Multi-Reference Benchmarks for Russian Grammatical Error Correction

Frank Palma Gomez*
Boston University
frankpalma12@gmail.com

Alla Rozovskaya
City University of New York
arozovskaya@qc.cuny.edu

Abstract

This paper presents multi-reference benchmarks for the Grammatical Error Correction (GEC) of Russian, based on two existing single-reference datasets, for a total of 7,444 learner sentences from a variety of first language backgrounds. Each sentence is corrected independently by two new raters, and their corrections are reviewed by a senior annotator, resulting in a total of three references per sentence. Analysis of the annotations reveals that the new raters tend to make more changes, compared to the original raters, especially at the lexical level. We conduct experiments with two popular GEC approaches and show competitive performance on the original datasets and the new benchmarks. We also compare system scores as evaluated against individual annotators and discuss the effect of using multiple references overall and on specific error types. We find that using the union of the references increases system scores by more than 10 points and decreases the gap between system and human performance, thereby providing a more realistic evaluation of GEC system performance, although the effect is not the same across the error types.[1]

1 Introduction

Grammatical Error Correction (GEC) is the task of detecting and correcting mistakes in text. Most of the GEC research effort has been devoted to correcting mistakes made by English language learners (Jianshu et al., 2017; Chollampatt and Ng, 2018; Grundkiewicz and Junczys-Dowmunt, 2019; Omelianchuk et al., 2020; Awasthi et al., 2019; Li and Shi, 2021; Rozovskaya and Roth, 2016).

The standard approach to evaluating GEC systems is to make use of reference-based measures, where system output is compared against a human-generated reference. A system is rewarded for proposing corrections that are in the reference, and penalized for proposing corrections not found in the reference. A sample sentence with errors from the NUCLE corpus of English language learners (Ng et al., 2013), along with two corrected versions (references 1 and 2) generated independently by two human experts, is depicted in Figure 1 (top part). When more than a single reference is available, system output is evaluated independently against each reference, and the reference that maximizes the score for the sentence is selected.

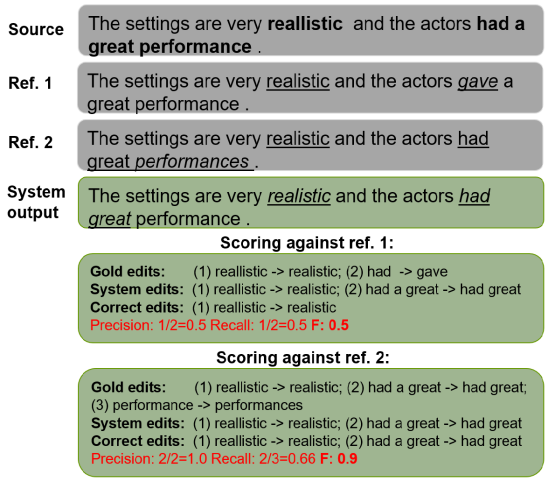

Figure 1: Top (grey): Original (source) sentence with errors, and two corrected versions (ref. 1 and ref. 2). Original erroneous tokens are in bold, and the changes in the references are underlined. Bottom (green): System-produced output and system scores with respect to ref.1 and ref. 2.

There are usually multiple ways of correcting a single sentence, but, since generating human annotations is expensive, many GEC benchmarks contain a single human reference. A large body of work strongly suggests that evaluating against a single reference severely underestimates system performance (Choshen and Abend, 2018b; Bryant et al., 2019; Mita et al., 2019), making it difficult to accurately evaluate GEC models and preventing progress in developing robust GEC systems that

*Work was done while the author was at Queens College, City University of New York.

[1]The annotations are available for research at https://github.com/arozovskaya/RULEC-GEC and https://github.com/arozovskaya/RU-Lang8

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1253–1270
March 17-22, 2024 ©2024 Association for Computational Linguistics

are not overly sensitive to the data from a single annotator. Using more than a single reference has been shown to provide a more realistic evaluation of GEC systems (Bryant and Ng, 2015). This is because having multiple references increases the chance that a valid system correction will match a correction proposed by a human expert. For this reason, system scores tend to increase with the number of annotators used in the gold standard, although at a diminishing rate, suggesting that *using three annotators might be sufficient for a more accurate GEC evaluation* (Bryant and Ng, 2015).

The goal of this work is to contribute to the task of building robust GEC systems by creating benchmark datasets with multiple reference annotations. We consider Russian that has two benchmark GEC datasets – RULEC (Rozovskaya and Roth, 2019) and RU-Lang8 (Trinh and Rozovskaya, 2021) – both annotated with a single reference, and construct two new reference annotations for each sentence.

Analysis of the generated corrections reveals that the new annotators have a high degree of variability in proposing corrections, and, compared to the original raters, propose more changes related to overall fluency.[2] We attribute this to the use of the annotation framework that does not focus on identifying error spans and error types, but, instead, encourages sentence-level re-writing.

The paper makes the following contributions: (1) We enhance two existing GEC benchmarks for Russian with two additional references, for a total of three references per sentence; (2) Using the multi-reference datasets, we benchmark two models that implement state-of-the-art techniques; the models show competitive performance on RULEC and achieve a new state-of-the-art performance on RU-Lang8, with original single-reference benchmarks; Using the union of 3 references increases system scores by 10 F-score points on average against a single annotator; (3) We analyze the effect of multiple references on individual error types and reveal interesting trends that are error specific: using multiple references substantially increases system scores on grammar and orthography, while the scores on lexical errors are affected only slightly; (4) We have the original annotator re-annotate a subset of RULEC in accordance with the new re-writing annotation paradigm and show that the direct re-writing approach negatively affects system scores.

[2]Following Napoles et al. (2017), we consider fluency changes as those that not only correct grammatical errors but *also* make the original text more *native sounding*.

2 Multi-Reference Annotation

Below, we start with an overview of the reference-based evaluation in GEC. Then we describe the annotation paradigms and motivate our choice of the direct-re-writing annotation approach. The rest of the section describes the Russian datasets and the multi-reference annotation.

2.1 Reference-Based Evaluation

The standard approach to evaluating GEC systems is to use reference-based measures, comparing system output to a *reference* generated by a human expert who corrected mistakes in the original *source sentence*.

Aligning the source sentence with a reference, a set of token-level *edits* required to transform the source into its corrected version, is generated. Similarly, the source is aligned with the *system output*. A *gold edit* is an edit between the source and a reference. A *system edit* is an edit between the source and system output. A *correct edit* is an edit in the intersection of gold and system edits. Given the sets of edits, precision, recall, and F-score are computed in a standard way, where precision is the percentage of system edits that are correct, and recall is the percentage of gold edits that are also part of the system edits. Top part of Figure 1 depicts a sample source sentence and two references. The bottom part shows system output, the corresponding edits, precision, recall, and F-scores. Ref. 2 scores would be picked for that sentence, as the F score is higher when ref. 2 is used. Please see Appendix A for an overview of evaluation metrics.[3]

Note that if only ref. 1 was available for the sentence in Figure 1, the resulting score for the sentence would be lower, resulting in performance underestimation.[4] When more than one reference is available, system output is compared independently with each reference, and the reference that maximizes the F-score for that sentence is selected. Having more references increases the chance that valid corrections in system output match those in one of the human-generated references, making that reference close to system output. This would

[3]Please see Choshen and Abend (2018a) for a good survey on the topic. We use M^2 scorer (Dahlmeier and Ng, 2012) that has been widely used in GEC research, with the default value of beta 0.5, i.e. weighting precision twice as high as recall, and refer to the result as $F_{0.5}$.

