Creating an Aligned Russian Text Simplification Dataset from Language Learner Data

Anna Dmitrieva
Department of Digital Humanities
University of Helsinki
Helsinki, Finland;
HSE University
Moscow, Russia
annadmitrieva252@gmail.com

Jörg Tiedemann
Department of Digital Humanities
University of Helsinki
Helsinki, Finland
jorg.tiedemann@helsinki.fi

Abstract

Parallel language corpora where regular texts are aligned with their simplified versions can be used in both natural language processing and theoretical linguistic studies. They are essential for the task of automatic text simplification, but can also provide valuable insights into the characteristics that make texts more accessible and reveal strategies that human experts use to simplify texts. Today, there exist a few parallel datasets for English and Simple English, but many other languages lack such data. In this paper we describe our work on creating an aligned Russian-Simple Russian dataset composed of Russian literature texts adapted for learners of Russian as a foreign language. This will be the first parallel dataset in this domain, and one of the first Simple Russian datasets in general.

1 Introduction

Automatic text simplification is a task of natural language processing aimed at making texts more readable and accessible to a broader audience. Nowadays, this task is most often viewed as a neural machine translation problem (Xu et al., 2015). Much like cross-lingual machine translation, intralingual neural text simplification requires a substantial amount of data in order to be able to train and test appropriate models.

Currently there are multiple datasets to choose from for English text simplification, such as Simple PPDB (Pavlick and Callison-Burch, 2016), Simple Wiki (Kauchak, 2013) and Newsela (Xu et al., 2015). However, not a lot of simplification-specific datasets are available for other languages. A few exceptions are, for example, the Spanish version of the Newsela dataset and the Alector parallel Simplified French corpus (Gala et al., 2020). For Russian, there are currently very limited options for publicly available simplification corpora. One such example is the dataset for one of the Dialogue Evaluation tasks (RuSimpleSentEval)[1]. However, this dataset consists mainly of texts from the Wiki-Large corpus (Zhang and Lapata, 2017) that were translated into Russian automatically and parallel data from the Russian paraphrase corpus (Gudkov et al., 2020). Only the texts from the development and test set have been simplified by human experts. We will, instead, produce a corpus of Russian texts adapted by human experts for language learners, a unique resource for further work on Russian text simplification.

Creating a new dataset for text simplification poses quite a few challenges. First of all, there needs to be a source of such data. Possible options include simplified Wikipedia articles like Simple English Wikipedia or Vikidia[2], a French website intended for young people with more accessible articles than in Wikipedia, both in terms of language and content (Brouwers et al., 2014). Other possible sources could be texts simplified for children or second language learners, since not many languages actually have a simplified Wikipedia equivalent. There is also an option of simplifying some texts manually with the help of human experts, like it was done in Newsela or Alector, but this is expensive and slow.

Another issue that gathering simplified data poses is the fact that there is no one simple language. Simplified texts can vary in complexity levels and intended target audience: for example, texts adapted for second language learners or children typically become more complex as the reader's age or L2 proficiency increases. The comprehensibility of the reading materials for users of easy language is typically higher than for plain language users (Maaß, 2019). Therefore, it is important to

[1] https://github.com/dialogue-evaluation/RuSimpleSentEval

[2] http://fr.vikidia.org

Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, pages 73–79

remember that it is impossible for one dataset to satisfy the needs of all simplified language users.

In this paper, we will describe the process of creating a Russian-Simple Russian parallel dataset made out of adapted Russian literature texts intended for learners of Russian as a foreign language (RaaFL). The corpus we produced is aligned on the paragraph level.

2 Data sources

The contents of the dataset described here are simplified Russian books and their original versions. The simplified books were adapted by professional RaaFL writers. They have been provided by the Zlatoust publishing house[3] for the purposes of this research.

The books and short stories in our dataset are all works by Russian writers, most of them being classical literature, such as novels by A. S. Pushkin, F. M. Dostoevsky, and A. P. Chekhov. There are also more modern works by writers like B. Akunin and V. Pelevin. Among the books provided for this dataset are 6 collections of novels and 16 separate books.

