learners of RaaFL and the original versions of these texts. We study the changes that occur during the adaptation process on lexical, morphological, and syntax level, and compare them to the methods usually described in methodological recommendations for teaching RaaFL.

USING RUGPT3-XL MODEL FOR RUNORMAS COMPETITION

Emelyanov A.†‡, **Shliazhko O.**†, **Katricheva N.**†, **Shavrina T.**†§ʃ
†SberDevices, Sberbank, Moscow, Russia; ‡Moscow Institute of Physics and Technology, Moscow, Russia; §National Research University Higher School of Economics, Moscow, Russia; ʃANO «AI Research Institute», Moscow, Russia

The paper presents a fine-tuning methodology of the RuGPT3-XL (Generative Pretrained Transformer-3 for Russian) language model for the normalization of text spans task. The solution is presented in a competition for two tasks: Normalization of Named Entities (Named entities) and Normalization of a wider class of text spans, including the normalization of different parts of speech (Generic spans). The best solution has achieved 0.9645 accuracy on the Generic spans task and 0.9575 on the Named entities task. The presented solutions are in the public domain at https://github.com/RussianNLP/RuNormAS-solution.

OCULOMOTOR EVERYDAY COMMUNICATION: HOW TO PICK A GOOD METRIC

Fedorova O. V., Interdisciplinary Scientific and Educational School of Moscow University "Brain, Cognitive Systems, Artificial Intelligence", Moscow, Russia

This paper contributes to the research field of bimodal linguistics that explores two modalities involved in eve-ryday communication — vocal and kinetic. When exploring almost any scientific phenomenon, one addresses two opposite issues: individual differences, on the one hand, and general patterns, on the other. We have focused on the individual differences and proposed a "portrait" approach to communication. We are faced with a difficult task to find a good metric for analyzing oculomotor behavior of people in everyday communication. In previous papers, starting from [14], the authors were looking for oculomotor patterns, but their results depend critically on the metric used. In this paper, we compared the most common metrics and showed that individual differences have a much more serious weight than general patterns. We then identified four coefficients that determine these individual differences: kaside, kvip, kchain, and dur75. By comparing these Core Oculomotor Portraits, we were able to make these individual differences more clear. However, a fact is a fact: there are far more individual differences than general patterns between our Narrators behavior. The proposed coefficients, in our opinion, clearly show (and even explain and predict) the observed individual differences.

TEXT SIMPLIFICATION WITH AUTOREGRESSIVE MODELS

Fenogenova A., Sberbank, SberDevices, Moscow, Russia

Text Simplification is the task of reducing the complexity of the vocabulary and sentence structure of the text while retaining its original meaning with the goal of improving readability and understanding. We explore the capability of the autoregressive models such as RuGPT3 (Generative Pre-trained Transformer 3 for Russian) to generate high quality simplified sentences. Within the shared task RuSimpleSentEval we present our solution based on different usages of RuGPT3 models. The following setups are described: 1) few-shot unsupervised generation with the RuGPTs models 2) the effect of the size of the training dataset on the downstream performance of fine-tuned model 3) 3 inference strategies 4) the downstream transfer and post-processing procedure using pretrained paraphrasers for Russian. This paper presents the second-place solution on the public leaderboard and the fifth-place solution on the private leaderboard. The proposed method is comparable with the novel state-of-the-art approaches. Additionally, we analyze the performance and discuss the flaws of RuGPTs generation.

RUSSIAN SUPERGLUE 1.1: REVISING THE LESSONS NOT LEARNED BY RUSSIAN NLP-MODELS

Fenogenova A.[1], **Shavrina T.**[1,2,3], **Kukushkin A.**[5], **Tikhonova M.**[1,2], **Emelyanov A.**[1,4], **Malykh V.**[6,7], **Mikhailov V.**[1,2], **Shevelev D.**[1], **Artemova E.**[2,6]
[1]SberDevices, Sberbank, Moscow, Russia; [2]National Research University Higher School of Economics, Moscow, Russia; [3]ANO "AI Research Institute", Moscow, Russia; [4]Moscow Institute of Physics and Technology, Moscow, Russia; [5]Alex Kukushkin Lab, Moscow, Russia; [6]Huawei Noah's Ark lab, Moscow, Russia; [7]Kazan Federal University, Kazan, Russia

In the last year, new neural architectures and multilingual pre-trained models have been released for Russian, which led to performance evaluation problems across a range of language understanding tasks.

This paper presents Russian SuperGLUE 1.1, an updated benchmark styled after GLUE for Russian NLP models. The new version includes a number of technical, user experience and methodological improvements, including fixes of the benchmark vulnerabilities unresolved in the previous version: novel and improved tests for understanding the meaning of a word in context (RUSSE) along with reading comprehension and common sense reasoning (DaNetQA, RuCoS, MuSeRC). Together with the release of the updated datasets, we improve the benchmark toolkit based on jiant framework for consistent training and evaluation of NLP-models of various architectures which now supports the most recent models for Russian. Finally, we provide the integration of Russian SuperGLUE with a framework for industrial evaluation of the open-source models, MOROCCO (MOdel ResOurCe COmparison), in which the models are evaluated according to the weighted average metric over all tasks, the inference speed, and the occupied amount of RAM. Russian SuperGLUE is publicly available at https: //russiansuperglue.com/.

TRADITIONAL MACHINE LEARNING AND DEEP LEARNING MODELS FOR ARGUMENTATION MINING IN RUSSIAN TEXTS

Fishcheva I. N., **Goloviznina V. S.**, Vyatka State University, Kirov, Russia; **Kotelnikov E. V.**, Vyatka State University, Kirov, Russia; ITMO University, Saint Petersburg, Russia

Argumentation mining is a field of computational linguistics that is devoted to extracting from texts and classifying arguments and relations between them, as well as constructing an argumentative structure. A significant obstacle to research in this area for the Russian language is the lack of annotated Russian-language text corpora. This article explores the possibility of improving the quality of argumentation mining using the extension of the Russian-language version of the Argumentative Microtext Corpus (ArgMicro) based on the machine translation of the Persuasive Essays Corpus (PersEssays). To make it possible to use these two corpora combined, we propose a Joint Argument Annotation Scheme based on the schemes used in ArgMicro and PersEssays. We solve the problem of classifying argumentative discourse units (ADUs) into two classes — "pro" ("for") and "opp" ("against") using traditional machine learning techniques (SVM, Bagging and XGBoost) and a deep neural network (BERT model). An ensemble of XGBoost and BERT models was proposed, which showed the highest performance of ADUs classification for both corpora.

RUBTS: RUSSIAN SENTENCE SIMPLIFICATION USING BACK-TRANSLATION

Galeev F., **Leushina M.**, Innopolis University, Innopolis, Russia; **Ivanov V.**, Innopolis University, Innopolis, Russia; Kazan Federal University, Kazan, Russia

Automatic text simplification is a crucial task enabling to reduce text complexity while preserving meaning. This paper presents our solution to the Russian Sentence Simplification Shared Task (RSSE) based on a backtranslation technique. We show that applying the simple back-translation approach for sentence simplification can give competitive results with the other methods without fine-tuning or training.