[4]We do not claim that evaluating against ref. 2 yields an accurate performance estimate, but we show that ref. 2 give a more accurate estimate than ref. 1. It is possible that there exists another reference that would result in an even higher score for that sentence.

allow for a more accurate evaluation of system performance (Bryant and Ng, 2015; Rozovskaya and Roth, 2021; Choshen and Abend, 2018b). Thus, building benchmarks with multiple references is crucial for an accurate evaluation of GEC systems. It has also been suggested that the score differences tend to even out with more than three references, so that *the use of three references is sufficient for providing a more realistic idea of system performance* (Bryant and Ng, 2015). Following these recommendations, there have been efforts to produce English GEC datasets with multiple references (Bryant et al., 2019; Napoles et al., 2017), however, benchmarks in other languages often have a single reference annotation, due to the effort involved in producing human-labeled GEC data, with a few exceptions (Zhang et al., 2022; Naplava et al., 2022; Syvokon and Nahorna, 2021).

2.2 Annotation Paradigm: Direct Re-Writing

There are two main approaches to GEC annotation: In the *error-coded* approach, a human expert corrects all mistakes in the original sentence, and also marks the error span and chooses the error type based on some linguistic taxonomy. This paradigm was adopted in the construction of several English GEC corpora (Yannakoudakis et al., 2011; Rozovskaya and Roth, 2010a) and corpora in other languages, including RULEC (Rozovskaya and Roth, 2019). A relaxed version of this approach consists of having the annotator correct all errors, while also marking the error spans (without having to specify the linguistic error type). This approach was used to annotate several GEC corpora, e.g., the Arabic QALB corpus (Mohit et al., 2014) and the RU-Lang8 dataset used in this work.

Sakaguchi et al. (2016) discuss the challenges of using the error-coded paradigm, specifically, an additional load on the human experts, which may impact annotation quality (Zhang et al., 2022), and the inconsistencies in annotations, when selecting error spans and error types, especially if the error taxonomy is large. In addition, there is the issue of inconsistencies in annotations for multiple datasets for the same language that may follow different error taxonomies (Bryant et al., 2017). To address these issues, Napoles et al. (2017) propose to use "holistic fluency edits to not only correct grammatical errors but also make the original text more fluent or native sounding."[5] This is the approach we adopt in our work, and we refer to it as *direct re-writing*, following Zhang et al. (2022): a human expert is asked to re-write the sentence and make it fully grammatical and fluent, while preserving the original meaning. Note that both annotation paradigms follow the "minimal-edit principle"[6] in that they aim to preserve the original sentence as much as possible. Nevertheless, the direct re-writing paradigm is more conducive to making the output text fluent, since it allows the annotator to focus on providing the appropriate corrections, without having to think about the linguistic error type and edit span boundaries (Napoles et al., 2017; Sakaguchi et al., 2016). The direct re-writing approach has been used in GEC annotation efforts for a variety of languages – English, Chinese, and Ukrainian (Syvokon and Nahorna, 2021; Napoles et al., 2017; Zhang et al., 2022). Figure A.1 in Appendix illustrates the difference between the annotation paradigms.

	Partition	Sents.	Tokens
RULEC	Train (gold)	4,980	83,404
	Dev (gold)	2,500	41,161
	Test (gold)	5,000	81,693
RU-Lang8	Dev (gold)	1,968	23,138
	Test (gold)	2,444	31,603

Table 1: Statistics on the Russian learner datasets. We add two new references to the test partitions of each benchmark.

> Please correct the following sentence to make it sound natural and fluent to a native speaker of Russian. You should fix grammatical mistakes, awkward phrases, spelling errors, etc. following standard written usage conventions, but your edits must be conservative. Please keep the original sentence (words, phrases, and structure) as much as possible.

Table 2: Annotation instructions (based on Napoles et al. (2017)).

2.3 Russian Learner Datasets

Two datasets of Russian learner data are available, that are manually corrected for errors: the RULEC-GEC corpus (Rozovskaya and Roth, 2019) (henceforth RULEC) and and RU-Lang8 (Trinh and Rozovskaya, 2021). Statistics are in Table 1.

RULEC contains essays written by learners of

[5]The use of error-coded paradigm also requires a certain level of linguistic background to be able to choose the appropriate error type.

[6]It is common to instruct the annotators to follow the principle of "minimal edits", that is making the smallest number of edits to render the sentence grammatical and well-formed.

Dataset	Error rate (%)		
	Rater S	Rater A	Rater B
RULEC	6.8	14.7	14.6
RU-Lang8	10.9	18.0	20.5

Table 3: Error rates by dataset and annotator.

Russian studying at the University of Oregon (Alsufieva et al., 2012). RU-Lang8 is a dataset of Russian learner writing collected from the online language learning platform Lang-8 (Mizumoto et al., 2011).[7] While RULEC consists of essays written in a University setting in a controlled environment, the majority of texts in RU-Lang8 are short paragraphs or questions posed by learners. Further, while RULEC data is relatively uniform in that it is all produced by native English speakers, the RU-Lang8 data comes from speakers with a diverse set of first language backgrounds (Mizumoto et al., 2012).

The original RULEC annotation uses the error-coded method. The annotation of RU-Lang8 is performed using a relaxed error-coded method: while errors in RULEC are also tagged with a linguistic error type at the level of syntax, morphology, and lexical usage (a total of 22 categories), the annotation of RU-Lang8 is performed at the level of four operations: Replace, Insert, Delete, and Word Order. In other words, error spans are marked manually in both RULEC and RU-Lang8, but error categories are not specified in RU-Lang8 (see Figure A.1 in Appendix).

2.4 Constructing New Annotations

We have generated two additional references for the *test* partitions of RULEC and RU-Lang8 (5,000 sentences of RULEC and 2,444 sentences of RU-Lang8), as shown in Table 1. The training and development data were not re-annotated. Two new annotators, native speakers of Russian, were recruited to perform the additional annotation. The annotators are college graduates without prior annotation experience. We have also used one of the annotators who participated in the original annotation of RULEC and RU-Lang8 (referred to as senior rater), for quality control and analyses. We denote the senior annotator as *S*, and the new annotators are referred to as *A* and *B*.[8]

The new annotators were given a trial set of 50 sentences each, and the annotation instructions (Table 2). The senior rater both reviewed the resulting annotations and performed *second-pass annotation*, which we also use for inter-annotator agreement (see Section 2.6). We computed the error rates on the second pass (shown in Table 5) and determined that the new annotators are eligible to perform the annotation, based on the error rates being below 10%. Following previous work of similar annotation in Russian and Ukrainian (Syvokon and Nahorna, 2021; Rozovskaya and Roth, 2019), we assume that the error rates below 10% are acceptable. Following the review of the senior annotator, the raters were also given additional instructions.

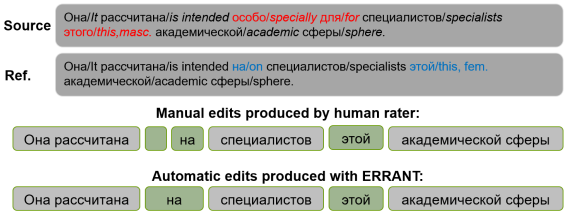

Figure 2: A comparison of manual edit spans and those automatically extracted with ERRANT. Green blocks are edits.

