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, приуроученного к конференции Диалог 2021. Соревнование RuSimpleSentEval посвящено упрощению предложений на русском языке. Специально для этого сорevания 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.