TRANSFER LEARNING FOR IMPROVING RESULTS ON RUSSIAN SENTIMENT DATASETS

Golubev A., Bauman Moscow State Technical University, Russia; **Loukachevitch N.**, Lomonosov Moscow State University, Russia

In this study, we test transfer learning approach on Russian sentiment benchmark datasets using additional train sample created with distant supervision technique. We compare several variants of combining additional data with benchmark train samples. The best results were achieved using three-step approach of sequential training on general, thematic and original train samples. For most datasets, the results were improved by more than 3% to the current state-of-the-art methods. The BERT-NLI model treating sentiment classification problem as a natural language inference task reached the human level of sentiment analysis on one of the datasets.

A CORPUS-BASED MODEL OF THE ENGLISH PHRASAL VERB CONSTRUCTION: ATTRACTION

Golubkova E., **Trubochkin A.**, Moscow State Linguistic University, Moscow, Russia

The article investigates the semantic of English phrasal verbs (PhVs) which are viewed as lexico-grammatical constructions. Triangulation of introspective, cognitive and corpus methods of analysis allows us to identify the semantic dimensions which feature the semantic pattern of the PhV-construction. The construction reveals the features of attraction involving new verbs provided the action or motion event is identical. Depending on the attraction strength level between the verb and the particle a new verb may be accepted to fill in the corresponding slot of the construction, which gives rise to a new phrasal verb. It allows us to categorise PhVs according to the attraction level and spot their PhV-patterns on corpus data.

RUSSIAN NEWS CLUSTERING AND HEADLINE SELECTION SHARED TASK

Gusev I. O., Moscow Institute of Physics and Technology, Moscow, Russia; **Smurov I. M.**, ABBYY, Moscow, Russia

This paper presents the results of the Russian News Clustering and Headline Selection shared task. As a part of it, we propose the tasks of Russian news event detection, headline selection, and headline generation. These tasks are accompanied by datasets and baselines. The presented datasets for event detection and headline selection are the first public Russian datasets for their tasks. The headline generation dataset is based on clustering and provides multiple reference headlines for every cluster, unlike the previous datasets. Finally, the approaches proposed by the shared task participants are reported and analyzed.

UNREASONABLE EFFECTIVENESS OF RULE-BASED HEURISTICS IN SOLVING RUSSIAN SUPERGLUE TASKS

Iazykova T., HSE University, Moscow, Russia; **Bystrova O.**, HSE University / Sberbank, Moscow, Russia; **Kapelyushnik D.**, HSE University, Moscow, Russia; **Kutuzov A.**, University of Oslo, Norway

Leaderboards like SuperGLUE are seen as important incentives for active development of NLP, since they provide standard benchmarks for fair comparison of modern language models. They have driven the world's best engineering teams as well as their resources to collaborate and solve a set of tasks for general language understanding. Their performance scores are often claimed to be close to or even higher than the human performance. These results encouraged more thorough analysis of whether the benchmark datasets featured any statistical cues that machine learning based language models can exploit. For English datasets, it was shown that they often contain annotation artifacts. This allows solving certain tasks with very simple rules and achieving competitive rankings.

In this paper, a similar analysis was done for the Russian SuperGLUE (RSG), a recently published benchmark set and leaderboard for Russian natural language understanding. We show that its test datasets are vulnerable to shallow heuristics. Often approaches based on simple rules outperform or come close to the results of the notorious pre-trained language models like GPT-3 or BERT. It is likely (as the simplest explanation) that a significant part of the SOTA models performance in the RSG leaderboard is due to exploiting these shallow heuristics and that has nothing in common with real language understanding. We provide a set of recommendations on how to improve these datasets, making the RSG leaderboard even more representative of the real progress in Russian NLU.

ON DEVELOPING A WEB RESOURCE TO STUDY ARGUMENTATION IN POPULAR SCIENCE DISCOURSE

Ilina D., Novosibirsk State University, Institute of Philology, RAS, Novosibirsk, Russia; **Kononenko I.**, **Sidorova E.**, A. P. Ershov Institute of Informatics Systems, RAS, Novosibirsk, Russia

This paper discusses the experience of developing a web resource intended to study argumentation in popular science discourse. Such type of argumentation is, on the one hand, the main mean of achieving a communicative goal and, on the other hand, often not expressed in explicit form. The web resource is built around a corpus of 2256 articles, distributed over 13 subcorpora. The annotation model, which is based on the ontology of argumentation and D. Walton's argumentation schemes for presumptive reasoning, underlies the argument annotation of the corpus. The distinc-tive features of the argument annotation model are the introduction of weighting characteristics into text markup through assessing the persuasiveness of the argumentation, as well as highlighting argumentative indicators visually. The paper considers a scenario of argument annotation of texts, which allows constructing an argumentative graph based on the typical reasoning schemes. The scenario includes a number of procedures that enable the annotator to check the quality of the text markup and assess the persuasiveness of the argumentation. The authors have annotated 162 texts, using the developed web resource, and as a result, identified the most frequent schemes of argumentation (Example Inference, Cause to Effect Inference, Expert Opinion Inference), as well as described some specific indicators of frequent schemes. Based on the above-mentioned outcomes, the authors listed the indicators of the most frequent schemes of argumentation and made some recommendations for annotators about identifying the main thesis.

DEFINING DISCOURSE RELATIONS: SUPRACORPORA DATABASE OF CONNECTIVES

Inkova O., Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland

The research is focused on definitions of discourse relations, a topic that is currently little-studied. The paper gives a brief overview of existing solutions for discourse relations definitions: Rhetorical Structure Theory (RST), Segmented Discourse Representation Theory (SDRT), Penn Discourse Treebank (PDTB), and Cognitive approach to Coherence Relations. The author shows criteria used to define a discourse relation, or, in case of a narrower definition, a logical-semantic relation, in these approaches and outlines the shortcomings of the described definitions. The author also describes the principles used to build the classification and the definitions of logical-semantic relations (LSR) in the Supracorpora Database of connectives (SDB). The classification is based on four basic semantic operations upon which rests every LSR's definition: implication, location on the chronological scale, comparison, correlation between specific and general or an element and a set. The classification consistently distinguishes the levels at which the LSR can be established: propositional, illocutionary, and metalinguistic. Each LSR is defined on the basis of these two criteria. Thus, for example, for the LSR of alternative based on the comparison operation, one has the choice between the LSR of propositional, illocutionary and metalinguistic alternative (We will go to the mountains or to the sea vs. Put the gun away, or are you scared? vs. The symbol of the year or, simply speaking, cutie-pie). In case of LSRs based on implication or comparison, the polarity criterion is added, distinguishing whether the LSR is established between p and q or their negative correlates ¬ p and ¬ q are also to be taken into account in order to obtain a correct interpre-tation (cf. well-known descriptions of how the Russian conjunction no 'but' functions). In addition, semantic and pragmatic characteristics of the context are also considered in the classification. For example, in the case of the LSR of specification and generalization, the semantic correlation between p and q (together with their intensional and extensional interpretations) is taken heed of. Several definitions of LSR and corresponding examples are provided. Thus, the LSR of extensional specification is defined as follows: based on the operation of correlation between the general and the particular; established at the propositional level; X contains a generalized notion or state of things p; Y contains a more particular q-notion, limiting p-extensional. And the LSR of intensional specification is defined as follows: based on the operation of correlation between the general and the particular; established at the metalinguistic level; X contains a generalized concept or state of things p; Y contains a more particular q-notion, limiting p-inten-sional. The definitions used in the SDB definitions make it possible to evaluate, on the basis of the proposed criteria, the semantic closeness of relations and increase the level of consistency in the work of experts and annotators. That in turn increases the value of the annotated material, and therefore its reliability.