2.5 Statistics on the Annotated Data

Identifying error spans and extracting edits in new annotations Since the error spans are not marked in the direct re-writing annotation paradigm, we apply the ERRANT tool (Bryant et al., 2017) to align the original sentence with each of its new references to get a list of edits. For consistency and direct comparison with the original references, we also apply ERRANT to obtain automatic error spans for the original sets of references in both datasets. The automatic error spans obtained with ERRANT do not always match the manual error spans in the original annotations. This is illustrated in Figure 2: the annotator marked three edits, while ERRANT produced two edits, merging the first two changes (word deletion and preposition replacement) into a single edit. Differences in error spans may result in different F-scores (see Section 3), but the changes are minor (about 5% of sentences have mismatches in error spans).

Computing error rates Table 3 shows the *error rates* (percentage of tokens that have been corrected). The senior rater made significantly fewer changes in each dataset, compared to the new annotators.

Distribution of edits by error type To assign error categories to the edits produced by ERRANT, we apply an error classification tool developed

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

[8] The senior rater previously participated in the RULEC annotation as one of the two raters (each rater corrected a different subset of RULEC sentences), and also performed the annotation of RU-Lang8.

for Russian (Rozovskaya, 2022), that uses a part-of-speech (POS) tagger and a morphological analyzer (Sorokin, 2017) to automatically classify the edits into appropriate linguistic types (the tool follows the error taxonomy adopted in the original RULEC annotation; please see further discussion on the choice of the tool in Appendix B). Note that because the original RULEC annotation follows the error-coded approach, the original references in RULEC come with manually-assigned error categories. For consistency and a fair comparison with the new references, we use ERRANT to get automatic error spans and apply the tool to obtain automatic error categories for the edits in the RULEC original references. Results on the top-12 most frequent error types are shown in Tables 4 and Appendix Table B1 for RULEC and RU-Lang8, respectively. Common Russian learner errors are illustrated in Appendix Table D3.

The new raters make more changes, compared to the senior rater. However, the difference is more pronounced on RULEC (an increase of 80-90%), vs. 50% in RU-Lang8. Further, the largest increase occurs in the open-class lexical categories.[9] We conjecture that the new annotators may have more strict criteria for grammaticality, and that the direct re-writing paradigm is conducive to the annotators making more changes, which would explain the larger increase in RULEC, where the original annotation used the error-coded approach. See also discussion in Section 4.

2.6 Inter-Annotator Agreement

We compute agreement in two ways. First, we follow the method used for computing agreement for English (Rozovskaya and Roth, 2010b), Russian (Rozovskaya and Roth, 2019), and Ukrainian (Syvokon and Nahorna, 2021), where the texts corrected by first annotator were given to the second annotator, and agreement was measured as the error rate relative to the text corrected on the first pass, as our goal is to make the sentence well-formed, without enforcing that errors are corrected in the same way. 100 sentences from each annotator were given to the other annotator for the second pass. Table 5 shows that the error rate of the sentences corrected by the senior rater on the second pass is lower than for the new raters. This is consistent with the earlier finding that rater S is more conservative. Overall, the numbers are higher than those reported for RULEC (0.67%-2.4%) and for Ukrainian (1.2%-2.9%). The highest error rates occur when a new rater re-annotated the texts originally corrected by the senior annotator, which we attribute to the new annotation strategy (see also Section 4).

[9] *Lex. (word)* and *Lex. (phrase)* denote lexical changes that involve single-token and multi-token replacements, respectively.

Error type	**Rel. freq. by rater(%)**			
	S (gold)	**S (auto)**	**A (auto)**	**B (auto)**
Spelling	**20.0**	**21.7**	15.1	15.9
Lex. (word)	11.7	11.8	9.7	10.3
Punc.	11.0	11.3	16.8	15.7
Noun case/num.	-	7.8	5.6	5.1
Prep.	3.3	5.3	4.2	4.0
Lex. (phrase)	4.2	9.6	**19.3**	**19.1**
Noun case	13.2	6.2	4.4	3.8
Insert	9.2	4.0	3.2	4.2
Adj. case	3.7	2.5	2.2	2.0
Verb agr.	2.9	2.5	1.6	1.5
Delete	5.6	1.2	3.5	3.4
Morph.	4.7	1.5	1.2	1.1
Total edits	5,283	5,093	9,741	9,819

Table 4: List of top-12 error types and their distribution in RULEC (by rater). *Gold* refers to the results obtained using manually-assigned edit spans and original gold error labels in RULEC. *Auto* denotes edit spans and error types obtained automatically: error categories are obtained when the automatic error classification tool is applied to the edit spans identified with ERRANT. Most frequent edit type for each rater is in bold.

Second pass	**First pass**		
	Rater S	**Rater A**	**Rater B**
Rater S	-	4.36	2.83
Rater A	7.37	-	3.68
Rater B	7.78	9.56	-

Table 5: Inter-annotator agreement using 100 sentences from RULEC. *Error rates* (the percentage of tokens that have been corrected on the second pass) based on the corrections on the second pass.

Diversity of annotations Our second evaluation measures agreement by treating one annotator as gold and another annotator as system output. This evaluation is expected to reveal *the degree of variability of corrections*. Results are presented in Table 6 for RULEC. Appendix Tables D4 and D5 show detailed results with Precision and Recall for RULEC and RU-Lang8, respectively.

The scores are lower than those reported previously for Russian (66.7 and 69.9, Trinh and Rozovskaya (2021)) but are similar to those reported

Gold annotator	$F_{0.5}$		
	S	A	B
Rater S	100.0	42.7	44.5
Rater A	48.9	100.0	42.2
Rater B	51.3	43.0	100.0

Table 6: Scoring one rater against another (RULEC).

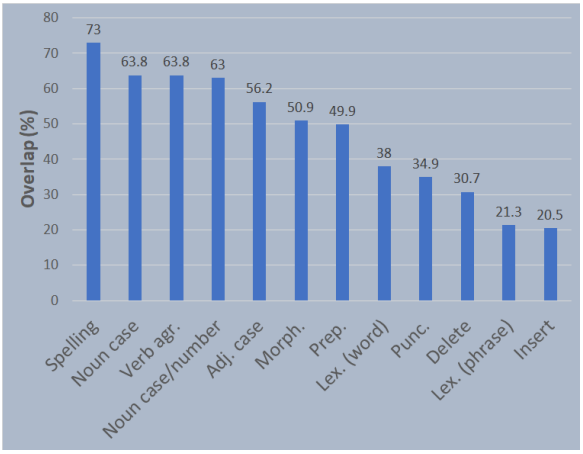

Figure 3: Overlap in edits based on pairwise annotator comparison for RULEC and RU-Lang8 combined (averaged over the 3 annotators). The top-12 error types.

for English (score of 45.91, Bryant and Ng (2015)). The scores indicate that the annotators exhibit a high degree of variability, which we attribute to the new annotation schema.

Overlap in edits by type To compare how often the raters agree on edits of various types, for each error type and rater, we compute the percentage of edits of that type, that is also found in the annotations of another rater. Combined results on the 12 most frequent error types for the two datasets, averaged over each pair of raters are shown in Figure 3. The error categories with the highest agreement are related to spelling and grammar. The errors with the lowest agreement are punctuation and mistakes related to lexical choice. Our results are consistent with previous findings in English suggesting that lexical choice errors have more correction options (Choshen and Abend, 2018b), which would result in lower human agreement on those mistakes.

3 Benchmarking Experiments

We implement two models, evaluate these on the original references and on the multi-reference benchmarks, and investigate how system scores are affected by a choice of a single rater. In Section 4, we perform additional analyses.

