New Dataset and Strong Baselines for the Grammatical Error Correction of Russian

Viet Anh Trinh
Department of Computer Science
CUNY Graduate Center
New York, NY
vtrinh@gradcenter.cuny.edu

Alla Rozovskaya
Department of Computer Science
Queens College (CUNY)
New York, NY
arozovskaya@qc.cuny.edu

Abstract

Motivated by recent advancements in grammatical error correction in English and existing issues in the field, we describe a new resource, an annotated learner corpus of Russian, extracted from the Lang-8 language learning website. This new dataset is benchmarked against two grammatical error correction models that use state-of-the-art neural architectures. Results are provided on the newly-created corpus and are compared against performance on another, existing resource. We also evaluate the contribution of the Lang-8 training data to the grammatical error correction of Russian and perform type-based analysis of the models. The expert annotations are available for research purposes.

1 Introduction

The task of Grammatical Error Correction (GEC) is concerned with correcting various grammatical and usage errors in text. Recently, much progress has been made, especially in English GEC, within the framework of neural machine translation (NMT) approaches. In spite of the progress, the issue of building *robust* models in GEC has been emphasized: Mita et al. (2019) showed that the performance of the models varies significantly across corpora and that single-corpus evaluation may be unreliable. While in English more efforts are being made in this direction, in other languages, due to lack of benchmark corpora and other resources, very little work has been done. More importantly, as learner data in other language is very hard to come by and expensive to annotate, few benchmark corpora exist in other languages.

We develop a benchmark corpus for Russian learner data, by providing expert quality annotations for a subset of the Russian subcorpus of Lang-8, henceforth *RU-Lang8* dataset. Lang-8 (Mizumoto et al., 2012) is a dataset collected from a language learning website and partially corrected by native language volunteers. In Russian GEC, Rozovskaya and Roth (2019) recently released an annotated learner corpus, RULEC. The expert annotations that we provide will allow researchers to use the created corpus as another evaluation benchmark corpus for non-English GEC. As we show, RU-Lang8 is more diverse than RULEC in terms of the first language backgrounds, and the genre of writing. We benchmark two state-of-the-art neural machine translation models on the new corpus: a convolutional neural network (CNN) and a Transformer model.

The paper makes the following contributions: (1) We generate gold annotations for Lang-8 data to create an additional evaluation dataset for Russian GEC, which is more diverse linguistically and contains data of different genre of writing, compared to the existing resource RULEC. We make the resource available for research purposes;[1] (2) We provide benchmark results on this new corpus, using state-of-the-art models that are trained on synthetic data and learner data; (3) We provide an error analysis showing that most of the grammar errors are still challenging for the current systems.

2 Related Work

Progress in English GEC There has been a lot of work on grammatical error correction, but most of the research has been done on English (Rozovskaya and Roth, 2011; Susanto et al., 2014; Yuan and Briscoe, 2016; Hoang et al., 2016; Chollampatt et al., 2016; Junczys-Dowmunt and Grundkiewicz, 2016; Mizumoto and Matsumoto, 2016; Rozovskaya and Roth, 2016; Jianshu et al., 2017; Chollampatt and Ng, 2018; Kaneko et al., 2020).

Recently, state-of-the-art results were obtained using statistical and neural machine translation approaches. The systems are typically trained on a combination of native data with synthetic er-

[1] https://github.com/arozovskaya/RU-Lang8

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4103–4111
August 1–6, 2021. ©2021 Association for Computational Linguistics

rors and naturally-occurring learner data from NUCLE (Dahlmeier et al., 2013) and the English part of the Lang-8 corpus (Mizumoto et al., 2012), even though the latter is only partially corrected and is known to contain a lot of noise.

Motivated by the issue of robustness, a recent shared task on English GEC (Bryant et al., 2019) released new evaluation data, both from learners of English and native speakers. Napoles et al. (2019) further addressed the issue of robustness of GEC models, by proposing novel evaluation metrics, and also released a diverse GEC dataset.