DIVERGENT TRANSLATION OF CONNECTIVES IN HUMAN AND MACHINE TRANSLATIONS

Inkova O., Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland;
Nuriev V., Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia

The paper is focused on divergent ways of conveying discourse relations in translation. For data collection, we used the supracorpora database of connectives storing parallel texts from the Russian-French subcorpus of the Russian National Corpus. These data show what logical-semantic relations tend to be translated using divergent ways, i.e. other than connectives (exclusion in its various gradations, propositional concomitance and substitution, the share of divergent translations ranging from 30% to 50%). Also, such data help define what causes divergent ways of transla-tion to be used. The causes may be as follows: (a) the lack of an adequate equivalent of a given connective in the target language; (b) differences in the syntactic structure of the source and target languages; (c) usage differences; (d) contextually determined use of divergent translation. If there is a prototypical indicator of logical-semantic relations (i.e. connective) in the source text, it also occurs in translation in more than 90% of cases. The data on human trans-lations are then compared with those on machine translations, which shows that the machine translation system also tends to keep a connective if there is one in the source text (it occurs in almost 98% of cases). However, there are cases where the machine translation system has difficulties processing a multiword connective (failing to perceive it as a whole) or a polyfunctional unit (failing to tell a connective from a non-connective) and thus uses divergent ways to translate it. Some causes of divergently translating connectives are likely to be the same for human and machine translations. These are differences in the syntactic structure of languages and usage differences. Further research of divergent means of conveying discourse relations will allow to draw a sharper border-line between explicitly ex-pressed and implicit discourse relations. The data collected from annotated corpora (both monolingual and multilin-gual and parallel) will help determine what the divergent ways of expressing logical-semantic relations are and how frequently they are used. The research results can be used both in automatic text processing and automatic text gen-eration. Also, the data on divergent translations of discourse relations can serve to improve the machine translation quality

THE RELATION OF CATEGORIES OF CONCRETENESS AND SPECIFICITY: RUSSIAN DATA

Ivanov V., Innopolis University, Innopolis, Russia; **Solovyev V.**, Kazan Federal University, Kazan, Russia

The categories of concreteness and specificity are important for understanding the mechanisms of information representation and processing in human brain. These two categories are quite close, but still different. A method for quantifying the degree of correlation of these categories for the English has recently been proposed. This paper deals with a similar research of the Russian. Ratings from the Concreteness/Abstractness Dictionary (RDCA) are taken as a measure of the words' concreteness. The degree of a word specificity is estimated by its location in the RuThes thesaurus. The paper represents the comparison with the English data and shows the similarity of the results for Rus-sian and English.

DATA PSEUDO-LABELING WHILE ADAPTING BERT FOR MULTITASK APPROACHES

Karpov D., **Burtsev M.**, Moscow Institute of Physics and Technology Dolgoprudny, Russia

Nowadays, BERT models have found wide use in the NLP field. However, standard BERT architecture training can be stifled by the lack of labels for different tasks while treating multitask settings as a one-task multilabel setting. For every example, we have labels from this example's source task but not from other tasks. This article addressed this issue, exploring eight different data pseudo-labeling approaches in the GLUE 4-task setting. These approaches do not require changes in samples or model architecture. One of the presented techniques excels results on RTE from the original article, by 6.2%, and falls behind the original article on QQP, MNLI, and SST only by 0.5–1.2%. This way also excels other pseudo-labeling approaches explored in the article by 0.5-2% on average if we consider similar tasks. However, for tasks that are dissimilar to each other, different proposed approach yields the best results.

ADJUNCT ROLE LABELING FOR RUSSIAN

Kazakov R., National Research University, Higher School of Economics, Moscow, Russia; **Lyashevskaya O.**, National Research University, Higher School of Economics; V. V. Vinogradov Russian Language Institute of RAS, Moscow, Russia

The task of the semantic role labeling usually focuses on identifying and classifying the core, obligatory arguments of the predicate. The adjuncts of Time, Location, etc. (noncore, modifier arguments) are considered on the periphery of the task [30] and even doing the easy part of it [44], despite the fact that they are highly integrated into the clause structure and may nontrivially interact with the meaning of the verb [4, 32]. In this paper, we present experiments on labeling the adjunct roles of LOCATION, TIME, MANNER, DEGREE, REASON, and PURPOSE, based on the manually annotated AdjunctsFrameBank data set. The results show an average F1score of 0.94 on the gold adjunct phrase annotations using the word2vec representations of adjuncts, word2vec representations of predicates, and the moprhosyntactic marking of adjuncts. Our findings generally corroborate the theoretical hypothesis on the structural and semantic autonomy and lexicomorphosyntactic specialization of adjuncts. Yet, more complicated organization of their network is revealed, pointing to the diversity of adjuncts in terms of their distribution and behavior.

A NEW ELECTRONIC SYSTEM FOR COMPARATIVE ANALYSIS OF VERSE AND PROSE

Kazartsev E., **Zemskova T.**, National Research University Higher School of Economics

This paper will focus on the development of a new computational system, Prosimetron, which enables compar-ative statistical studies of the rhythm of verse and prose in different languages (currently 10 languages are operative, with the possibility of adding more). The results of the analysis can be used not only for studying the processes for the genesis, expansion, and modification of various versification systems, but also for commenting on and interpreting the verse rhythm in different national poetic traditions in comparison with their foreign sources and language prosody. In addition, the possibility to model various processes of poetic speech generation and to analyze rhythmic vocabu-laries of prose allows hypotheses about the cognitive mechanisms of verse generation. This system operates in a semi-automatic mode and, by minimizing errors and enabling the processing of large amounts of data, provides a unique tool for computer research on the rhythm of different modes of speech.

BERT FOR RUSSIAN NEWS CLUSTERING

Khaustov S. V., **Gorlova N. E.**, **Kalmykov A. V.**, **Kabaev A. S.**, MTS AI, Moscow, Russia

This paper provides results of participation in the Russian News Clustering task within Dialogue Evaluation 2021. News clustering is a common task in the industry, and its purpose is to group news by events. We propose two methods based on BERT for news clustering, one of them shows competitive results in Dialogue 2021 evaluation. The first method uses supervised representation learning. The second one reduces the problem to binary classification.

LOWRESOURCEEVAL2021: A SHARED TASK ON SPEECH PROCESSING FOR LOWRESOURCE LANGUAGES

Klyachko E., HSE University, RAS Iling, Moscow, Russia; **Grebenkin D.**, **Nosenko D.**, NSU, NSU SDAML, Novosibirsk, Russia; **Serikov O.**, HSE University, DeepPavlov, MIPT, Moscow, Russia

This paper describes the results of the first shared task on speech processing for lowresource languages of Russia. Speech processing tasks are notoriously dataconsuming. The aim of the shared task was to evaluate the performance of stateoftheart models on low-resource language data as well as draw the attention of experts to field linguistics data (using Lingovodoc project data). The tasks included language identification and IPA transcription, with three teams participating in them. The paper also provides a description for the datasets as well as an analysis of the participants' solutions. The datasets created as a result of the shared task can be used in other tasks to enhance speech processing and help develop modern NLP tools for both speech communities and field linguists.