Most of the books and collections of novels are aimed at readers with an intermediate level of Russian, with some of them being aimed at lower and higher levels. All together, there are 2 books in this dataset dedicated for A2 level, 2 for levels A2-B1, 11 for B1, 1 for B1-B2, 3 for B2, and 3 collections of novels that fall between B1 and C1 levels.

For evaluation and as an additional reference, we have also created a smaller dataset that is made of 300 adapted texts from textbooks for children who are learners of Russian as a native language (1-4 school grades). Those are short texts from exercises that we manually aligned with their original versions.

At the moment, we are negotiating the possibility of giving open access for research purposes to parts of the collection. We aim at publishing a large portion of the parallel data with permissive licenses depending on the agreement we can establish. We will make that part available on GitHub or similar websites.

3 Data collection and preprocessing

The original format of the books and novel collections that were used for this dataset is PDF. Extracting the text from PDF files creates the first challenge in our process. The conversion was done with the help of the Apache Tika software[4] that allowed us to turn PDF documents into accessible XML files. After that, the texts were further cleaned and normalized to remove unnecessary line breaks, diacritics and other noisy symbols with the help of a dedicated Python script. The original texts that were downloaded from open sources were also processed in the same way and all texts were then stored in plain text files for further analysis and subsequent alignment.

While studying the characteristics of simplified texts, we performed word and sentence tokenization and, in certain cases, lemmatization on our dataset. For lemmatization, we used pymystem3[5], a wrapper for Yandex Mystem 3.1, which is a morphological analyzer for Russian (Segalovich, 2003). It is a dictionary-based algorithm, which has been proven to achieve up to 96,43% accuracy on POS tagging (Dereza et al., 2016), and up to 82,08% accuracy on lemmatization (Akhmetov et al., 2020). We also relied on Mystem for word tokenization, considering only words that it recognizes as tokens. Therefore, for, example, digits and punctuation were not considered words (tokens). For sentence tokenization, we used the NLTK sentence tokenizer for Russian which is based on ru_punkt[6]. It is important to note that these instruments were used only for obtaining dataset statistics and not during text alignment or neural text simplification.

4 Characteristics of the simplified texts

First of all, we wanted to see how different the adapted texts are compared to the corresponding original versions. In this study, we focus on automatic measures to provide essential properties of the texts and we leave a deeper analysis of the simplification level to future work. We looked at a number of metrics that are commonly used for determining the complexity of texts for RaaFL learners (Laposhina et al., 2018). The results are listed in Table 1, where we also list some dataset statistics. Some of the scores are weighted: for example, when computing average paragraph length in the whole dataset, average paragraph length in each text is weighted by the number of paragraphs in this text. For Flesch-Kincaid grade level (FKGL) we used a formula with constants that were adapted

[3] http://zlat.spb.ru/

[4] https://tika.apache.org/

[5] https://github.com/nlpub/pymystem3

[6] https://github.com/Mottl/ru_punkt

Metric	Original	Adapted
Words	885167	268409
Unique words	89318	32762
Sentences	69737	29003
Par length / text*	250.45	180.29
Punctuation / sent*	2.40	1.66
Sentences / par*	3.15	3.19
Average TTR	0.42	0.43
Words / par*	39.06	28.70
Word length / text*	5.08	4.89
Average FKGL	6.04	4.49

Table 1: Characteristics of adapted and original texts. Sent = sentence, par = paragraph. * - weighted average.

for Russian texts[7]. In this formula, the number of vowels in the word is considered to be the number of syllables.

It can be seen that original and adapted texts vary in length with the originals having more words on average. However, the number of sentences per paragraph stays almost the same. Average word length and sentence complexity (approximated by the average number of punctuation characters per sentence) are slightly bigger in the original versions. Adapted texts also have higher readability according to the Flesch-Kincaid grade level. Nevertheless, the degree of lexical variation does not seem to change drastically according to the type-token ratio (TTR).