GEC on Other Languages Two most prominent attempts at GEC in other languages include shared tasks on Arabic and Chinese text correction. In Arabic, a large-scale corpus (2M words) was collected and annotated as part of the QALB project (Zaghouani et al., 2014). There have also been three shared tasks on Chinese grammatical error diagnosis (Lee et al., 2016; Rao et al., 2017, 2018). In other languages, attempts at automatic grammar detection and correction have been limited to identifying specific types of misuse (grammar or spelling) (Imamura et al., 2012; Israel et al., 2013; de Ilarraza et al., 2008; Vincze et al., 2014).

The most relevant to us is the work of Rozovskaya and Roth (2019) that made available an annotated corpus of Russian learner essays. The data released in Rozovskaya and Roth (2019) is relatively uniform, as it is all produced by native English speakers, whereas the RU-Lang8 data comes from a diverse set of speakers.

3 The RU-Lang8 Dataset

RU-Lang8 was created using data collected as part of the Lang-8 corpus (Mizumoto et al., 2012). The Lang-8 learner corpus is a dataset compiled from a language learning website.[2] It contains data from learners of a variety of foreign languages and is weakly annotated (partial corrections are provided by volunteers, but these are quite noisy). The Lang-8 corpus consists of pairs of sentence (source, target), where the *source* denotes the original sentence, and the *target* refers to the modified version that may contain partial corrections and volunteer commentaries. While the English subcorpus contains over 30 million tokens, the Russian learner subcorpus is small, containing about 633,000 tokens. We created a subset of 54,000 tokens and had it manually corrected by expert annotators. This newly-created resource with expert annotations is comparable in size to existing GEC datasets, and should be a valuable addition to multi-lingual resources in GEC. The RU-Lang8 corpus differs from RULEC: the latter consists of essays written in a University setting in a controlled environment, while the Lang-8 data was collected online; the majority of texts in RU-Lang8 are short paragraphs or questions posed by learners.

First language (%)	First language (%)
Japanese (37.9)	Portuguese (7.5)
English (14.0)	German (4.2)
Korean (11.8)	Polish (3.0)
Trad. Chinese (8.1)	Spanish (1.0)
Mandarin (7.6)	Mongolian (0.7)

Table 1: 10 most common first language backgrounds for data from Russian learners in the Lang-8 corpus.

Partition	Sentences	Tokens in the source side
Train	43,848	578,383
Dev	1,968	23,138
Test	2,444	31,603

Table 2: Statistics on the Russian data from Lang-8. The development and the test partitions are manually re-annotated. The training partition includes original noisy corrections.

RU-Lang8 Preprocessing From the Lang-8 corpus, we extract all sentence pairs, where the source and the target sentences are in Russian, using the tool which we modified for Russian (Chollampatt and Ng, 2018). The Lang-8 corpus also contains information about the author's first language. Overall, there are 34 first languages in the Russian subcorpus. Table 1 shows the distribution of the most common first languages in the dataset. The most common first language is Japanese (37% of all writers). Other common first languages are English, Korean, Traditional Chinese, and Mandarin.

The extracted sentence pairs are then tokenized using an in-house tokenizer and further cleaned up, by removing sentence pairs where the target side includes corrector's comments. As a result, 51,575 sentence pairs are kept. These sentence pairs are randomized and split up into training, development, and test partitions. The development and test partitions are manually re-annotated, as described below. The sentence pairs from the training partition are not re-annotated but contain the original noisy Lang-8 corrections. The sizes of the subcorpora are shown in Table 2.

[2] https://lang-8.com

Error type	**Example**
Lexical choice	предлагает "proposes" → утверждает "claims"
Extra word (open-class)	был "was" → ∅
Prep. (ins.,del.,repl.)	в "in" → из "from, out of"
Word form	вдохновленным "inspired" → вдохновенной "inspiring"
Noun:case	специалисты "experts" (pl.,nom) → специалистам "experts" (pl.,dat.)
Adj.:case	главная "main" (sg., fem., nom.) → главную "main" (sg., fem., acc.)
Verb:number/gender	живут "live" (3rd person pl.) → живет "lives" (3rd person sg.)
Verb:aspect	чувствовала "was feeling" (past, imperf.) → почувствовала "felt" (past, perf.)
Verb:voice	продолжала "continued" (past, active) → продолжалась "continued" (past, reflexive)

Table 3: **Some common grammatical error types in Russian learner data.** A complete set of errors with examples is shown in Appendix Table 11.