THE INTONATION OF 'YES' AND 'NO' IN AN ARCHAIC RUSSIAN DIALECT

Knyazev S. V., Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia; **Pronina M. K.**

Universitat Pompeu Fabra, Barcelona, Spain

The present paper analyzes the intonation of pragmatic particles da "yes" and net "no" found in the spontaneous dialogue speech corpus of a Northern Russian dialect, in which each word bears a pitch accent. Intonation that marks such particles sounds unusual for speakers of Standard Russian and is perceived by them as blunt and impolite. The main aim was to find a consistent pattern explaining the distribution of falling and rising pitch accents on such particles in a dialect of Vaduga (Arkhangelsk region). We tested three hypotheses that can account for this distribution: (a) semantic explanation (the type of pitch accent depends on the semantics of the very particle); (b) communicative explanation (it depends on the communicative function of the preceding utterance, that is, whether it is a question or not); (c) phonetic explanation (it depends on the pitch accent of the preceding utterance). A total of 240 utterances from 3 speakers were analyzed. Results showed that the semantics of the particle is not a relevant factor, while the communicative type and the pitch accent of the preceding utterance are significant predictors of the pitch accent that marks the particle, with the latter better explained the data. We propose that when analyzing the intonation of a dialect, semantic interpretation of the intonational constructions of the standard dialect should not be taken into account. Moreover, we suggest that a new approach of collecting prosodic data with elderly people while controlling for pragmatic context is needed

PARENTHETICAL CONSTRUCTIONS IN RUSSIAN SPOKEN DISCOURSE: BASIC TYPES AND PROSODIC FEATURES

Korotaev N. A., Russian State University for the Humanities (RSUH), Moscow, Russia

The paper discusses the notion of parentheticals in Russian spoken discourse. Using data from two prosodically annotated corpora — "Stories about presents and skiing" and "Russian Pear Chats & Stories" — I advocate for a discourse-oriented approach to parenthetical constructions. I define a parenthetical construction as consisting of three elements: the left context, the parenthetical unit, and the right context. Each element constitutes a separate discourse unit and is thus prosodically autonomous. I rely on the notion of projection [Auer 2005] to account for the discourse relationships between these three components. When the speaker pronounces the left context, she projects a continuation that is to be realized in the right context, while the parenthetical unit provides a digressive discourse step.

Typically (around 50% in my data), parentheticals are anchored to their left contexts and are pronounced with a falling or level pitch accent. Noted deviations from this prototype include free parentheticals, parenthetical uses of vot, and parentheticals pronounced with a rising pitch accent. Furthermore, I explore two prosodic features frequently associated with parentheticals, namely, increased articulation rate and pitch range narrowing. I show that, while both these tendencies are statistically significant, the latter has a larger effect size than the former.

AUDIO AND TEXT-DRIVEN APPROACH FOR CONVERSATIONAL GESTURES GENERATION

Korzun V. A., MIPT, Moscow, Russia; **Dimov I. N.**, MSU, Moscow, Russia; **Zharkov A. A.**, MIPT, Moscow, Russia

This paper describes FineMotion's gesture generating system entry for the GENEA Challange 2020. We start by using simple baselines and expand them by using context and combining both audio and textual features. Among the participating systems, our entry attained the highest median score in the human-likeness evaluation and second highest median score in appropriateness.

CURRENT LANDSCAPE OF THE RUSSIAN SENTIMENT CORPORA

Kotelnikov E. V., Vyatka State University, Kirov, Russia; ITMO University, Saint Petersburg, Russia

Currently, there are more than a dozen Russian-language corpora for sentiment analysis, differing in the source of the texts, domain, size, number and ratio of sentiment classes, and annotation method. This work examines publicly available Russian-language corpora, presents their qualitative and quantitative characteristics, which make it possible to get an idea of the current landscape of the corpora for sentiment analysis. The ranking of corpora by annotation quality is proposed, which can be useful when choosing corpora for training and testing. The influence of the training dataset on the performance of sentiment analysis is investigated based on the use of the deep neural network model BERT. The experiments with review corpora allow us to conclude that on average the quality of models increases with an increase in the number of training corpora. For the first time, quality scores were obtained for the corpus of reviews of ROMIP seminars based on the BERT model. Also, the study proposes the task of the building a universal model for sentiment analysis.

NO WAY! DISCOURSE FORMULAE OF DISAGREEMENT IN RUSSIAN AND ENGLISH: A COMPARATIVE STUDY

Koziuk E., **Badryzlova Y.**, HSE University Moscow, Russia

The study explores the discourse formulae (DFs) of disagreement in Russian and English belonging to the sub-classes of refusal and prohibition. Starting with a subset of six Russian target DFs, we establish their English equivalents using corpus analysis. We also define the typical speech acts to which the DFs in both languages react, and design model contexts that exemplify these types of speech acts. We use the model contexts as stimuli in our Russian and English surveys where we look at the preferences of native speakers in choice of DFs across the speech acts. We use the data of the surveys to establish the pragmatic function of each DF, (i.e. refusal or prohibition, or both), and their potential in each subclass (strong, medium, or weak). For each DF, we also identify the types of speech acts to which they react most readily. We compare the results of our analysis to the lexicographic description of the target DFs as presented in the Russian-English Dictionary of Idioms.

THE TYPES OF INFINITIVE CONSTRUCTIONS WITH PREDICATIVES (ACCORDING TO THE RUSSIAN NATIONAL CORPUS)

Kustova G. I., Vinogradov Russian Language Institute of the Russian Academy of Sciences; Moscow, Russia

The paper considers constructions «predicative + infinitive». For the first time, a class of interpretive infinitive constructions (opposed to emotional reactions) is introduced. For emotional reactions, the predicative and the infinitive refer to the same subject, the infinitives of the perception, mental, speech verbs are typical for them: It hurts / scares to see how forests are dying ('X sees, X is scared') → It hurts that forests are dying. For interpretive constructions, the subjects of the predicative and the infinitive do not coincide: It is heartless to separate the mother from the children — 'X separates, Y evaluates such an act as heartless'. The infinitives of perceptual and mental verbs in such a construction are either not used, or they denote a kind of action: It is tactless to listen to private conversations.

ANIMACY IN THE USE OF ANAPHORIC AND DEMONSTRATIVE PRONOUNS IN RUSSIAN AND FRENCH

Letuchiy A., **Nikishina E.**, HSE University, Moscow, Russia

The article focuses on the role of animacy in Russian and French pronominal systems. Although animacy is a grammatical category only in Russian, while in French it is not reflected in the behavior of nouns, it turns out that some animacy-based restrictions on the use of anaphoric and demonstrative pronouns are common for the two lan-guages. We address syntactic restrictions that affect the following types of uses: (i) use of anaphoric pronouns in copular constructions; (ii) repetition of anaphoric pronouns for the sake of clearness and / or emphasis; (iii) deictic use of anaphoric pronouns; (iv) anaphoric use of demonstrative pronouns. In all the four cases, except, perhaps, the fourth one, pronouns tend to have an animate referent, while inanimate ones are more problematic. We conclude that these restrictions mainly result from the fact that animate objects have a greater discourse importance and more often become the main subject of the discourse than inanimate ones. At the same time, degree of strictness of restrictions sometimes differ between the two languages: for instance, demonstrative pronouns in the anaphoric use tend to have an animate antecedent in Russian, while for French, this tendency is weaker.