We also calculated the number of words in the texts that are supposed to be known to the reader on a certain level of language acquisition. There are established lexical minimums with appropriate vocabulary to every level of RaaFL (see Andryushina and Kozlova (2012), Andryushina and Kozlova (2011), Andryushina et al. (2018), Andryushina et al. (2019a), and Andryushina et al. (2019b)). These lists consist of lemmatized words, so the texts were also lemmatized before counting the percentages of known words. We also considered people's names (that is, first names, surnames and patronymics) and location names to be known to the reader, since in literature the audience quickly becomes acquainted with characters and the places where the plot is set up. We identified those words with the help of pymystem3. For example, in Table 2 A1 vocab is the percentage of words in the text that are supposed to be in an A1 level speaker of Russian's vocabulary. Because most of our texts have more experienced readers as their target audience, the amount of known vocabulary for A1 and A2 levels is lower on average.

[7] https://github.com/infoculture/plainrussian

Level	Original	Adapted
A1	57.48	62.97
A2	65.62	71.81
B1	74.3	80.69
B2	81.82	87.24
C1	89.67	92.97

Table 2: Mean amount of known vocabulary (percentage of words in the lexical minimum).

We can see that even at the C1 level, an average speaker will most likely not recognize all words even in the adapted versions of texts. However, the percentage of known words is higher in the adapted books at all levels.

5 Text alignment

Further analysis and the use of the data for training automatic text simplification tools require an explicit alignment between original and simplified text segments. Many alignment algorithms and tools have been proposed for the creation of parallel corpora mainly in the context of machine translation research.

We used multiple alignment approaches ranging from semi-automatic to fully automatic tools. First of all, we tried InterText (Vondřička, 2014) - an editor for managing parallel texts. It has multiple options for automatic pre-alignment of the text that can then be adjusted using a graphical interface. In our work, we tried Hunalign (Varga et al., 2007). This aligner turned out to have some shortcomings, especially in aligning big texts where one version is significantly longer than the other. This is not a surprise as Hunalign (as most of the other algorithms) is developed for cross-lingual sentence alignment and not for linking texts in the same language but different levels of complexity. Nevertheless, the InterText interface provides a convenient tool that enabled a semi-automatic alignment of a smaller dataset that we can use for the evaluation of fully automatic alignment algorithms later on. Using this method we created a gold standard of 302 aligned segments taken from textbook exercises.

We tested two methods to align the entire collection of books and novels, Bleualign (Sennrich and Volk, 2010) and CATS-Align: a tool for customised

Data	Words	Vocab	Sents
Bleualign orig	345779	52010	32593
Bleualign adapt	254000	31831	30553
CATS orig	589715	55380	50673
CATS adapt	268396	32746	32382

Table 3: Alignment statistics. Vocab = vocabulary, sents = sentences, orig = original versions, adapt = adapted versions.

alignment of text simplification corpora[8] (Stajner et al., 2017). Bleualign[9] is designed for sentence-to-sentence alignment of parallel texts used for training machine translation models. The alignment is performed based on the modified BLEU score between source sentences translated into the target language and the original target language sentences. The principle of matching sentences in the same language suits our needs very well and in the case of simplified language alignment we can, therefore, skip the translation step.

CATS-Align is a tool developed specifically for aligning simplification datasets, particularly for Newsela. It offers multiple options for similarity strategies, alignment levels, alignment strategies and other parameters. Our choices were the character trigram similarity strategy that uses the log TF-IDF weighting and compares vectors with the cosine similarity, paragraph alignment level, and closest similarity alignment strategy.

As mentioned above, we chose paragraph alignment for our texts. We believe that text simplification should not consider isolated sentences only but also look at additional context. We are interested in cases where larger portions of the text are removed, sentences are merged or split or even some information is expanded. Also, creating sentence alignments out of an existing paragraph alignment is more effective and reliable than just aligning sentences from texts (Stajner et al., 2017). Hence, we can easily extend the dataset with additional linking between individual sentences later on.

Using Bleualign on our data resulted in 7452 aligned paragraphs, and 9352 with CATS-Align. Other statistics, such as total numbers of word tokens, sentences, and the number of unique words (vocabulary) can be found in Table 3.

