








<https://doi.org/10.17323/jle.2024.22272>

# Fighting Evaluation Inflation: Concentrated Datasets for Grammatical Error Correction

Vladimir Starchenko <sup>1</sup>, Darya Kharlamova <sup>1</sup>, Elizaveta Klykova <sup>2</sup>, Anastasia Shavrina <sup>1</sup>,  
Aleksy Starchenko <sup>1</sup>, Olga Vinogradova <sup>2</sup>, Olga Lyashevskaya <sup>1,3</sup>

<sup>1</sup> HSE University, Moscow, Russia

<sup>2</sup> independent researcher

<sup>3</sup> Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia

## ABSTRACT

**Background:** Grammatical error correction (GEC) systems have greatly developed over the recent decade. According to common metrics, they often reach the level of or surpass human experts. Nevertheless, they perform poorly on several kinds of errors that are effortlessly corrected by humans. Thus, reaching the resolution limit, evaluation algorithms and datasets do not allow for further enhancement of GEC systems.

**Purpose:** To solve the problem of the resolution limit in GEC. The suggested approach is to use for evaluation concentrated datasets with a higher density of errors that are difficult for modern GEC systems to handle.

**Method:** To test the suggested solution, we look at distant-context-sensitive errors that have been acknowledged as challenging for GEC systems. We create a concentrated dataset for English with a higher density of errors of various types, half-manually aggregating pre-annotated examples from four existing datasets and further expanding the annotation of distant-context-sensitive errors. Two GEC systems are evaluated using this dataset, including traditional scoring algorithms and a novel approach modified for longer contexts.

**Results:** The concentrated dataset includes 1,014 examples sampled manually from FCE, CoNLL-2014, BEA-2019, and REALEC. It is annotated for types of context-sensitive errors such as pronouns, verb tense, punctuation, referential device, and linking device. GEC systems show lower scores when evaluated on the dataset with a higher density of challenging errors, compared to a random dataset with otherwise the same parameters.

**Conclusion:** The lower scores registered on concentrated datasets confirm that they provide a way for future improvement of GEC models. The dataset can be used for further studies focusing on distant-context-sensitive GEC.

## KEYWORDS

Grammatical error correction, L2 errors, ESL, concentrated datasets, cross-sentence GEC

**Citation:** Starchenko, V., Kharlamova, D., Klykova, E., Shavrina, A., Starchenko, A., Vinogradova, O., & Lyashevskaya, O. (2024). Fighting evaluation inflation: Concentrated datasets for grammatical error correction. *Journal of Language and Education*, 10(4), 112-129. <https://doi.org/10.17323/jle.2024.22272>

### Correspondence:

Vladimir Starchenko,  
vmstarchenko@edu.hse.ru

**Received:** August 18, 2024

**Accepted:** December 16, 2024

**Published:** December 30, 2024

## INTRODUCTION

Grammatical error correction (GEC) is an important task of applied natural language processing (NLP). It involves identifying and correcting errors in word spelling and punctuation, modifying syntactic patterns, as well as suggesting the right word and word order to improve the readability and clarity of text. The definition of the task includes not only detection, classification, and correction of forms and structures that are "strictly grammatical in nature" (Bryant et al.,

2023) but also broader contextual analysis and fluency enhancement that ensure that the correction is consistent with the intended meaning and style of the text (Du & Hashimoto, 2023). GEC technologies can be used to assist children or second language (L2) learners, they can save language teachers' time, as well as optimize the work of proofreaders, editors, and other specialists dealing with texts.

GEC systems have greatly developed over recent decades. Qorib and Ng



(2022) note that the state-of-the-art GEC models GECToR (Omelianchuk et al., 2020) and T5 (Rothe et al., 2021) exhibit better results than human experts do from the point of view of common metrics, and yet, these systems still fail to detect and/or correct some errors that are easily handled by an educated native speaker. GEC is thus facing the crisis of metric resolution limit: while there is room for growth regarding the observed quality for various types of errors, metrics appear to have reached the ceiling.

At the moment, no solution to this problem has been implemented in the research field. The practice that could pave the way to the solution is to give the scores of a model for various types of errors separately. It allows GEC systems to reveal the more challenging types of errors, but it does not overcome the problem of challenging errors being underrepresented in the existing datasets. Additionally, this practice is scarce in the research (see Yuan & Bryant, 2021; Zhang et al., 2022 as examples) and, crucially, has not been used for further comparison and tuning of models. The present study makes a step towards solving the resolution limit problem.