THE SEMANTIC COMPONENT *SCALE* IN THE MEANING OF A DISCOURSE PARTICLE *UZH*

Levontina I. B., Russian Language Institute RAS

The modal particle uzh is perhaps the most difficult Russian discourse word to describe since its semantics is highly elusive. The existing descriptions are rather abstract and poorly correlate with various cases of usage of uzh. Besides, they do not take into consideration several crucial components of this particle's meaning. For instance, in phrases like Uzh ya-to znayu ('I do know') one can notice a hugely important component of meaning—the idea of a scale. One can say Ya-to etot sekret znayu, a vot drugim nevdomek ('I do know the secret, whereas others have no idea about it'), and in this example, uzh would be irrelevant. Uzh ya-to eto znayu

presupposes that others probably know it too, but it's me who knows it for sure. This very idea of a scale and poles together with the idea of the exceedance of expectations (which is also important for the meaning of uzh) constitutes the semantic contribution that this particle makes. Moreover, uzh partly smooths the opposition between the central and other elements of a multi-tude, because it does not exclude them from consideration, it just gives emphasis to that one.

The aim of this research is to examine those types of uzh usage, where the idea of a scale is most clearly actual-ized. Probably, if we understand how the significant components of this particle's meaning function, we will get closer to the development of a complete picture of its usage. For example, the idea of a scale within the meaning of uzh is expressed in the context of a special question (Zachem uzh tak zlo? 'Why so mean?'). In an argument uzh often implies that the speaker was almost ready to back down, but not to this extent — like in a famous poem by Daniil Kharms called «Liar» (1930). The idea of a scale is vividly realized in the context of an implicit (Gde uzh mne!, 'How can I…') or explicit negation. It is especially interesting to pay attention to the peculiar effects of the combination of uzh with comparative forms (luchshe uzh, 'it would be better...'). The usage of uzh in standard word combinations raz uzh, esli uzh, togda uzh has its restrictions, also connected with the idea of a scale. The development of a modal meaning in a temporal word, which brings the transformation of a timeline into a scale of expectations or possibilities, is quite typical.

GENDER AND CASE IN RUSSIAN NOUNS DENOTING PROFESSIONS AND SOCIAL ROLES

Magomedova V. D., Independent researcher, Russia; **Slioussar N. A.**, NRU HSE, Moscow, SPbU, Saint Petersburg, Russia

In the present paper, we analyzed a group of Russian nouns denoting professions and social roles. Historically, these nouns were masculine; in modern Russian, they can also be used with feminine agreement, but only nominative forms are regarded as normative (e.g. etot / eta vrač 'thisM/F doctor'). We showed that oblique case feminine forms occur naturally using the Web-as-corpus approach and conducted three experimental studies. We discovered that offline rating and online processing of such forms depends on their case. Firstly, this is a unique example of the properties of the form influencing the properties of the lexeme. Secondly, the fact that all oblique forms are regarded as marginal and that locative was found to be significantly worse than other oblique cases points to a deep connection between grammatical gender and inflectional classes and to the crucial role of affix syncretism in morphological processing. This presents a challenge for different approaches in theoretical morphology.

MORPHOLOGICAL ANNOTATION OF SOCIAL MEDIA CORPORA WITH REFERENCE TO ITS RELIABILITY FOR LINGUISTIC RESEARCH

Michurina M., **Ivoylova A.**, ABBYY Lab, MIPT, Dolgoprudny, Russia; RSUH, Moscow, Russia; **Kopylov N.**, **Selegey D.**, ABBYY Lab, MIPT, Dolgoprudny, Russia

This paper presents the results of the study devoted to the applicability of SOTA methods for morphological corpus annotation (based on GramEval2020) for analytical sociolinguistic research. The study shows that statistically successful technologies of morphosyntactic annotation for such purposes create a number of problems for researchers if they are used purely i.e. without any linguistic knowledge. In this paper, methods for improving the morphological annotation, successfully implemented in GICR, from the point of view of its reliability are presented.

EXPERIMENTS ON HUMAN INCREMENTAL PARSING OF ENGLISH

Mityushin L., **Iomdin L.**, A. A. Kharkevich Institute for Information Transmission Problems Russian Academy of Sciences

Experiments have been carried out in which human subjects incrementally constructed dependency trees of English sentences. The subjects were successively presented with growing initial segments of a sentence, and had to draw syntactic links between the last word of the segment and the previous words. They were also shown a fixed number of lookahead words following the last word of the segment. The results of the experiments show that lookahead of 1 or 2 words is sufficient for confident incremental parsing of English declarative sentences.

COMMUNICATION FAILURES IN EVERYDAY CONVERSATIONS: A CASE STUDY BASED ON THE "RETROSPECTIVE COMMENTING METHOD"

Mustajoki A., HSE University, Moscow, Russia; Helsinki University, Finland; **Cherkunova N.**, **Sherstinova T.**, HSE University, St. Petersburg, Russia

The paper deals with communication failures in everyday spoken discourse. The spontaneous character of oral speech is its basic property and becomes a prerequisite for the appearance of such a phenomenon as communicative failures. By communicative failures, we mean speech situations when the recipient of a speech message does not understand it correctly, i.e., in the way the speaker intended. The purpose of this pilot study is 1) to assess the total number of communication failures that occur with a person during a single day and 2) to determine the dependence of communication failure frequency on the communication settings and conditions. The main result of the study is a qualitative and quantitative assessment of communication failures during a subjects's day. The research is based on a special experiment based on 24-hour monitoring of the subject's speech and his subsequent retrospective commentary on all recorded data. Such an approach allows one to reduce the subjectivity inherent in much linguistic work. The research continues a series of studies devoted to the effectiveness of spoken communication and is important not only for understanding the fundamental processes of speech perception but is also crucial for the development of artificial intelligence systems involving human-computer speech dialogue systems and for speech technologies of the next generation.

RUSIMSCORE: UNSUPERVISED SCORING FUNCTION FOR RUSSIAN SENTENCE SIMPLIFICATION QUALITY

Orzhenovskii M. V., Saint Petersburg, Russia

We propose an unsupervised complex scoring function (RuSimScore) to measure simplification quality of Russian sentences, and a model for text simplification based on this function. The function allows to score simplicity and original meaning preservation. First, filtered a noisy parallel corpus (machine translated WikiLarge) and extracted good simplification examples. After that, a pretrained language model was fine-tuned on these examples. We generate multiple outputs from the language model and select the best one according to the scoring function. The weights in the scoring function can be adjusted to balance between better content preservation and getting simpler sentences (controllable simplification).

RUSHIFTEVAL: A SHARED TASK ON SEMANTIC SHIFT DETECTION FOR RUSSIAN

Pivovarova L., University of Helsinki, Finland; **Kutuzov A.**, University of Oslo, Norway

We present the first shared task on diachronic word meaning change detection for the Russian. The participating systems were provided with three sub-corpora of the Russian National Corpus — corresponding to pre-Soviet, Soviet and post-Soviet periods respectively — and a set of approximately one hundred Russian nouns. The task was to rank those nouns according to the degrees of their meaning change between periods.

Although RuShiftEval is in many respects similar to the previous tasks organized for other languages, we introduced several novel decisions that allow for using novel methods. First, our manually annotated semantic change dataset is split in more than two time periods. Second, this is the first shared task on word meaning change which provided a training set.

The shared task received submissions from 14 teams. The results of RuShiftEval show that a training set could be utilized for word meaning shift detection: the four top-performing systems trained or fine-tuned their methods on the training set. Results also suggest that using linguistic knowledge could improve performance on this task. Finally, this is the first time that contextualized embedding architectures (XLM-R, BERT and ELMo) clearly outperform their static counterparts in the semantic change detection task.