	Repl. (%)	**Ins.** (%)	**Del.** (%)	**Punc.** (%)	**Word order**(%)	**Total**
Dev	71.3	7.9	7.1	12.4	1.4	3,434
Test	74.2	8.4	8.4	7.7	1.3	3,354

Table 4: Statistics on corrections in RU-Lang8.

Correction of the RU-Lang8 Data The development and the test partitions of the RU-Lang8 corpus were manually annotated by a native Russian speaker, with a Master's degree and with prior annotation experience. To estimate the quality of the annotations, a second expert annotator with a similar background and native proficiency was used (see next section). In contrast to Rozovskaya and Roth (2019), where the errors are also tagged with error type at the level of syntax, morphology, lexical usage, and orthography, the annotation of RU-Lang8 is performed at the level of four operations: *Replace*, *Insert*, *Delete*, and *Word Order*. This new annotation framework speeds up the annotation process significantly and allows the annotator to focus on providing the appropriate correction, without having to think also about the linguistic error type. This approach was also used in the annotation of other GEC corpora (Reznicek et al., 2012; Boyd et al., 2014; Mohit et al., 2014).

The annotation was performed with a publicly-available tool used in other annotation efforts (Rozovskaya and Roth, 2010). Table 4 shows the distribution of errors in terms of edit operations.

A subset of the test data was also marked with error type, using the error classification schema of RULEC. This was done to allow for a comparison of the error distributions between RULEC and RU-Lang8. Table 3 shows examples of some common Russian learner errors in the RU-Lang8 dataset. The complete set of error categories is shown in Appendix Table 11. The error distributions in the two corpora are shown in Appendix Table 12. Because RULEC contains data from two groups of learners – foreign language learners of Russian and heritage speakers[3] – we show statistics for each RULEC group separately. The distribution of errors in RU-Lang8 is very similar to that of the foreign group in the RULEC corpus, even though in RULEC the learners come from the English-speaking language background, while in RU-Lang8, there is a lot of diversity with respect to the first language background. As for error rates,[4] the RU-Lang8 data has significantly higher error rates than both foreign and heritage parts of RULEC (Table 5). We attribute this to the overall higher proficiency level of the RULEC corpus writers. Finally, note that, as shown in Table 5, the error rates in development and test partitions of RU-Lang8 vary substantially: the error rates are 15.6% and 11.3% in the development and test sets, respectively. Since the sentences were selected uniformly at random, we do not have an explanation for the reason why the error distribution is different in the two subsets. We do believe that the varying error distributions might be useful, as they would reflect realistic scenarios where the test data may not have exactly the same distribution as the development/training data.

Inter-Annotator Agreement Computing inter-rater agreement in grammatical error correction

[3] The heritage group in RULEC includes native Russian speakers who grew up in the United States; these speakers have a different error distribution from that of foreign learners of Russian, with the majority of errors being of type punctuation and spelling.

[4] Error rate denotes the percentage of the tokens that have been modified by the annotator.

Corpus	Total words	Incorr. words	Error rate (%)
RULEC (Foreign)	164,071	11,343	6.9
RULEC (Herit.)	42,187	1,705	4.0
RU-Lang8 (dev)	23,138	3,605	15.6
RU-Lang8 (test)	31,603	3,558	11.3

Table 5: Error rates in RULEC (foreign and heritage speakers shown separately) and in RU-Lang8. *Error rates* throughout the paper refer to the percentage of tokens that have been modified.