One of the types of errors which are affected by the resolution limit problem are errors that require information from a distant context (i.e., context broader than one clause) for detection or correction. There is a consensus in the literature that such errors are particularly challenging for models to correct, both for technical reasons (such as the common practice of training models at the sentence level rather than the text level) and due to the difficulty of taking into consideration long-distance dependencies (Chollampatt et al., 2019; Yuan & Bryant, 2021; Qorib & Ng, 2022). Resolution limit makes the advancement of GEC systems with respect to the context-sensitive errors problematic, if common benchmarks and metrics are used.

The present study suggests using evaluation datasets with a higher ratio of the errors which still cause problems for GEC systems. Such datasets are expected to lower the scores of models and allow tuning them for challenging errors. We have selected context-sensitive errors as the material for testing the suggested approach to the resolution limit problem. The concentrated dataset we created comprises 1,014 examples collected from widely used GEC datasets. It consists of manually selected and additionally annotated examples, each containing at least one error that requires distant context for correction. To verify that the concentrated dataset provides higher resolution, we applied two neural networks, BART and T5, to solve the GEC task on the created dataset. We showed that the two GEC systems produce low scores when evaluated across the concentrated dataset, despite the fact that they show competitive results for GEC in general. Thus, creation of the concentrated dataset paves the way for GEC results to grow, as lower (but more accurate) scores make evaluation more distinctive and leave room for improvement. The present study also contributes

to the area of applying machine learning approaches to the problem of wide-context dependency, providing a tool for the evaluation and comparison of models with respect to context-sensitive errors.

## LITERATURE REVIEW

### GEC Task

Researchers have been trying to improve the results of error correction in texts since the beginning of the computer era. Initially, the practically-oriented studies focused on spelling correction (Cargill, 1980; Bentley, 1985), while GEC in a wider sense was mostly discussed as a preprocessing step for NLP systems that failed to process grammatically incorrect input (Kwasny & Sondheimer, 1981; Jensen et al., 1983). The first GEC tools created for practical use emerged later (Burstein et al. 2003; Leacock et al. 2009, among others), primarily relying on rule-based approaches.

Practically-oriented systems quickly moved to data-driven supervised machine learning designs relying on classification (e.g., Lee 2004; Rozovskaya & Roth, 2010; Dahlmeier & Ng, 2011) and statistical machine translation (SMT) architectures (e.g., Brockett et al., 2006; Yuan & Felice, 2013). A detailed survey of studies dedicated to GEC at this stage can be found in Leacock et al. (2014).

Since that survey, GEC systems have rapidly advanced with the development of deep learning and large language models (LLMs). GEC systems based on various architectures were implemented, including Recurrent Neural Networks (RNN – cf. Yuan & Briscoe, 2016; Xie et al., 2016; Wang et al., 2017), Convolutional Neural Networks (CNN – cf. Chollampatt & Ng, 2018), and Transformers (Edunov et al., 2018; Wang et al., 2019, and many subsequent studies). More detailed discussions of the recent advancement of GEC systems are presented by Wang et al. (2021) and Bryant et al. (2023).

### Resolution limit in GEC validation

Modern GEC models seem to have reached the resolution limit: they have previously received even higher scores in terms of common metrics ( $F_{0.5}$  proposed for the GEC task by Ng et al., 2014) than human experts (Qorib & Ng, 2022). However, more recent studies (e.g., Zhou et al., 2023; Li & Wang, 2024) claim even further improvements of GEC systems.

We must emphasize two crucial notes at this point. Firstly, when we refer to the “low” scores of annotators, we do not mean those that are caused by the inaccuracies in their annotation. Imperfect annotators’ agreement mostly results from the fact that they choose different options equally suitable for the correction of errors in the original text.

Secondly, the output of GEC systems cannot be claimed to be perfect either. Qorib and Ng (2022, pp. 2795–2797) list several types of errors that GEC models recurrently fail to locate and correct. Among them are such common classes of errors as inaccuracies in syntactic patterns (e.g., subject-verb agreement); errors in long sentences; sentences with high error rates; cross-sentence errors; errors that require paraphrasing of a sentence segment (like correcting phrases that do not sound authentic); and some others. Another notable error type that does not challenge a native speaker, but is repeatedly discussed as being problematic for modern GEC systems, is spelling errors (Chollampatt & Ng, 2018; Starchenko & Starchenko, 2023).