Models We have selected GEC models that draw on methods that showed competitive performance in multilingual GEC. Broadly speaking, there are two leading approaches to GEC: sequence-to-sequence (seq2seq) and edit-based (Bryant et al., 2023). We chose the seq2seq framework for both of our models, as it demonstrated superior performance on multiple languages (e.g., Rothe et al. (2021); Palma Gomez et al. (2023)). Indeed, we show below that our results are competitive with previous work on the original RULEC benchmark (Table 7) and outperform state-of-the-art on RU-Lang8 (Table 8).

Regarding the edit-based framework GECToR (Omelianchuk et al., 2020), it was shown to be competitive on English, however, attempts to use it with other languages proved to be less successful (Syvokon and Romanyshyn, 2023). This is because GECToR requires language-specific knowledge to develop rules, while the seq2seq approach does not require special knowledge and can be implemented by researchers not proficient in the target language. A recent survey of GEC research notes the following, as it discusses edit-based approaches such as GECToR (Bryant et al., 2023): "Their main disadvantages, however, are that they generally require human engineering to define the size and scope of the edit label set and that it is more difficult to represent interacting and complex multi-token edits with token-based labels."

Model 1: seq2seq model Our first (smaller) model is a seq2seq Transformer (henceforth, *seq2seq*): the erroneous sentences are treated as the source language, and their corrected counterparts are treated as the target language. Seq2seq approaches have demonstrated strong empirical results in GEC (Chollampatt and Ng, 2018; Yuan and Briscoe, 2016; Grundkiewicz et al., 2019; Grundkiewicz and Junczys-Dowmunt, 2019; Kiyono et al., 2019; Zhao et al., 2019; Jianshu et al., 2017; Yuan and Briscoe, 2016; Katsumata and Komachi, 2019; Xie et al., 2018).

Model 2: mT5 model For our second seq2seq model, we adopt the approach of Rothe et al. (2021) and make use of mT5 (Xue et al., 2021), pre-trained on a subset of Common Crawl, covering 101 languages and composed of about 50 billion documents (Xue et al., 2021). Rothe et al. (2021) finetune mT5 on GEC gold data, although state-of-the-art results are only achieved, when they re-train mT5 with a different objective and use an extremely large model xxl with 13B parameters. We use the original mT5 model of smaller sizes (mT5-base, 580M parameters, and mT5-large, 1.2B parameters). We refer to this model as *mT5*.

Generating synthetic data with morphological transformations Both models are *pre-trained* on native data where the source side has been corrupted with artificial noise. Common data corruption strategies include spelling-based transformations (Grundkiewicz and Junczys-Dowmunt, 2019) and morphology-based transformations (Choe et al., 2019). The latter utilizes morphological variants of the same word, when generating synthetic noise. In this work, we adopt the morphology-based transformations, and generate noise based on the output of a morphological analyzer for Russian (Sorokin et al., 2016). Please see Appendix C for more details about the method.

Experimental Setup The seq2seq models are trained on 15M sentences with synthetic errors, while the mT5 models use 10M synthetic sentences (due to computational constraints). The mT5 models are further finetuned on the RULEC training data. Please see more detail on the experimental setup in Appendix D. Appendix Table D2 summarizes the gold and synthetic data used to train the models.

3.1 Results on the Original References

RULEC Table 7 shows that our results are comparable to or better than previously reported. The top segment of the table lists models trained in this work. The remaining three segments show results of previous work broken down by the amount of gold data used in training and fine-tuning. The special symbols next to each model indicate the type and amount of gold data used (explained in the table caption). Our mT5-base result is comparable to gT5 xxl (13B parameters, last table section); with mT5-large, we obtain a 2-point improvement. Our smaller seq2seq model outperforms all models of similar sizes (section 2 in the table) that also use RULEC training data. Sorokin (2022) uses ruGPT-3 and RoBERTa-large. Their model is comparable to mT5-large, in terms of parameters, but is trained on Russian data, whereas mT5 is multilingual.

RU-Lang8 There are only two results available (Trinh and Rozovskaya, 2021). Comparison is shown in Table 8. Both models show competitive performance, and both the mT5-base and mT5-large model improve over existing state-of-the-art.

3.2 Results on Multi-Reference Benchmarks

Main results In the remainder of the paper, we report results on seq2seq as a smaller model and an mT5-large as a larger model. Tables 9 and 10 show

Model	$F_{0.5}$
This work seq2seq ✧	47.4
This work mT5-base ★	51.0
This work mT5-large ★	53.2
Rothe et al. (2021) gT5 base ★	26.2
Grundkiewicz and Junczys-Dowmunt (2019) ★	34.5
Naplava and Straka (2019) ★	47.2
Flachs et al. (2021) ★	44.7
Katsumata and Komachi (2020) ★	44.4
Naplava and Straka (2019) ✷	50.2
Rothe et al. (2021) gT5 xxl ✦	51.6
Sorokin (2022) 'scorer-only' ✦	53.4
Sorokin (2022) 'combined' ✦	55.0

Table 7: Comparison with previous work for RULEC, using original reference. The top segment shows models trained in this work. The remaining segments show results obtained in previous work, broken down by the amount of gold data used. Extra large models are grouped in the bottom segement. ✧ denotes models that do not use RULEC gold training data; ★ refers to models that use RULEC training data for fine-tuning. ✷ denotes models that use RULEC training and dev data for fine-tuning; ✦ denotes extra large models in terms of parameters and native data used that also use RULEC training data.

Model	$F_{0.5}$
This work seq2seq ✧	47.7
This work mT5-base ★	49.8
This work mT5-large ★	54.5
Trinh and Rozovskaya (2021) ★	47.0
Trinh and Rozovskaya (2021) ✪	49.1

Table 8: Comparison with previous work for RU-Lang8, using original reference. ✧ denotes a model that does not use RULEC training data. ★ refers to models that use RULEC training data for fine-tuning. ✪ denotes a model that uses RULEC training and data from Lang8.

Model	Rater	Performance		
		P	R	$F_{0.5}$
seq2seq	S (gold)	58.8	26.7	47.4
	S (auto)	58.3	27.2	47.4
	A	55.2	13.5	34.1
	B	56.9	13.8	35.1
	S,A,B	69.9	33.8	**57.6**
mT5	S (gold)	64.1	31.7	53.2
	S (auto)	63.7	32.3	53.4
	A	61.9	16.4	39.8
	B	62.1	16.3	39.7
	S,A,B	76.7	39.9	**64.8**

Table 9: Performance on RULEC (test) by individual rater and when using a union of all three. Best result against original reference and the union of 3 in bold.

Model	Rater	Performance		
		P	R	$F_{0.5}$
seq2seq	S	57.6	28.2	47.7
	A	55.9	18.7	40.0
	B	52.9	16.2	36.4
	S,A,B	65.3	35.5	**56.0**
mT5	S	65.1	33.0	54.5
	A	62.0	21.2	44.8
	B	57.5	18.0	40.0
	S,A,B	71.6	40.5	**62.1**

Table 10: Performance on RU-Lang8 (test) by individual rater and when using a union of all three. Best result against original reference and the union of 3 in bold.