Second pass	Error rate (%)	Judged correct (%)
Annotator A	1.55	90.0
Annotator B	1.02	97.4

Table 6: Inter-annotator agreement. *Error rates* based on the corrections on the second pass. *Judged correct* denotes the percentage of sentences in the agreement set that the second rater did not change.

is not trivial, as the space of possible corrections for a sentence is extremely large (Choshen and Abend, 2018; Bryant and Ng, 2015; Rozovskaya and Roth, 2021). To estimate the quality of the annotation, we have a second annotator independently re-annotate a subset of the data, 120 sentences. We compute inter-annotator agreement in two ways. First, we follow the method used for RULEC (Rozovskaya and Roth, 2019) where the texts corrected by one annotator were given to the second annotator. Agreement is computed as the percentage of sentences that did not have additional corrections on the second pass, as our goal is to make the sentence well-formed. 120 sentences from each annotator were given to the other annotator for the second pass. Table 6 shows that the error rate of the sentences corrected by annotator A (original annotator) on the second pass was 1.55%, with 90% of the sentences remaining unchanged. The sentences corrected by annotator B (second annotator) on the second pass had an error rate of less than 1.02%, and over 97% of the sentences did not have additional corrections. These agreement numbers are comparable to and even slightly higher than in RULEC (68.5% and 91% of unchanged sentences). The error rates are also in the same ballpark (0.67%-2.4% for RULEC).

We also measure agreement by treating reference corrections made by one annotator as gold and scoring the second annotator against them. .Results are shown in Table 7. The scores of 66.7 and 69.9 are higher than those reported on English CoNLL-14 (score of 45.91, Bryant and Ng (2015)).

Gold annotator	P	R	$F_{0.5}$
Annotator A as gold	66.0	69.6	66.7
Annotator B as gold	72.2	61.9	69.9

Table 7: Scoring one annotator against another.

4 Experiments

GEC Models We benchmark two state-of-the-art neural machine translation models: a Convolutional Encoder-Decoder Neural Network model (*CNN*) (Chollampatt and Ng, 2018) and a *Transformer* model (Naplava and Straka, 2019).[5] The Transformer model achieved the highest F-score on RULEC. Both models make use of the RULEC training and dev data (about 5K sentences) and native data with synthetic errors. The CNN model is trained jointly on RULEC and synthetic data, while the Transformer is pre-trained on synthetic data and finetuned on RULEC data.

The models make use of the two approaches of generating synthetic data that showed state-of-the-art performance on English GEC. The Transformer model makes use of the Aspell confusion sets method (Grundkiewicz et al., 2019) to generate synthetic errors in native data. In line with Grundkiewicz and Junczys-Dowmunt (2019), the word error rate used is that of 15%, where on average 15% of the tokens are perturbed, and, on top of these Aspell-generated confusions, characters are perturbed in 10% of the word tokens to account for spelling mistakes. We refer the reader to Naplava and Straka (2019) for details about the model implementation. The CNN models are trained on synthetic data that use part-of-speech (POS)-based confusion sets (Choe et al., 2019) (which we re-implemented for Russian). In all cases, results of single models are compared.

The models are trained on similar amounts of synthetic data: the CNN model uses 13 million sentences from the Web (Borisov and Galinskaya, 2014), while the Transformer model uses 10 million sentences from the Web (Bojar et al., 2017). Although we did not directly compare the two data sources, we assume the native data used by both models is of similar quality as the data comes from the Web in both cases. Since the synthetic data used in the CNN model is not focused on spelling errors, we run an off-the-shelf spellchecker for Russian,

[5] We thank Jakub Náplava for kindly agreeing to run the Transformer model on the RU-Lang8 data.

Model	**RULEC**			**RU-Lang8**		
	P	**R**	$\boldsymbol{F_{0.5}}$	**P**	**R**	$\boldsymbol{F_{0.5}}$
CNN	55.8	26.6	45.7	57.9	26.8	**47.0**
Transf.	59.1	26.1	**47.2**	-	-	-
Transf. (+dev)	63.3	27.5	***50.2***	55.3	28.5	46.5

Table 8: Results on the test of the CNN and Transformer models. Best result for each test set is in bold. Transformer+dev shows performance of Transformer model finetuned on both training and dev RULEC data.

Training data	**RULEC**			**RU-Lang8**		
	P	**R**	$\boldsymbol{F_{0.5}}$	**P**	**R**	$\boldsymbol{F_{0.5}}$
RU+synth.	55.8	26.6	45.7	57.9	26.8	47.0
RU+L8+synth.	54.5	27.6	**45.6**	58.8	29.6	**49.1**

Table 9: Results on RULEC and RU-Lang8 of the CNN models trained jointly on synthetic and learner data. RU stands for RULEC, and L8 stands for RU-Lang8.

developed following Flor (2012) for English.