SEMANTICS, GRAMMAR AND PROSODY OF PARENTHETICALS INTRODUCED BY THE SUBORDINATOR 'KAK' *AS*

Podlesskaya V. I., **Pozhilov Ju. M.**, Russian State University for the Humanities, Moscow, Russia

Based on data from the multimedia subcorpus of the Russian National Corpus, the paper addresses syntactic, sematic and prosodic features of the particular type of quotations with the reporting frame headed by the subordinator kak 'as' (kak skazal mne staryj rab pered tavernoj…). Our data show mixed evidence regarding the parenthetical status of the construction. On the one hand, typically for parentheticals, its function is clearly pragmatized, since it expresses speaker's attitude towards the quote. On the other hand, typical parentheticals have only loose syntactic connection with their "host", while the kak-phrase is introduced by the subordinator and has the form of the standard adverbial clause. Further on, while typical parentheticals are characterized by grammatical and prosodic reduction, grammatical and prosodic restrictions operating in the kak-phrase are optional and context (e.g., word order) sensitive. The kind of data we present supports the approach to parenthesis that doesn't favor either/or decisions, but rather is based on multifactorial analysis that considers the whole range of possible parameters and isolates their observed language-specific clusters.

SEMSKETCHES2021: EXPERIMENTING WITH THE MACHINE PROCESSING OF THE PILOT SEMANTIC SKETCHES CORPUS

Ponomareva M.†‡, **Petrova M.**†, **Detkova J.**†, **Serikov O.**‡ʃ, **Yarova M.**§
†ABBYY, Moscow, Russia; ‡National Research University Higher School of Economics, Moscow, Russia; §Moscow Institute of Physics and Technology, Moscow, Russia; ʃDeeppavlov MIPT, Moscow, Russia

The paper deals with elaborating different approaches to the machine processing of semantic sketches. It presents the pilot open corpus of semantic sketches. Different aspects of creating the sketches are discussed, as well as the tasks that the sketches can help to solve. Special attention is paid to the creation of the machine processing tools for the corpus. For this purpose, the SemSketches2021 Shared Task was organized. The participants were given the anonymous sketches and a set of contexts containing the necessary predicates. During the Task, one had to assign the proper contexts to the corresponding sketches.

SHORT TEXT CLUSTERING WITH TRANSFORMERS

Pugachev L., **Burtsev M.**, Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Recent techniques for the task of short text clustering often rely on word embeddings as a transfer learning component. This paper shows that sentence vector representations from Transformers in conjunction with different clustering methods can be successfully applied to address the task. Furthermore, we demonstrate that the algorithm of enhancement of clustering via iterative classification can further improve initial clustering performance with classifiers based on pre-trained Transformer language models.

ZERO-SHOT CROSS-LINGUAL TRANSFER OF A GLOSS LANGUAGE MODEL FOR SEMANTIC CHANGE DETECTION

Rachinskiy M., HSE University, Moscow, Russia; **Arefyev N.**, Samsung Research Center Russia; Lomonosov Moscow State University; HSE University, Moscow, Russia

Consulting word definitions from a dictionary is a familiar way for a human to find out which senses a particular word has. We hypothesize that a system that can select a proper definition for a particular word occurrence can also naturally solve Semantic Change Detection (SCD) task. To verify our hypothesis, we followed an approach previously proposed for Word Sense Disambiguation (WSD) and trained a system that embeds word definitions and word occurrences into the same vector space. In this space, the embedding of the most appropriate definition has the largest dot product with a contextualized word embedding.

The system is trained on an English WSD corpus. To make it work for the Russian language, we replaced BERT with the multilingual XLMR language model and exploited its zeroshot crosslingual transferability. Despite not finetuning the encoder model on any Russian data, this system achieves the second place in the competition, and likely works for any of one hundred other languages XLMR was pretrained on, though the performance may vary. We then measure the impact of such WSD pretraining and show that this procedure is crucial for our results. Since our model was trained to choose a proper definition for a word, we propose an algorithm for the interpretation and visualization of the semantic changes through time.

By employing additional labeled data in Russian and training a simple regression model, that converts the distances between output contextualized embeddings into more humanlike scores of sense similarity between word occurrences, we further improve our results and achieve the first place in the competition.

SWITCHING TO WORK IN AN INCLUSIVITY WORKSHOP: MULTIMODAL ANALYSIS OF INTERACTION

Rudneva E. A., Institute for Linguistic Studies, St. Petersburg, Russia

The study focuses on switching from talk to work in an "inclusivity workshop" for people with mental disabili-ties. Work activities and conversation about general topics can be approached from the perspective of multiactivity and considered courses of actions intertwined in social interaction. The order of activities is negotiated among partic-ipants using both linguistic and non-linguistic means. The data are extracts of video recordings containing a partici-pant getting others to do things. The paper provides multimodal analysis of 6 cases of an instructor getting an autistic participant to switch to work, which occurred within a 17-minute conversation about animals. In the data, the autistic participant never provides a second-pair response to a directive. In 5 out of 6 cases analysed in the paper he fulfils the action to different extents, demonstrating various degrees of involvement. Getting the autistic person to switch to work is more effective when suggesting actions one by one, through concrete embodied actions, and when orienting to phases of the ongoing talk. The study highlights differences between autistic and non-autistic participants switching from one course of actions to another. Considering goals of an inclusivity workshop, success of switching to work can be also determined by the opportunities for the smooth conversation.

DETECTION OF SEMANTIC CHANGES IN RUSSIAN NOUNS WITH DISTRIBUTIONAL MODELS AND GRAMMATICAL FEATURES

Ryzhova A. A., Federal Research Center "Computer Science and Control" of the RAS, Lomonosov Moscow State University, Moscow, Russia; **Ryzhova D. A.**, **Sochenkov I. V.**, HSE University, Moscow, Russia

The paper presents the models detecting the degree of semantic change in Russian nouns developed by the team 'aryzhova' within the RuShiftEval competition of the Dialogue 2021 conference. We base our algorithms mostly on unsupervised distributional models and additionally test a model that uses vectors representing morphological preferences of the words in question. The best results are obtained by the model built on the ELMo architecture with a small window, while the quality of performance of the "grammatical" model is comparable to that of the models based on much more sophisticated algorithms.

RUSIMPLESENTEVAL-2021 SHARED TASK: EVALUATING SENTENCE SIMPLIFICATION FOR RUSSIAN

Sakhovskiy A.†, **Izhevskaya A.**‡, **Pestova A. S.**‡, **Tutubalina E. V.**†‡, **Malykh V. A.**†§, **Smurov I. M.**ʃ◊, **Artemova E. L.**‡§
†Kazan Federal University, Kazan, Russia; ‡National Research University Higher School of Economics, Moscow, Russia; §Huawei Noah's Ark lab, Moscow, Russia; ʃABBYY, Moscow, Russia; ◊Moscow Institute of Physics and Technology, Moscow, Russia

This report presents the results from the RuSimpleSentEval Shared Task conducted as a part of the Dialogue 2021 evaluation campaign. For the RSSE Shared Task, devoted to sentence simplification in Russian, a new middlescale dataset is created from scratch. It enumerates more than 3,000 sentences sampled from popular Wikipedia pages. Each sentence is aligned with 2.2 simplified modifications, on average. The Shared Task implies sequence-to-sequence approaches: given an input complex sentence, a system should provide with its simplified version. A popular sentence simplification measure, SARI, is used to evaluate the system's performance.