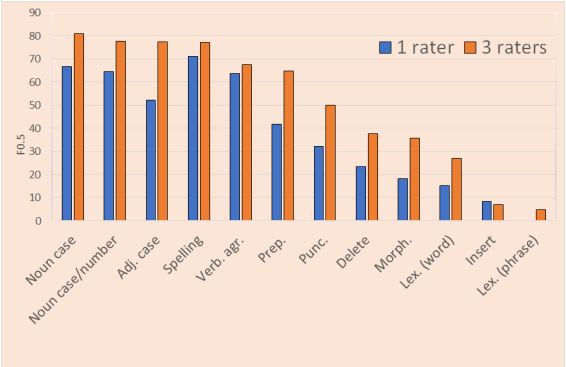

Figure 4: $F_{0.5}$ on RULEC for the top-12 automatic error types. Results when one annotator (senior) is used vs. a union of 3. mT5 model.

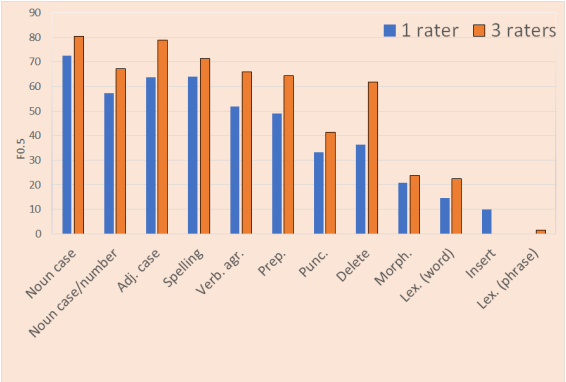

Figure 5: $F_{0.5}$ on RU-Lang8 for the top-12 automatic error types. Results when one annotator (senior) is used vs. a union of 3. mT5 model.

results on RULEC and RU-Lang8, respectively. For RULEC original reference, we show performance using original edit spans (gold) and automatic spans produced with ERRANT (auto).

System scores, when evaluating against a single annotator, vary widely, and scores are much higher, when evaluated against the original reference. Using 3 references increases the scores for both models and benchmarks. Similar behavior has been observed for English (Bryant and Ng, 2015). Also of note is that while the use of multiple references does not change the ranking of the systems, the *gap between the system scores* increases when three references are used, suggesting that multiple references provide more robust results. The results support the view that the use of multiple references helps account for variability in GEC system corrections, thereby providing a more accurate evaluation of GEC system performance (Bryant and Ng, 2015).

4 Further Analysis and Discussion

Effect of using multiple references on individual error types Choshen and Abend (2018b) show that the performance on some error types in English are more severely underestimated than on others. This happens because some errors, such as lexical errors, have a larger set of correction options.

We compare the effect of using multiple references on individual error types. Results are shown in Figure 4 and 5 for mT5-large model on RULEC and RU-Lang8, respectively. Across categories, the best performance is obtained on spelling errors and inflectional grammar errors. The highest gains when 3 references are used are observed on punctuation errors, preposition errors, deletion errors, morphology and adjective case. The smallest gains are on lexical errors and insertions. Comparing system performance on individual error types, higher scores seem to correlate with higher human agreement on those errors (see Fig. 3).

Impact of multiple references and comparison to human performance We evaluate system performance using 1, 2, and 3 references. When using 1, and 2 references, we average the results across different (subsets of) annotators. We perform a similar experiment scoring one human against another or a set of 2 human raters (Figure 6).

The scores increase for both models with the number of references used. Human performance also increases with 2 references, compared to a single one. Note also that the gap between human performance and system is larger for a single-reference evaluation, compared to 2 references used. The gap between the system performance also increases as the number of references used increases from 1 to 3. This suggests that a multi-reference dataset reduces the risk of underestimating performance, and thus provides more robust model evaluation.

Comparison of the annotation paradigms To evaluate the effect of the annotation guidelines on the corrections and on evaluation, as well as to

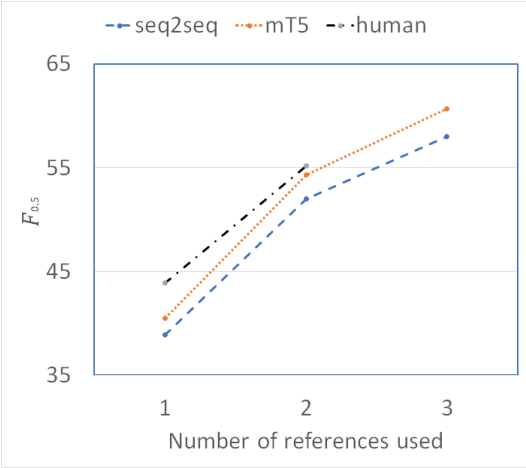

Figure 6: Effect of the number of references on $F_{0.5}$.

Model	Rater/ number of edits	Performance P	R	$F_{0.5}$
seq2seq	S (326) ❑	60.4	26.7	48.2
	S (388) ❍	53.8	20.1	40.3
	A (382) ❍	48.6	18.1	36.3
	B (515) ❍	53.8	15.1	35.6
mT5	S (326) ❑	63.9	28.2	51.0
	S (388) ❍	64.6	24.0	48.2
	A (382) ❍	54.6	20.2	40.7
	B (515) ❍	57.0	15.7	37.4

Table 11: Performance on 200-sentence RULEC subset. ❑ denotes error-coded annotation paradigm, and ❍ stands for direct re-writing.

perform a fair comparison with the new raters, we have the senior annotator re-annotate a subset of the data, using the direct re-writing approach. To this end, the senior annotator is asked to perform a re-annotation of a 200-sentence subset from each dataset, following the new guidelines. Column 2 in Table 11 and Appendix Table 12 show the number of automatic edits in the original and newly corrected files, and compare these to the other 2 raters. The direct re-writing paradigm results in a higher number of edits (20% increase in RULEC and 10% increase in RU-Lang8). The lower increase in RU-Lang8 can be attributed to the relaxed error-coded approach used in the annotation for this dataset (see Section 2.3). It should also be noted that the senior annotator still makes fewer edits compared to the new annotators, which we can attribute to the personal preference of that annotator. In Tables 11 and 12, we show the performance on the 200-sentence subset for each annotator. The general trend is that the new annotation schema results in a lower $F_{0.5}$ score on both datasets, although the difference is more pronounced on RULEC.

Model	Rater/ number of edits	Performance P	R	$F_{0.5}$
seq2seq	S (320) ❑	54.7	23.8	43.4
	S (352) ❍	55.8	21.9	42.6
	A (443) ❍	58.8	18.1	40.5
	B (506) ❍	50.0	13.4	32.4
mT5	S (320) ❑	60.9	26.2	48.2
	S (352) ❍	60.7	24.1	46.6
	A (443) ❍	58.3	17.4	39.6
	B (506) ❍	52.6	14.0	33.9

Table 12: Performance on 200-sentence RU-Lang8 subset. ❑ denotes error-coded annotation paradigm, and ❍ stands for direct re-writing.

Recommendations We believe the findings of this work should be useful for thinking about how to modify evaluation, as well as the training and tuning paradigms in GEC, and we would like to propose several ideas. One recommendation is to develop different strategies for evaluating performance on different error types. Specifically, one finding of the paper is that using 3 references alleviates the problem of performance underestimation on grammar and orthography errors, whereas performance on lexical errors remains unchanged. This suggests that, perhaps, a different approach to evaluating performance on lexical errors should be used, one that considers paraphrasing instead of simple edit matching. Another recommendation and a direction for future work is understanding how training and finetuning on data with a single reference affect system performance, and whether it would be valuable to develop validation and training sets with multiple references.