Evaluation The models are evaluated on the test partitions of RULEC and RU-Lang8. Comparing system output against human-generated reference is a standard practice in GEC. Several measures have been proposed, such as M^2 scorer (Dahlmeier and Ng, 2012), GLEU (Napoles et al., 2015), ERRANT (Bryant and Ng, 2015), and I-measure (Felice and Briscoe, 2015). M^2 computes precision, recall, and F-score and has been widely used in evaluating GEC systems, and we use it here to compare with previous results on RULEC. M^2 has been used with different beta parameter values, the default is $beta = 0.5$, weighing precision twice as high as recall, which we use here.

Overall Performance We show results for the two models in Table 8.[6] The Transformer finetuned on RULEC train data outperforms the CNN on RULEC *by 1.5 points*, and by almost 5 points when finetuned on the union of training and dev RULEC data. However, on RU-Lang8, the CNN model outperforms the transformer slightly.

Contribution of Lang8 Training Data This is shown in Table 9. Interestingly, performance on RULEC does not improve, while performance on RU-Lang8 improves by 2 points. The latter is not surprising since the data comes from the same domain, however, it is surprising that there is no effect on the RULEC corpus.

Performance Analysis by Error Type We also evaluate the best models shown in Table 8 per error type. Evaluating the precision requires classifying the edits proposed by the system, however, recall can be computed, using the types of the gold edits. In Table 10 in the Appendix, we show the recall on the most common error types. The type-based performance analysis reveals which errors are more challenging for the systems. The highest recall by far (ranging between 52.0% and 70.5%) on both datasets is achieved on spelling errors by both models. The Transformer also achieves a recall of 60% on verb agreement errors on both datasets. The following error categories are more challenging for both models: noun case, preposition, adjective case errors, and punctuation. Finally, the most challenging errors are lexical choice errors, where the recall is below 5% for both models and on both datasets. This supports the observation that current models perform best on spelling errors and currently struggle with other phenomena, which is further exacerbated by the morphological complexity of Russian: the performance on Russian falls behind that on English and German (Grundkiewicz and Junczys-Dowmunt, 2019). Our error-type-based analysis is also in line with the findings in the study on two English corpora, as well as RULEC and RU-Lang8 Rozovskaya and Roth (2021). Specifically, while lexical errors are some of the most common learner mistakes, only a small fraction of system edits are of lexical type.

5 Conclusion

This paper presents an annotated Russian learner corpus based on data from the Lang-8 website. The dataset is more diverse than the existing resource RULEC from the point of view of the first language backgrounds, and also differs in the genre of writing. We benchmark two state-of-the-art models that are trained on learner data and synthetic data, using two competitive noisification techniques.

We believe that the RU-Lang8 corpus with expert annotations is a valuable contribution to the GEC field, where a lot of progress has been made in English, due to a large number of resources and benchmark corpora, but where very few works focus on non-English GEC.

Acknowledgments

We are grateful to Jakub Náplava for the help running his system on the RU-Lang8 data. We thank the anonymous reviewers for their comments.

[6] The last row shows the result of finetuning the Transformer with both train and dev RULEC data.

References

O. Bojar, C. Rajen, Federmann C, Y. Graham, B. Haddow M. Huck, P. Koehn, Q. Liu, V. Logacheva, and C. Monz. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Second Conference onMachine Translation*.

A. Borisov and I. Galinskaya. 2014. Yandex school of data analysis russian-english machine translation system for WMT14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schone, B. Štindlová, and C. Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *LREC*.

C. Bryant, M. Felice, Ø. Andersen, and T. Briscoe. 2019. The BEA-19 shared task on grammatical error correction. In *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA-19)*.

C. Bryant and H. T. Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *ACL*.

Y. J. Choe, J. Ham, K. Park, and Y. Yoon. 2019. A neural grammatical error correction system built on better pre-training and sequential transfer learning . In *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA-19)*.

S. Chollampatt and H.-T. Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI*. Association for the Advancement of Artificial Intelligence.