A likely explanation for these pitfalls of GEC systems is that the number of challenging errors in training and evaluation datasets is not large enough to noticeably affect the metrics. Yet the percentage of challenging errors is quite high both in manual data processing and in the application of GEC systems (see, e.g., the discussion of character-level errors in Starchenko & Starchenko, 2023). A possible reason for this is that the corpora used for the training and evaluation of GEC systems are created on the basis of non-native speakers' texts, which are often overloaded with basic grammatical errors.

Some researchers report the evaluation results for different types of errors separately (e.g., Yuan & Bryant, 2021; Zhang et al., 2022), providing special procedures for evaluating the efficiency of the model for errors identified worse than others. This practice is becoming more common, especially after the emergence of the ERRANT scorer (Bryant et al., 2017), which implements separate evaluations for various error types. There are currently no approaches that directly leverage such breakdown evaluation statistics in model enhancement. They are usually presented as a hindsight observation rather than used for tuning a GEC system, while they constitute the material which can be directly used for training and evaluation.

## Concentrated Datasets in NLP

Concentrated datasets are successfully applied in various domains of NLP and not only in GEC. One example is the handling of ethics-related biases by LLMs. While these biases are not frequent in the natural data, even a singular appearance in the output greatly impacts the use of such models in commercial practices. As a result, models are additionally fine-tuned and evaluated on concentrated datasets containing biased data (e.g., Nangia et al., 2020; Zhao et al., 2023).

In the GEC domain, concentrated datasets are currently not a widespread tool in evaluation or training. Starchenko (2024) created a concentrated synthetic dataset for fine-tuning an LLM for the GEC task, while Starchenko and Starchenko (2023) proposed a synthetic evaluation dataset. Both

studies are, however, restricted to spelling errors, which are arguably the most basic type of aforementioned challenging errors, naturally allowing for wide-scale synthesizing of “error — correction” pairs. Chollampatt et al. (2019) generated a synthetic dataset with tense errors. To the best of our knowledge, no concentrated dataset of natural language production has been used for GEC. The present study fills this gap by creating and applying a concentrated evaluation dataset consisting of annotated examples from several learner corpora.

## Context Dependency in GEC

The problem of context dependency is crucial for GEC. Identifying a grammatical error and suggesting a correction for it is highly dependent on the context, such as parts of speech of neighboring words, their lexical semantics, and word order. Discourse type and the general intentions of the author are also relevant to the way an error is corrected.

Since early on, GEC systems have greatly relied on the context, which has been achieved either by passing some of its features to a model (for classifiers) or by using architectures incorporating context-sensitivity (SMT, RNN, CNN, Transformers). It is, however, the local context around the error that usually receives more attention. Models are often trained for correcting sentences out of context (e.g., the state-of-the-art (SOTA) model by Rothe et al., 2021, pp. 703–704). Moreover, some of the commonly used GEC datasets contain sampled sentences, rather than paragraphs or full texts (cf. Napoles et al., 2017 for a relatively recently released dataset JFLEG). As a result, even the most powerful modern GEC models often fail to correct some types of errors that are more sensitive to the wider context, e.g., pronouns, verb tenses, modality, and usage of discourse markers.

Only a few studies pay special attention to the broader context in GEC. This problem is usually formulated in terms of cross-sentence errors, or “errors that require cross-sentence context to [be] correct[ed]” (Qorib & Ng, 2022). Chollampatt et al. (2019) created a CNN model that includes an additional encoder, preserving information from the previous sentences, and incorporated the encoding in the decoder via attention and gating mechanisms. Yuan & Bryant (2021) compared various Transformer-based architectures by measuring the performance on longer-context-sensitive errors.

## METHOD

### Working Definition of Context-Sensitive Errors

The narrower practical scope of this paper concerns errors that require taking into account distant context. The most straightforward case of distant-context-sensitive errors are

cross-sentence errors. Consider the discourse presented in (1):

(1) *I go for a walk to a park every day with my two lovely Corgi dogs. I met → [meet] many people in the park.*

The second sentence in (1) is correct when regarded on its own, but from the first sentence in the given context it is clear that the verb in the second sentence cannot be used in Past Simple and must be given in Present Simple. As discussed, it is such examples that are problematic for modern neural networks. Henceforth, we call such errors context-sensitive, subsuming distant rather than local context by the single term “context”.

Notably, context-sensitive errors may also emerge within one sentence. The modification (1)' minimally differs from (1), and the context required for correcting the tense error is practically the same as in (1).