5 Conclusion

We enriched two Russian GEC benchmarks with additional annotations. We have analyzed and compared the resulting annotations and the original references and shown that the new annotators make more changes compared to the original raters, especially at the lexical level. We computed inter-annotator agreement and human-vs-human performance. We implemented two strong GEC models and evaluated their performance on the new benchmarks. The gap between the model scores increases with the use of more references, whereas gap between human performance and system scores decreases, suggesting an improvement in the robustness of the results when multiple references are used.

Limitations

This work presented annotations for two Russian GEC benchmarks. This resource should help develop robust GEC systems for Russian that provide a more realistic evaluation of system performance and are not overly sensitive to the data from a single annotator. We note that a limitation of this work is that it does not completely solve the issue of performance underestimation, especially in the context of correcting lexical errors that have a large set of possible corrections.

Another limitation of this work is that, while we consider the models to be low-resource for GEC, the methods that we used for creating synthetic training data rely on language-specific resources, such as a POS tagger and a morphological analyzer. Finally, we adopt the most common approach to GEC that operates at the sentence level, and we do not investigate error correction that looks at broader context.

Ethical Considerations

The annotation presented in this work is performed using data from an existing dataset that is publicly available for research (Mizumoto et al., 2012), and, more specifically, a subset of that data that was previously extracted and pre-processed (Trinh and Rozovskaya, 2021), which is also publicly available. The annotation presented in this work was manually generated by two native Russian speakers that were hired to perform the annotation for a compensation. The amount was set according to a compensation that was offered for similar annotation efforts, and that pay was deemed acceptable by the annotators.

The resulting annotations are expected to contribute to the development of robust systems for the grammatical error correction of Russian and should benefit learners of Russian as a foreign language. The dataset could also be of use to linguists working on second language acquisition, as it could provide insight about the types of errors made by learners of Russian. The authors are not aware of any potential problems that could result from the use of the data and the annotations.

Acknowledgments

The authors thank Galina Khabas, Mikhail Khabas, and Michael Svoysky for their annotation work. The authors are grateful to the anonymous reviewers for their insightful comments. This work was partly supported by the PSC-CUNY grant 64487-00 53.

References

A. Alsufieva, O. Kisselev, and S. Freels. 2012. Results 2012: Using flagship data to develop a russian learner corpus of academic writing. *Russian Language Journal*, 62:79–105.

A. Awasthi, S. Sarawagi, R. Goyal, S. Ghosh, and V. Piratla. 2019. Parallel iterative edit models for local sequence transduction. In *EMNLP-IJCNLP*.

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.

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, M. Felice, and T. Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *ACL*.

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

C. Bryant, Z. Yuan, M. R. Qorib, H. Cao, H. T. Ng, and T. Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*.

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 *BEA Workshop*.

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.

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

L. Choshen and O. Abend. 2018b. Inherent biases in reference-based evaluation for grammatical error correction and text simplification. In *ACL*.

L. Choshen, D. Nikolaev, Y. Berzak, and O. Abend. 2020. Classifying syntactic errors in learner language. In *CoNLL*.

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

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

S. Flachs, F. Stahlberg, and S. Kumar. 2021. Data strategies for low-resource grammatical error correction. In *BEA Workshop*.

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)*.

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*.

S. Katsumata and M. Komachi. 2019. (almost) unsupervised grammatical error correction using synthetic comparable corpus. In *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA-19)*.

S. Katsumata and M. Komachi. 2020. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

S. Kiyono, J. Suzuki, M. Mita, T. Mizumoto, and K. Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *EMNLP-IJCNLP*.

P. Li and S. Shi. 2021. Tail-to-tail non-autoregressive sequence prediction for Chinese grammatical error correction. In *ACL*.

A. D. McCarthy, C. Kirov, M. Grella, A. Nidhi, P. Xia, K. Gorman, E. Vylomova, S. J. Mielke, G. Nicolai, M. Silfverberg, T. Arkhangelskiy, N. Krizhanovsky, A. Krizhanovsky, E. Klyachko, A. Sorokin, J. Mansfield, V. Ernštreits, Y. Pinter, C. L. Jacobs, R. Cotterell, M. Hulden, and D. Yarowsky. 2020. UniMorph 3.0: Universal Morphology . In *LREC*.

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, M. Komachi, M. Nagata, and Y. Matsumoto. 2011. Mining revision log of language learning SNS for automated japanese error correction of second language learners. In *IJCNLP*.

B. Mohit, A. Rozovskaya, N. Habash, W. Zaghouani, and O. Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *ANLP Workshop*.

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

J. Naplava, M. Straka, J. Strakova, and A. Rosen. 2022. Czech Grammar Error Correction with a Large and Diverse Corpus. In *TACL*.

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

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

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

C. Napoles, K. Sakaguchi, and J. Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *EACL*.

H. T. Ng, S. M. Wu, Y. Wu, Ch. Hadiwinoto, and J. Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of CoNLL: Shared Task*.

K. Omelianchuk, V. Atrasevych, A. Chernodub, and O. Skurzhanskyi. 2020. GECToR ? Grammatical Error Correction: Tag, Not Rewrite . In *Building Educational Applications Workshop (BEA)*.

F. Palma Gomez, A. Rozovskaya, and D. Roth. 2023. A low-resource approach to the grammatical error correction of ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop in conjunction with EACL*.

S. Rothe, J. Mallinson, E. Malmi, S. Krause, and A. Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *ACL*.

A. Rozovskaya. 2022. Automatic Classification of Russian Learner Errors. In *LREC*.

A. Rozovskaya and D. Roth. 2010a. Annotating ESL errors: Challenges and rewards. In *BEA*.

A. Rozovskaya and D. Roth. 2010b. Generating confusion sets for context-sensitive error correction. In *Proceedings of EMNLP*.

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*.

K. Sakaguchi, C. Napoles, M. Post, and J. Tetreault. 2016. Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality. In *TACL*.

A. Sorokin. 2017. Spelling correction for morphologically rich language: a case study of russian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*.

A. Sorokin. 2022. Improved grammatical error correction by ranking elementary edits. In *EMNLP*.

A. Sorokin, A. Baytin, I. Galinskaya, E. Rykunova, and T. Shavrina. 2016. SpellRuEval: the first competition on automatic spelling correction for russian. In *Proceedings of the International Conference "Dialogue 2016"*.

O. Syvokon and O. Nahorna. 2021. UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language.

O. Syvokon and M Romanyshyn. 2023. The UNLP 2023 shared task on grammatical error correction for Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*.

V. A. Trinh and A. Rozovskaya. 2021. New dataset and strong baselines for the grammatical error correction of russian. In *ACL Findings*.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. 2017. Attention is all you need. *In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems*.

M. White and A. Rozovskaya. 2020. A comparative study of synthetic data generation methods for grammatical error correction. In *BEA*.

Z. Xie, G. Genthial, S. Xie, A. Y. Ng, and D. Jurafsky. 2018. Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction. In *NAACL*.

L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. . In *NAACL*.

H. Yannakoudakis, T. Briscoe, and B. Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

J. Ye, Y. Li, Q. Zhou, Y. Li, S. Ma, H.-T. Zheng, and Y. Shen. 2023. CLEME: Debiasing multi-reference evaluation for grammatical error correction. In *EMNLP*.

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

Y. Zhang, Z. Li, Z. Bao, J. Li, B. Zhang, C. Li, F. Huang, and M. Zhang. 2022. MuCGEC: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. In *NAACL*.