S. Chollampatt, K. Taghipour, and H. T. Ng. 2016. Neural network translation models for grammatical error correction. In *IJCAI*.

L. Choshen and O. Abend. 2018. Automatic metric validation for grammatical error correction. In *ACL*.

D. Dahlmeier and H. T. Ng. 2012. A beam-search decoder for grammatical error correction. In *Proceedings of EMNLP-CoNLL*.

D. Dahlmeier, H. T. Ng, and S. M. Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

M. Felice and T. Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *NAACL-HLT*.

M. Flor. 2012. Four types of context for automatic spelling correction. *Traitement Automatique des Langues (TAL). (Special Issue: Managing noise in the signal: error handling in natural language processing)*, 3(53):61–99.

R. Grundkiewicz and M. Junczys-Dowmunt. 2019. Minimally-augmented grammatical error correction. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT*.

R. Grundkiewicz, M. Junczys-Dowmunt, and K. Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA-19)*.

D.-T. Hoang, S. Chollampatt, and H. T. Ng. 2016. Exploiting n-best hypotheses to improve an SMT approach to grammatical error correction. In *IJCAI*.

A. Díaz de Ilarraza, K. Gojenola, and M. Oronoz. 2008. Detecting erroneous uses of complex postpositions in an agglutinative language. In *Proceedings of COLING*.

K. Imamura, K. Saito, K. Sadamitsu, and H. Nishikawa. 2012. Grammar error correction using pseudo-error sentences and domain adaptation. In *ACL*.

R. Israel, M. Dickinson, and S.-H. Lee. 2013. Detecting and correcting learner korean particle omission errors. In *IJCNLP*. Association for Computational Linguistics.

J. Jianshu, Q. Wang, K. Toutanova, Y. Gong, S. Truong, and Jianfeng J. Gao. 2017. A nested attention neural hybrid model for grammatical error correction. In *ACL*.

M. Junczys-Dowmunt and R. Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *EMNLP*.

M. Kaneko, M. Mita, S. Kiyono, J. Suzuki, and K. Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *ACL*.

L.-H. Lee, G. Rao, L.-C. Yu, E. Xun, B. Zhang, and L.-P. Chang. 2016. Overview of NLP-TEA 2016 shared task for chinese grammatical error diagnosis. In *3rd Workshop on Natural Language Processing Techniques for Educational Applications*. Association for Computational Linguistics.

M. Mita, T. Mizumoto, M. Kaneko, R. Nagata, and K. Inui. 2019. Cross-corpora evaluation and analysis of grammatical error correction models – is single-corpus evaluation enough? In *NAACL*.

T. Mizumoto, Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of esl writings. In *COLING*.

T. Mizumoto and Y. Matsumoto. 2016. Discriminative reranking for grammatical error correction with statistical machine translation. In *NAACL*.

B. Mohit, A. Rozovskaya, N. Habash, W. Zaghouani, and O. Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP Workshop on Arabic Natural Language Processing (ANLP)*.

J. Naplava and M. Straka. 2019. Grammatical error correction in low-resource scenarios. In *W-NUT Workshop*.

C. Napoles, M. Nădejde, and J. Tetreault. 2019. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. In *Transactions of ACL*.

C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault. 2015. Ground truth for grammatical error correction metrics. In *ACL*.

G. Rao, Q. Gong, B. Zhang, and E. Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proceedings of the Fifth Workshop on Natural Language Processing Techniques for Educational Applications*.

G. Rao, B. Zhang, and E. Xun. 2017. IJCNLP-2017 task 1: Chinese grammatical error diagnosis. In *IJCNLP*. Association for Computational Linguistics.

M. Reznicek, A. Ludeling, and C. Krummes. 2012. Das falkohandbuch. korpusaufbau und annotationen version 2.0.

A. Rozovskaya and D. Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*.

A. Rozovskaya and D. Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of ACL*.

A. Rozovskaya and D. Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *ACL*.

A. Rozovskaya and D. Roth. 2019. Grammar error correction in morphologically-rich languages: The case of russian. In *Transactions of ACL*.