Fourteen teams participated in the Shared Task, submitting almost 350 runs involving different sentence simplification strategies. The Shared Task was conducted in two phases, with the public test phase allowing an unlimited number of submissions and the brief private test phase accepting one submission only. The post-evaluation phase remains open even after the end of private testing. The RSSE Shared Task has achieved its objective by providing a common ground for evaluating state-of-the-art models. We hope that the research community will benefit from the presented evaluation campaign.

SENTENCE SIMPLIFICATION WITH RUGPT3

Shatilov A. A., **Rey A. I.**, RANEPA, Moscow, Russia

This paper describes our solution for the RuSimpleSentEval shared task on sentence simplification held together with Dialogue 2021 conference. Our approach was to filter the provided dataset, finetune the pretrained ruGPT3 model on it and select generated simple candidates based on cosine similarity and ROUGEL with a complex sentence as an input. The system achieved SARI 38.49 and took third place in the competition. We have reviewed and analyzed examples of simplified sentences produced by the model. The analysis showed that the sentences produced by the system lose the original meaning of the input sentence in about half of the cases.

THE RUSSIAN PARTICLE 'ZHE' IN THE LIGHT OF PARALLEL CORPORA

Shmelev A., Vinogradov Russian Language Institute of the RAS, Moscow, Russia

This paper deals with the Russian particle zhe and its use in the Russian translations from English and demon-strates the possibilities of "one-focus analysis" in contrastive studies based on the parallel corpora. It correlates the explications of zhe given in earlier studies (it makes special reference to the Active Dictionary of Russian) with the stimuli to translation, that is, fragments of the original English text that might cause the appearance of zhe in a Russian translation as a reaction to those stimuli. The study sought to validate, disprove or improve the semantic analysis of zhe made without recourse to electronic corpora.

The analysis of the stimuli that have led Russian translators to use the particle zhe reveals important characteris-tics of this word. It turns out that the Russian particle zhe is often pragmatically obligatory as its absence would violate the idiomatic nature of the utterance and change its illocutionary force. It is often the case that if a translator had given word-for-word translation, that is without a particle, they would convey the precise meaning, but the translation would be inadequate: the wrong implicature would appear. On the other hand, when they add the particle, they may impart new shades of meaning which the original text did not contain.

AUTOMATIC DETECTION OF IMPLICIT AGGRESSION IN RUSSIAN SOCIAL MEDIA COMMENTS

Shulginov V. A., **Mustafin R. Zh.**, **Tillabaeva A. A.**, Higher School of Economics, Laboratory of Linguistic Conflict Resolution Studies and Contemporary Communicative Practices, Moscow, Russia

This article studies the characteristics of implicit and explicit types of aggression in the comments of a Russian social network with the means of machine learning. As it is hypothesized that expression of aggression depends on local norms, the dataset contains the comments collected from a single social media community. These comments were divided into three classes: polite communication, implicit aggression, and explicit aggression. Trying different combinations of data preprocessing, we discovered that lemmatization

and replacement emojis with placeholders contribute to better results. We tested several models (Naive Bayes, Logistic Regression, Linear Classifiers with SGD Training, Random Forest, XGBoost, RuBERT) and compared their results. The study describes the misclassifications and compares the keywords of each class of comments. The results can be helpful while enhancing the algorithm of detection of implicit aggression.

THE DYNAMICS OF VOCABULARY IN RUSSIAN PROSE (BASED ON FREQUENCY DICTIONARIES OF THE CORPUS OF RUSSIAN SHORT STORIES 1900–1930)

Skrebtsova T. G., **Grebennikov A. O.**, Saint Petersburg State University Saint Petersburg, Russia;
Sherstinova T. Yu., National Research University Higher School of Economics Saint Petersburg, Russia

The paper presents the results of a study that is part of a large-scale project aimed at studying the changes that took place in the Russian language during the first three decades of the 20th century. In the history of Russia, this period was marked by stormy events that led to a radical change in the state system and the formation of a new society. To quantify the scale of changes that occurred in the language in the result of these dramatic events, it is necessary to analyze the representative volume of linguistic data and to compare different chronological periods in dynamics using quantitative methods. The research was carried out on the data of an annotated sample from the Corpus of the Russian Short Stories of 1900–1930, which contains texts by 300 Russian writers. All the texts in the Corpus are divided into three time frames: 1) the pre-war period (1900–1913), 2) the war and revolutionary years (1914–1922) and 3) the early Soviet period (1923–1930). Frequency distribution of significant vocabulary in dynamics was analyzed, which made it possible to identify the main tendencies in the change of individual words and lexical groups frequencies from one historical period to another and to correlate them with the previously identified dynamics of literary themes. The technique used allows to trace the influence of large-scale political changes on the vocabulary of literary language, to note the peculiarities and tendencies of the writers' worldview in a certain historical period, and also makes it possible to significantly supplement the analysis of the dynamics of literary themes in fiction.

ON SLAVIC COGNATE RECOGNITION IN CONTEXT

Stenger I., **Avgustinova T.**, Saarland University, Saarbrücken, Germany

This study contributes to a better understanding of reading intercomprehension as manifested in the intelligibility of East and South Slavic languages to Russian native speakers in contextualized cognate recognition experiments using Belarusian, Ukrainian, and Bulgarian stimuli. While the results mostly confirm the expected mutual intelligibility effects, we also register apparent processing difficulties in some of the cases. In search of an explanation, we examine the correlation of the experimentally obtained intercomprehension scores with various linguistic factors, which contribute to cognate intelligibility in a context, considering common predictors of intercomprehension associated with (i) morphology and orthography, (ii) lexis, and (iii) syntax.

WHAT HAVE I SEEN? ON THE MEANING AND DISTRIBUTION OF AN EXPERIENTIAL DISCOURSE MARKER

Tatevosov S. G., Lomonosov Moscow State University, Moscow, Russia; **Kisseleva X. L.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper explores the discourse marker ja vižu (lit. 'I see') and its cross-linguistic counterparts. We argue that it presents its scope proposition as the product of abduction, a logical inference that derives the optimal explanation for the observed state of affairs. This view is supported by the set of observations suggesting that restrictions on the distribution of ja vižu are mostly derivable as restrictions on abuctive reasoning, which involve informativeness, likelihood and parsimony considerations.

META-EMBEDDINGS IN TAXONOMY ENRICHMENT TASK

Tikhomirov M. M., **Loukachevitch N. V.**, Lomonosov Moscow State University, Moscow, Russia

In this paper we consider the taxonomy enrichment task based on a recently appeared dataset, called Diachronic wordnets, created on the basis of English and Russian wordnets. We study meta-embeddings approaches, which combine several source embeddings, to the hypernym prediction of novel words and show that meta-embedding approaches obtain the best results for this task if compared to other methods based on different principles. When combining with automatically extracted features from theWiktionary online dictionary, the joint approach improves the results.

RUSSIAN NEWS SIMILARITY DETECTION WITH SBERT: PRE-TRAINING AND FINE-TUNING

Vatolin A. S., **Smirnova E. Y.**, **Shkarin S. S.**, SberBank, Moscow, Russia

Computation of text similarity is one of the most challenging tasks in NLP as it implies understanding of semantics beyond the meaning of individual words (tokens). Due to the lack of labelled data this task is often accomplished by means of unsupervised methods such as clustering. Within the DE2021: "Russian News Clustering and Headline Selection" we propose a method of building robust text embeddings based on Sentence Transformers architecture, pretrained on a large dataset of in-domain data and then fine-tuned on a small dataset of paraphrases leveraging GlobalMultiheadPooling.