W. Zhao, L. Wang, K. Shen, R. Jia, and J. Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *NAACL*.

A Evaluation Metrics

Reference-based evaluations include several measures (Napoles et al., 2019; Dahlmeier and Ng, 2012; Bryant et al., 2017; Napoles et al., 2015a; Felice and Briscoe, 2015), and their comparison is out of the scope of this work.

The M2 metric used in this work is the most commonly used metric in GEC. M2 is edit-based and computes F-score (typically $F_{0.5}$ score, with precision being weighted higher than recall). ERRANT (Bryant et al., 2017), another commonly used metric, is very similar and also computes F-score. I-measure (Felice and Briscoe, 2015) calculates a weighted accuracy of edits, and GLEU (Napoles et al., 2015b) (inspired by BLEU from machine translation, as BLEU itself is not appropriate for GEC) calculates a weighted precision of overlapping n-grams. GLEU rewards n-gram overlap of the correction with the reference and penalizes unchanged incorrect n-grams in the correction. There is currently no consensus on the best reference-based evaluation metric for GEC. For example, GLEU seems to correlate better with human judgments than M2 (Napoles et al., 2017), but it is also a metric that tends to penalize most error types and discourages system from proposing changes (Choshen and Abend, 2018a).

The multi-reference evaluation we consider applies both to M2 and ERRANT. We expect the multi-reference benchmarks to be also useful for GLEU, as previous research suggests that the error types being penalized in GLEU and M2 are due to the edits being under-represented in the reference sets (Choshen and Abend, 2018a).

Finally, note that the F-scores are calculated independently for each reference in the presence of multiple references, and then the reference that maximizes the score is selected. One reason why taking a reference with the highest score makes sense is because we assume that the reference that is closest to the system output (in terms of edit overlap between system output and gold reference) will provide a more realistic evaluation of system performance. However, it can also be argued that because the set of gold references does not include all possible corrected versions of the source sentence, that there exists a reference that is even closer that might contain a set of (independent) changes from a union of two or more references. A recent work by Ye et al. (2023) attempts to address this issue by identifying independent changes in a set of multiple references. They show experiments on English datasets, and we leave it to future work to apply their framework to other languages, including Russian.

Error type	Rel. freq. by rater(%)		
	S (auto)	A (auto)	B (auto)
Spelling	**24.3**	**20.3**	17.9
Lex. (word)	10.4	6.9	9.1
Punc.	7.6	15.8	12.7
Noun case/num.	4.1	3.3	2.9
Prep.	5.5	4.2	4.0
Lex. (phrase)	10.1	16.8	**21.3**
Noun case	6.2	4.4	4.0
Insert	4.0	4.8	4.5
Adj. case	2.5	1.8	1.5
Verb agr.	2.2	1.8	1.4
Delete	2.7	3.6	4.7
Morph.	0.5	0.6	0.5
Total edits	3,383	4,910	5,257

Table B1: List of top-12 automatic error types and their distribution in the RU-Lang8 dataset (by rater). *Auto* denotes edit spans and error types obtained automatically: error categories are obtained when the automatic error classification tool is applied to the edit spans identified with ERRANT. Most frequent edit type for each rater is in bold.

B Annotation Statistics

To assign error categories to the edits produced by ERRANT, we apply a tool developed for Russian (Rozovskaya, 2022), that uses a POS tagger and a morphological analyzer (Sorokin, 2017) to automatically classify the edits into appropriate linguistic types. It should be noted that there is a language-agnostic error classification tool, SErCL (Choshen et al., 2020), which can also be used to classify errors. We chose the error classification tool that was specifically designed for Russian learner errors. The tool was also evaluated against gold error types in RULEC, while SErCL performance against Russian error types is not known. Finally, SErCL mainly considers syntactic error types, while we also include errors in derivational morphology.

Table B1 shows distribution of errors by type and annotatator in RU-Lang8.

C Generating Synthetic Data with Morphological Transformations

Spelling-based transformations (Grundkiewicz and Junczys-Dowmunt, 2019; Grundkiewicz et al., 2019) include highly confusable words based on edit distance obtained from a dictionary available

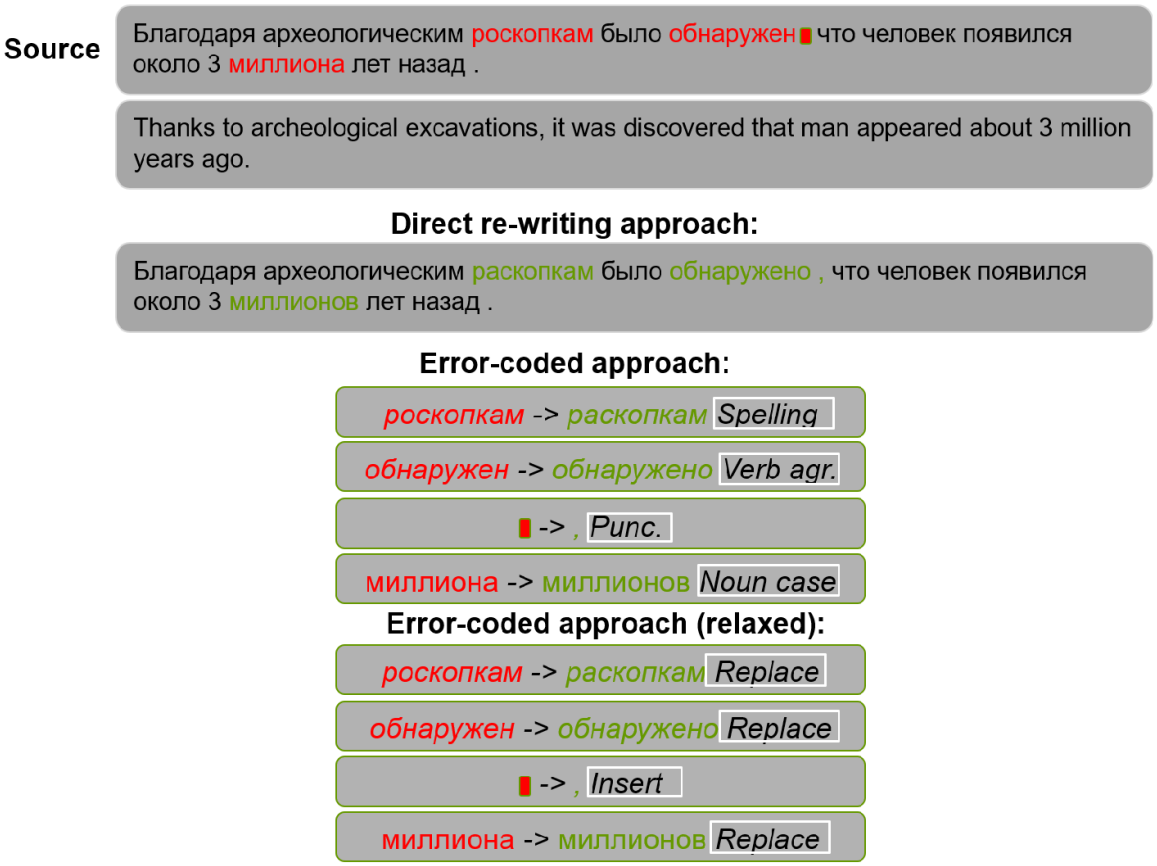

Figure A.1: Comparison of the annotation paradigms. The erroneous tokens are shown in red, and the changes are in green for emphasis. The error spans are marked manually in both the error-coded and the relaxed error-coded annotation, but the relaxed version does not specify linguistic error types; instead, changes are labeled as operations (Insert/Replace/Delete).

in a spellchecker, while morphology-based transformations utilize morphological variants of the same word. The latter method showed competitive results in English (Choe et al., 2019). However, Flachs et al. (2021) find that the morphology-based method underperforms across several languages, but they use Unimorph (McCarthy et al., 2020) to generate morphological confusions.