(1)' *I go for a walk to a park every day with my two lovely Corgi dogs, and I met → [meet] many people in the park.*

As discussed later, we show that it is not the sentence borders that make errors challenging for GEC systems; cases like (1)' are also problematic for them. Thus, it is important to include errors that depend on a context which is distant yet is located within the same sentence. This removes a clear-cut border between local and distant context-sensitive GEC, which cannot be set up at the sentence border.

In order to operationalize the annotation of examples for the dataset, we use the following working definition:

**Definition:** *Errors that cannot be detected or corrected without access to the material from another clause headed by a finite verb are context-sensitive.*

Our dataset thus includes not only cross-sentence errors, but also cross-clause errors. Clearly, this definition excludes some possible (and arguably more debatable) cases of context-sensitive GEC, with the distant context located within the same clause or in a different non-finite clause. What is crucial for the present study is that the suggested definition allows us to include only uncontroversial cases of context-dependent GEC, while not limiting ourselves to cross-sentence examples. We leave a more theoretically-grounded definition of context-sensitive errors for further research.

## Creation of a Concentrated Dataset with a Higher Ratio of Context-Sensitive Errors

The concentrated dataset with a higher rate of cross-clause errors is built with the data extracted from existing error-annotated datasets. In this section, we focus on the algorithm of its creation and the characteristics of the four datasets that have formed it, while the resulting features of the dataset are presented in the Results section.

## Characteristics of the Non-Concentrated Datasets Used

The concentrated dataset comprises examples annotated for grammatical errors from the following four datasets.

**The First Certificate in English (FCE)** dataset (Yannakoudakis et al., 2011) contains texts of B1–B2 English learners in the style of a short essay, letter, or description, with each text corrected by one annotator. It is split into training, development, and evaluation subsets.

**The CoNLL-2014** dataset (Ng et al., 2014) is a part of the National University of Singapore Corpus of Learner English (NUCLE; Dahlmeier et al., 2013). It was created as an evaluation dataset for the CoNLL-2014 shared task and contains essays of C1 English learners. Different versions of CoNLL-2014 present annotations by 18 different experts.

**The Write & Improve (W&I) and LOCNESS (BEA-2019)** dataset (Bryant et al., 2019) was created for the BEA-2019 shared task and includes essays by A1–C2 English learners and by undergraduate native speakers. It is split into training, development, and evaluation subsets, with the latter annotated by 5 experts.

These datasets are frequently used for training and evaluation in GEC studies, including various shared tasks (Dale et al., 2012; Ng et al., 2014; Bryant et al., 2019).

**The Russian Error-Annotated Learner English Corpus (REALEC)** dataset (Vinogradova & Lyashevskaya, 2022) comprises essays of university English learners, most of them at B1–B2 levels of English proficiency. A single annotation approach for each type of error is described in the annotation guide, which has been used by four experts. While this corpus has been released recently and has only been used once in large-scale GEC studies (Volodina et al., 2023), it is particularly useful for the present research, as its annotations include discourse-related error types that are highly relevant for context-dependent GEC.

More detailed information about the datasets is summarized in Appendix A.

For FCE and BEA-2019, only evaluation subsets are taken into consideration. As CoNLL-2014 is an evaluation dataset, and since REALEC has not been actively used for GEC model training yet, there is no expectation that models could learn relevant examples from them during training. Therefore, the concentrated dataset should not be problematic for evaluation in this respect.

## Annotation of the Concentrated Dataset

Context-sensitive errors do not have any common features that allow for their easy automatic extraction, have not been

annotated in most existing datasets, and are not frequent. As a result, their extraction from the corpora requires a substantial amount of manual annotation.

To ensure that all the annotations conformed to the same standard and all sentences could be compared regardless of their source, we normalized the annotations for all the datasets by processing the corrections and automatically applying the Error Annotation Toolkit (ERRANT) tags (Bryant et al., 2017) to them. Additionally, we preserved the annotation of discourse-related errors from REALEC.