AUTOMATIC DETECTION OF DECEPTIVE AND TRUTHFUL PARALINGUISTIC INFORMATION IN SPEECH USING TWO-LEVEL MACHINE LEARNING MODEL

Velichko A. N., **Karpov A. A.**, SPC RAS, Saint-Petersburg, Russia

In this work, we present a novel approach to one of computational paralinguistic tasks — automatic detection of deceptive and truthful information in human's speech. This task belongs to the aspects of destructive behaviour and was first presented at the International INTERSPEECH Computational Paralinguistics Challenge ComParE in 2016. The need of contactless method for deception detection follows from the fact that existing contact-based approaches such as polygraphs and lie detectors have multiple restrictions, which significantly limit their usage. Both for training and testing of the proposed models we used two English-language corpora (Deceptive Speech Database and Real-Life Trial Deception Detection Dataset). We extracted tree sets of acoustic features from those

audio samples using openSMILE toolkit. The proposed approach includes preprocessing of the extracted acoustic features with the usage of methods for data augmentation and dimensionality reduction of feature space. We have got 1680 speech utterances and 986-dimensional informative feature vector for each utterance. The main part of the proposed approach is two-level recognition model, where the first level includes three models of gradient boosting (Catboost, XGBoost and LightGBM). The second level consists of logistic regression-based model for final prediction on truthfulness or deceptiveness that takes into account predictions from the first level. Using this approach, we have achieved the result of classification in terms of F-score = 85.6%. The proposed approach can be used both independently and as a component of multimodal systems for detection of deceptive and truthful utterances in speech, as well as in systems for detection of a destructive behaviour.

TRANSFORMERS FOR HEADLINE SELECTION FOR RUSSIAN NEWS CLUSTERS

Voropaev P. M., **Sopilnyak O. A.**, Moscow Institute of Physics and Technology, Moscow, Russia

In this paper, we explore various multilingual and Russian pre-trained transformer-based models for the Dialogue Evaluation 2021 shared task on headline selection. Our experiments show that the combined approach is superior to individual multilingual and monolingual models. We present an analysis of a number of ways to obtain sentence embeddings and learn a ranking model on top of them. We achieve the result of 87.28% and 86.60% accuracy for the public and private test sets respectively.

THE PROSODY OF SPOKEN DIALOGUE

Yanko T. E., Institute of Linguistics, Moscow, Russia

This paper is aimed at establishing the parameters of the dialogic communication expressed through Russian prosody. The linguistic and extra-linguistic constituents of dialogue are analyzed. These are: the illocutionary mean-ings that generate speech acts, characteristic of the dialogic communication; the discourse links that combine the successive speech acts of one interlocutor if his/her current contribution into the dialogue is not limited to a single speech act; the prosodic characteristics of genre typical for a concrete type of communication (a friendly talk, an exam, a press conference, a scientific presentation, or an interrogation). The proposed taxonomy is based on the analysis of the minor working corpus of spoken dialogues from the Russian National corpus (Multimodal sub-corpus Murko), the annotated database Spokencorpora.ru, video-hosting Youtube.com, films, scientific conferences, and press conferences. The computer system Praat is used to analyze the sound data. The paper is illustrated with tracings of sound records.

RUSSIAN DISCOURSE MARKERS 'VIDIMO' AND 'PO-VIDIMOMU' (*APPARENTLY*): SYNCHRONIC AND DIACHRONICAL SEMANTICS

Zalizniak Anna A., Institute of Linguistics of the RAS, Institute of Informatics Problems of the RAS, Moscow, Russia

The article analyzes the meaning of Russian discursive words vidimo and po-vidimomu ('apparently'), and reconstructs the ways of their semantic evolution over the past two centuries. It is shown that the meaning of an inference made by the speaker on the basis of some data, which is the only one for both words in modern language, arose in different ways. The semantic evolution of both words includes the the replacement of the meaning of visual perception with the meaning of epistemic evaluation and the acquisition of egocentric semantics. The word vidimo initially served as a marker of a true visual impression; the word po-vidimomu which initially included an interpretative component, acquired the meaning of a potentially false judgment, which was subsequently lost. The research is based on texts included in the Russian National Corpus (www.ruscorpora.ru).

LINKING THE ANCIENT GREEK WORDNET TO THE HOMERIC DEPENDENCY LEXICON

Zanchi C., **Luraghi L.**, University of Pavia, Pavia PV, Italy; **Biagetti E.**, University of Pavia / Bergamo, Pavia PV, Italy

The Ancient Greek WordNet is a new resource that is being developed at the Universities of Pavia and Exeter, based on the Princeton WordNet. The Princeton WordNet provides sentence frames for verb senses, but this type of information is lacking in most WordNets of other languages. In fact, exporting sentence frames from English to other languages is not a trivial task, as sentence frames depend on the syntax of individual languages. In addition, the information provided by the Princeton WordNet is not corpus-based but relies on native speakers' knowledge. This type of information is not available for dead languages, which are by definition corpus languages. In this paper, we show how sentence frames can be extracted from morpho-syntactically parsed corpora by linking an existing depend-ency lexicon of Homeric verbs (HoDeL) to verbs in the Ancient Greek WordNet. Given its features, HoDeL allows automatically extracting all subcategorization frames available for each verb along with information concerning their frequency as well as semantic information regarding the possible arguments occurring in specific frames. In the paper, we show our method to automatically link the two resources and compare some of the resulting sentence frames with the English sentence frames in the Princeton WordNet.

RUSSIAN PREDICATIVES AND THE ONTOLOGY OF STATES

Zimmerling A. V., Institute of linguistics, Russian Academy of Science/ Puskhin state Russian language Institute, Moscow, Russia

Basing on the frequency dictionary of Russian predicatives, I measure the volume of the lexical class of non-agreeing predicatives licensing the productive dative-predicative sentence pattern, where the predicative assigns da-tive case to its animate subject. The tested vocabulary includes 422 elements. Their frequency rates are derived from the main corpus of RNC using an approximation — the number of hits in the context "predicative + dative subject in 1Sg" in the window {-1; 1}. I argue that the Russian dative-predicative construction has an invariant meaning of internal state, i.e. spaciotemporal stative situation with a priority argument. However, most predicatives licensing dative-predicative structures in Russian also express external states, i.e. spaciotemporal stative situations without a priority argument, if used without overt referential dative subject. This can be proved both for words denoting physical sensations, cf. X-y kholodno 'X is cold' vs kholodno 'It is cold' and for some words denoting affections, cf. tosklivo 'dreary', 'sad', X-y tosklivo 'X feels sad' vs zdes' tosklivo 'It's dreary here'. The shift from internal state to external state is licensed in Russian. If a lexical item has regular uses in the dative-predicative structure, it generally can express the meaning of external state outside this structure. The reverse if false: if a lexical item has regular uses as an external state, cf. vetreno 'windy', pyl'no 'dusty', it only can have infrequent side uses with a dative subject. This asymmetry is confirmed by the corpus data. I check an additional list of words with the meaning of external state, measure their frequency rate in the context "predicative + dative subject in 1Sg" in the window {-1; 1} and compare them to standard dative predicatives.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог» (2021)

Выпуск 20

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**

Уч.-изд. л. 63,7. Заказ № 1327

Издательский центр
Российского государственного
гуманитарного университета
125047, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06