In this work, we use morphological transformations. Our intuition is that this method should perform well, given the rich morphology of Russian and our use of a language-specific analyzer. In addition, we employ a Russian spellchecker prior to running the GEC model (and a spelling-based synthetic data tends to correct spelling errors that the spellchecker already takes care of (White and Rozovskaya, 2020)). To generate morphological transformations, we compile a dictionary based on a large native corpus of Russian (250M tokens collected over the web (Borisov and Galinskaya, 2014) that has been pre-processed with the morphological analyzer (Sorokin, 2017). The dictionary is keyed on the base form for each word, and contains all wordforms (inflectional variants) corresponding to that base form that occurred in the 250M corpus. We use this dictionary to generate synthetic errors as follows: given a word in the monolingual training data, we use a POS tagger to obtain its POS tag and the morphological analyzer to obtain the base form. The token is then replaced with its inflectional variant that corresponds to the same base form.[10]

D Experimental Setup

Seq2seq models are trained using Transformer (Vaswani et al., 2017) implemented in the Fairseq toolkit. We use the "Transformer (big)" settings and the parameters specified in (Kiyono et al., 2019) for `Pretrain` setting. The models are trained until convergence and 10 checkpoints are averaged during inference. Results are averaged over 2 runs. The synthetic data is created by corrupting monolingual Russian data from the Yandex corpus (Sorokin, 2017) collected over the Web.

Data used to train the models Table D2 summarizes the gold and synthetic data used to train the models.

Model 1 (seq2seq) is pre-trained on 15M synthetic sentences (RULEC-dev data is used as validation data during the pre-training stage). Seq2seq models are typically further finetuned on gold training data, but finetuning on RULEC-train did not improve the scores on the RULEC dev data, and thus we skip the finetuning stage. We hypothesize that this happens because RULEC-train is relatively small, compared to the sizes of gold training data in other languages, e.g. 30K sentences in Ukrainian, or 10K sentences in Spanish, while only 4,800 sentences in RULEC-train.

Model 2 (mt5) is trained in 2 stages: (1) In stage 1, it is pre-trained on 10M synthetic sentences; (2) In stage 2, the model is further finetuned on RULEC-train. Both in (1) and (2) RULEC-dev is used as validation data.

Gold data	(1) RULEC-train is used for training and finetuning (2) RULEC-dev is used as validation data
Synth. data	(1) 15M sentences used to train seq2seq models from scratch (2) 10M sentences used to pretrain mT5 models in stage 1, before finetuning on RULEC-train

Table D2: Description of gold and synthetic data used to train the models.

[10] 15% of tokens are modified in this way, to mimic the errors rates in the learner data.

Error type	Example
Punc.	→,
Delete (open-class)	был “was” →
Insert (open-class)	→ для того “with the purpose of”
Prep. (ins.,del.,repl.)	в “in” → из “from, out of”
Conjunction	и (“and”) →
Noun case/number	иде-и (“idea” (sg.,gen/pl.,nom.)) → иде-й (“idea” (pl.,gen))
Noun case	специалист-ы “experts” (pl.,nom) → специалист-ам (pl.,dat.)
Noun number	пол-а“gender” (sg.,gen.) → пол-ов “gender” (pl.,gen.)
Adj. case	главн-ая “main” (sg., fem., nom.) → главн-ую (sg., fem., acc.)
Adj. number	дальнейш-ие “future” (pl.,nom.) → дальнейш-ее “future” (sg.,nom.))
Verb number/gender	жив-ут “live” (3rd person pl.) → жив-ет (3rd person sg.)
Verb other	соблазн-ить “to seduce” → соблазн-ил “seduced”
Verb aspect	чувствовала “feel” (past, imperf.) → по-чувствовала (past, perf.)
Verb voice	продолжала “continue” (past, active) → продолжала-сь (past, reflexive)
Verb tense	предлаг-ал “offered” (past tense) → предлаг-ает “offers” (present tense)
Deriv. morph.	вдохнов-ленным “inspired” → вдохнов-енной “inspiring”
Lex. (word)	предлагает “proposes” → утверждает “claims”

Table D3: Some common grammatical error types in Russian learner data. Partial changes on a word are shown with a hyphen.

Noun case

Это зависит от *показания/показаний очевидцев
This depends from $testimony_{\text{gen,*}\mathbf{sg}/\text{gen,}\mathbf{pl}}$ $eyewitness_{\text{gen,pl}}$
'This depends on the testimony of eyewitnesses'

Preposition

Слова *от/из прошлых уроков
$word_{\text{nom,pl}}$ *from**/out of** $previous_{\text{gen,pl}}$ $lesson_{\text{gen,pl}}$
'Words from previous lessons'

Verb number agreement

Все новые здания *разваливается/разваливаются
All $new_{\text{nom,pl}}$ $building_{\text{nom,pl}}$ $*fall_{\text{pres,imperfect,}\mathbf{sg}}/fall_{\text{pres,imperfect,}\mathbf{pl}}$ $apart$
'All new buildings are falling apart'

Verb gender agreement

Лера *пробовал/пробовала флиртовать с ним
Valerie $*try_{\text{past,imperfect,}\mathbf{masc}}/try_{\text{past,imperfect,}\mathbf{fem}}$ to flirt with him
'Valerie tried flirting with him'

Lexical choice

Тогда люди стали *спрашивать/задавать вопросы
Then $people_{\text{nom,pl}}$ started *to inquire/**to ask** $questions_{\text{acc,pl}}$
'Then people started to ask questions'

Word form

Такие окна не *пускают/пропускают свет
Such $windows_{\text{nom,pl}}$ do not $*allow_{\text{animate}}/allow_{\mathbf{inanimate}}$ light
'Such windows do not allow light'

Verb aspect

Она мне сразу *нравилась/понравилась
She I_{dat} immediately $*like_{\text{past,}\mathbf{imperfect}\text{,sg}}$ $like_{\text{past,}\mathbf{perfect}\text{,sg}}$
'I liked her immediately'

Missing word

Много необходимо сделать * /чтобы решить эту проблему
$Much_{\text{nom}}$ must to do * **/in order** to solve this $problem_{\text{acc,sg}}$
'A lot needs to be done to solve this problem'

Table A.3: Examples of common errors in the Russian learner corpus. Incorrect words are marked with an asterisk.

Gold annotator	**Rater S**			**Rater A**			**Rater B**		
	P	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**
Rater S	100.0	100.0	100.0	41.2	49.7	42.7	42.3	55.5	44.5
Rater A	55.1	33.6	48.9	100.0	100.0	100.0	43.0	39.4	42.2
Rater B	59.5	33.0	51.3	5.6	34.8	43.0	100.0	100.0	100.0

Table D4: Scoring one annotator against another (RULEC dataset).

Gold annotator	Rater S			Rater A			Rater B		
	P	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**
Rater S	100.0	100.0	100.0	48.6	54.5	49.7	42.8	51.0	44.3
Rater A	57.7	39.5	52.9	100.0	100.0	100.0	39.9	36.6	39.2
Rater B	54.0	30.0	46.5	42.8	30.4	39.6	100.0	100.0	100.0

Table D5: Scoring one annotator against another (RU-Lang8 dataset).