Among the CATS alignments, multiple original paragraphs end up being aligned with more than one different adapted paragraph, and vice versa. Therefore, we only have 7671 unique original paragraphs and 9202 unique adapted paragraphs. However, sometimes having multiple simplified options for one paragraph can be helpful. And, if needed, the cosine similarity scores that CATS provides for each pair of paragraphs can help choose between multiple versions of one paragraph alignment (that is, the pair with the highest score can be chosen from multiple pairs with good scores). With Bleualign, non-unique alignments happen only with paragraphs that are repeated in the original and adapted texts - for example, the word "Pause." in plays.

Aligner	Strict F1	Lax F1
Bleualign	0.90	0.90
CATS	0.98	0.98

Table 4: Alignment evaluation.

In order to evaluate the alignment quality, we employed strict and relaxed (lax) F1 scores, as in Sennrich and Volk (2010). We considered precision to be the number of correct alignments per the number of proposed alignments, and recall to be the number of correct alignments per the number of alignments in the reference corpus. However, for strict F1 score we only considered an alignment correct if it matched the reference alignment exactly, and for lax F1 score an alignment was considered correct if both sides overlapped with the original paragraph pair. We used the small, hand-aligned dataset for alignment evaluation. The results are presented in Table 4.

As can be seen, although both aligners showed good results, CATS-Align performed better on hand-aligned data. When investigating the errors of Bleualign, we found out that most of the erroneous alignments happen because the aligner takes some additional sentences from the next paragraphs. This seems to only happen with adapted paragraphs. One possible reason is the size mismatch between original and adapted texts. However, when aligning longer books and novels, this might be appropriate, since sometimes longer paragraphs from the original texts are split into multiple paragraphs during adaptation, so it can be good to grab additional adapted paragraphs during alignment to match a longer original one.

To study the impact of different aligners on

[8] https://github.com/neosyon/SimpTextAlign

[9] https://github.com/rsennrich/Bleualign

downstream applications, we then decided to compare the performance on simplification models when using different alignment tools (see below).

6 Neural text simplification

In order to evaluate the impact of alignment, we built neural text simplification models from the collected training data. We chose the architecture based on (Nisioi et al., 2017), which has proven to perform well on English-Simple English data. This architecture is openly available online[10] and has some modifications that have further improved its performance (Cooper and Shardlow, 2020). We used OpenNMT-py (Klein et al., 2017) to build our models.

Similar to the original paper, we used an architecture with 2 LSTM layers with hidden states of 500 and 500 hidden units. The dropout probability was set to 0.3. SGD was used as an optimizer, and global attention and input feeding were employed. We also employed the default learning rate of 1.0 with the decay of 0.7. The vocabulary size was set to 50000, which also happen to suit our needs, since, as can be seen from Table 3, the vocabularies of the original paragraphs only slightly exceed this number, and the adapted vocabularies are even smaller.

For evaluation, we used the BLEU and SARI metrics from the EASSE library (Alva-Manchego et al., 2019) and the aforementioned Flesch-Kincaid Grade Level score with constants optimized for Russian. We evaluated our models on both test sets that consisted of data from the same books and novels but which the models have not seen during training, and the small hand-crafted dataset used for alignment evaluation, which consists predominantly of excerpts from children's literature. The test and development set sizes for the dataset aligned with Bleualign were both 1000 paragraphs, and for the data aligned with CATS - 1500 paragraphs. The best results for each system on larger test sets can be seen in Table 5.

Because there are no state-of-the-art Russian automatic text simplification systems yet, we don't have anything to compare these scores to. In the original paper, the highest BLEU score was 87.50, and the highest SARI was 38.59. FKGL scores were not reported. However, this system was built for a different language and trained on a bigger corpus with sentence-level alignment.

[10] https://github.com/senisioi/NeuralTextSimplification

Aligner	BLEU	SARI	FKGL
Bleualign	21.68	42.97	2.82
CATS	14.69	40.94	2.82

Table 5: Simplification evaluation – larger automatically aligned test sets.

Aligner	BLEU	SARI	FKGL
Bleualign	10.86	35.53	3.33
CATS	7.51	33.84	2.72

Table 6: Simplification evaluation – small manually aligned test set.