A. Rozovskaya and D. Roth. 2021. How good (really) are grammatical error correction systems? In *EACL*.

R. H. Susanto, P. Phandi, and H. T. Ng. 2014. System combination for grammatical error correction. In *EMNLP*.

V. Vincze, J. Zsibrita, P. Durst, and M. Katalin Szabó. 2014. Automatic error detection concerning the definite and indefinite conjugation in the hunlearner corpus. In *Proceedings of LREC*.

Z. Yuan and T. Briscoe. 2016. Grammatical error correction using neural machine translation. In *NAACL*.

W. Zaghouani, B. Mohit, N. Habash, O. Obeid, N. Tomeh, A. Rozovskaya, N. Farra, S. Alkuhlani, and K. Oflazer. 2014. Large scale arabic error annotation: Guidelines and framework. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

A Additional Results

Error type	**RULEC**		**RU-Lang8**	
	CNN	**Trans.**	**CNN**	**Trans**
Punc.	**24.4**	7.3	**15.4**	10.3
Noun:case	**38.4**	36.7	**40.5**	35.1
Spelling	62.0	**66.7**	52.0	**70.5**
Lex. choice	1.0	3.4	0.5	2.6
Prep.	30.1	29.6	10.9	**17.4**
Adj.:case	20.5	22.0	22.5	27.5
Verb:agr.	27.1	**59.4**	42.4	**60.6**

Table 10: Recall by error type of the model trained on native data combined with learner data. Best result for each error type and dataset is in bold.

Error type	**Example**
Spelling	моркое → морское “marine”
Lexical choice (single-token)	предлагает “proposes” → утверждает “claims”
Replace (multi-token)	грозит “threatens” → создает угрозу “creates a threat”
Punc.	∅ →,
Extra word (open-class)	был “was” → ∅
Missing word (open-class)	∅ → для того “with the purpose of”
Prep. (ins.,del.,repl.)	в “in” → из “from, out of”
Conjunction	и “and” → ∅
Word form	вдохновленным “inspired” → вдохновенной “inspiring”
Noun:case	специалисты “experts” (pl.,nom) → специалистам “experts” (pl.,dat.)
Noun:number	пола“gender” (sg.,gen.) → полов “gender” (pl.,gen.)
Adj.:case	главная “main” (sg., fem., nom.) → главную “main” (sg., fem., acc.)
Adj.:number	дальнейшие “future” (pl.,nom.) → дальнейшее “future” (sg.,nom.))
Adj.:gender	которое “which” (sg.,neutral) → которая “which” (sg. fem.)
Verb:number/gender	живут “live” (3rd person pl.) → живет “lives” (3rd person sg.)
Verb:aspect	чувствовала “was feeling” (past, imperf.) → почувствовала “felt” (past, perf.)
Verb:voice	продолжала “continued” (past, active) → продолжалась “continued” (past, reflexive)
Verb:tense	предлагал “offered” (past tense) → предлагает “offers” (present tense)
Verb:other	соблазнить “to seduce” → соблазнил “seduced”

Table 11: **Grammatical error types in Russian learner data.**

Error type	**Percentage** (%)		
	RULEC-Foreign	**RULEC-Heritage**	**RU-Lang-8**
Spelling	18.6	42.4	19.2
Lexical choice	13.3	5.5	11.6
Punctuation	7.6	22.9	10.3
Missing word	8.9	4.7	7.3
Replace	6.3	2.8	1.7
Extra word	5.7	2.4	6.6
Preposition	3.3	1.5	4.6
Word form	3.1	2.1	1.0
Pronoun	1.0	0.5	1.0
Conjunction	0.8	0.1	1.0
Noun:case	14.0	7.8	12.6
Noun:number	2.5	1.8	0.7
Noun:gender	0.3	0.2	0.7
Adj.:case	3.9	2.1	6.3
Adj.:number	1.0	0.3	-
Adj.:gender	1.4	0.5	-
Verb:number/gender	2.5	1.6	1.7
Verb:aspect	2.0	0.2	3.6
Verb:tense	1.2	0.3	4.6
Verb:voice	1.2	0.2	0.7
Verb:other	0.5	0.1	-

Table 12: Distribution by error type. *Replace* includes phenomena not covered by other categories, e.g., additional morphological phenomena, replacing multi-word expressions, and word order.