As can be seen, models trained on the data aligned with Bleualign tend to have higher BLEU scores. However, the SARI scores for both datasets are close, and the outputs are equally readable according to FKGL.

As for the performance on completely unseen out-of-domain data, both kinds of models did not show great results. The best results for each system on the small test set are presented in Table 6. It can be seen that, although the BLEU scores halve compared to the bigger test set, the SARI and FKGL scores do not show such a rapid decline. In fact, the change in readability is very little in comparison to Table 5.

Clearly, the models need further work to be able to perform well on new domains. However, the performance on larger test sets indicates that our data can be used for neural text simplification, perhaps not only as a single dataset, but also as material for fine-tuning existing models.

7 Conclusions

We have described the process of creating a parallel Russian-Simple Russian dataset made of literature texts adapted by human experts. This dataset is still a work in progress. For now, we have described some linguistic properties of our texts and established differences between various alignment strategies. We also provide initial tests on the use of the data for neural text simplification. In the future we plan to add sentence alignments and further improvements to the automatic text simplification derived from the data. Data and models will be publicly released if copyrights permit.

References

Iskander Akhmetov, Alexander Krassovitsky, Irina Ualiyeva, Alexander Gelbukh, and Rustam Mussabayev. 2020. An open-source lemmatizer for russian language based on tree regression models. *Research in Computing Science*, 149:147–153.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. 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. Association for Computational Linguistics.

N. P. Andryushina and T. V. Kozlova. 2011. *Lexical minimum for Russian as a foreign language. Base level*. Zlatoust Publishing House, Russia, Saint Petersberg.

N. P. Andryushina and T. V. Kozlova. 2012. *Lexical minimum for Russian as a foreign language. Elementary level*. Zlatoust Publishing House, Russia, Saint Petersberg.

Natalia Andryushina, Irina Afanaseva, Galina Bitekhtina, Liubov Klobukova, and Irina IAtsenko. 2019a. *Lexical minimum for Russian as a foreign language. Second certification level*. Zlatoust Publishing House, Russia, Saint Petersberg.

Natalia Andryushina, Irina Afanaseva, Larisa Dunaeva, Liubov Klobukova, Lidiia Krasilnikova, and Irina IAtsenko. 2019b. *Lexical minimum for Russian as a foreign language. Third certification level*. Zlatoust Publishing House, Russia, Saint Petersberg.

Natalia Andryushina, Galina Bitekhtina, Liubov Klobukova, Larisa Noreiko, and Irina Odintsova. 2018. *Lexical minimum for Russian as a foreign language. First certification level*. Zlatoust Publishing House, Russia, Saint Petersberg.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. Syntactic sentence simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 47–56, Gothenburg, Sweden. Association for Computational Linguistics.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5588–5594, Marseille, France. European Language Resources Association.

Oksana Dereza, Dmitry Kayutenko, and Alyona Fenogenova. 2016. Automatic morpho-logical analysis for russian: A comparative study. In *Proceedings of the International Conference Dialogue 2016. Computational linguistics and intellectual technologies. Student session (online publication)*.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France. European Language Resources Association.

Vadim Gudkov, Olga Mitrofanova, and Elizaveta Filippskikh. 2020. Automatically ranked Russian paraphrase corpus for text generation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, Online. Association for Computational Linguistics.

David Kauchak. 2013. 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. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Antonina Laposhina, Tatiana Veselovskaya, Mariia Lebedeva, and Olga Kupreshchenko. 2018. Automated text readability assessment for Russian second language learners. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 396–406.

Christiane Maaß. 2019. Easy language and beyond: How to maximize the accessibility of communication. pages 1 – 41, Hildesheim.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. 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. Association for Computational Linguistics.

Ellie Pavlick and Chris Callison-Burch. 2016. 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. Association for Computational Linguistics.

Ilya Segalovich. 2003. Fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications)*, pages 73–280, Las Vegas, Nevada, USA.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of AMTA 2010*.

Sanja Stajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. 2017. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 97–102.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Recent Advances in Natural Language Processing IV*.

Pavel Vondřička. 2014. Aligning parallel texts with InterText. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1875–1879, Reykjavik, Iceland. European Language Resources Association (ELRA).

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.