To make the annotation feasible, we focused on several error types that were expected to be context-sensitive more often (building on suggestions in Bryant et al., 2021). We chose to consider errors with ERRANT tags CONJ (conjunctions), DET (determiners), NOUN:INFL (nominal inflection), PRON (pronouns), PUNCT (punctuation), VERB:SVA (subject-verb agreement), VERB:TENSE (verbal tense), WO (word order), and REALEC tags *Inappropriate\_register* (stylistic errors), *Linking\_device* (discourse linking tools), and *Ref\_device* (usage of anaphoric expressions). For most of these errors, sensitivity to the information in the preceding and/or subsequent context does not have to be explained and is demonstrated by the examples in Appendices B and C. To mention just a couple of the types of such errors: the use of definite article for the first mentioning or of indefinite article for any further mentioning (annotated with DET); the use of predicates in present tenses when there is a reference to the specific time in the past in the context (annotated with VERB:TENSE); etc.

For each of the types, around 50 examples were annotated, and the tags with the highest ratio of context-sensitive errors were chosen for further annotation: ERRANT tags PRON, PUNCT, VERB:TENSE, and REALEC tags *Inappropriate\_register*, *Linking\_device*, and *Ref\_device*. Next, 140–260 examples of each of these error types were annotated. The number of annotated examples and the ratio of context-sensitive errors for each tag are presented in Appendix B. The description of less frequent tags is provided in Appendix C.

For each sentence, the initial tag assigned either by ERRANT or by an annotator (for REALEC) was displayed. An expert from the team of authors had to examine the sentence in context and decide whether it is necessary to take into account information from other clauses or sentences to locate and/or correct the error. For such examples, additional annotation had to be provided:

- whether context from another clause/sentence is required to locate the error;
- whether context from another clause/sentence is required to correct the error;
- the type of context required for locating or correcting the error, namely, whether it is a cross-clause or cross-sentence error;

- the distance in sentences or clauses (if the context is within the same sentence) from the one containing the error (i.e., the number of sentences or clauses that need to be considered to locate and correct the error);
- the direction in which this context is located: to the left, to the right, to any direction or to both directions from the erroneous sentence or clause;
- the type of error (see Appendices B and C).

As a result, we processed and annotated 3,403 errors in the extended context from four corpora of English learner texts and selected a total of 1,014 context-sensitive errors.

### ***Inter-Annotator Agreement***

To get a better understanding of the validity of our results, we calculated inter-annotator agreement. For this, we randomly chose 100 sentences representing all the initial error types. All of the sentences were marked up by the four annotators who worked on the whole dataset. We used this subset (henceforth called the agreement dataset) to calculate the inter-annotator agreement for the column “whether context from another clause/sentence is required to locate the error”.

Since the agreement dataset did not have empty values and there were more than two annotators, we used Krippendorff’s Alpha (Krippendorff, 2011), Fleiss’ Kappa (Fleiss, 1971), and Randolph’s Kappa (Randolph, 2005; Warrens, 2010). The main challenge connected with the metric calculation was that the two classes in the dataset were extremely unbalanced: the proportion of context-sensitive errors was relatively small compared to the whole body of errors. This disrupted the estimation of annotator agreement by random chance and resulted in an underestimation of agreement by the commonly used Krippendorff’s Alpha and Fleiss’ Kappa.

The rapid degradation of Krippendorff’s Alpha for the annotation of a small and unbalanced dataset (Marzi et al., 2024) is illustrated in Table 1. The table shows the application of Krippendorff’s Alpha to an imaginary dataset, annotated by three groups of experts. The first group shows perfect agreement; in the second group there is one error in the annotation; and in the third group two experts made one error each. One can see that even one error causes the score to drop to 0.429, and the second error makes it zero, despite the fact that intuitively the annotator’s agreement is relatively high.

To compensate for this, we resorted to using Randolph’s Kappa, which is less affected by class imbalance. Additionally, we provided a custom estimation of agreement: we compared each annotation in the agreement dataset to the annotation that ended up in the main dataset and calculated the percentage of annotators that agreed with the label from the main dataset. After that, we calculated the mean of this percentage across the subset we were working with:

**Table 1**

*Illustration of Rapid Degradation of Krippendorff's Alpha on Unbalanced Datasets*

| Expected value <sup>a</sup> | Case of annotation 1 <sup>b</sup> |      |      | Case of annotation 2 <sup>b</sup> |      |      | Case of annotation 3 <sup>b</sup> |      |      |
|-----------------------------|-----------------------------------|------|------|-----------------------------------|------|------|-----------------------------------|------|------|
|                             | A1.1                              | A1.2 | A1.3 | A2.1                              | A2.2 | A2.3 | A3.1                              | A3.2 | A3.3 |
| 1                           | 1                                 | 1    | 1    | 1                                 | 1    | 1    | 1                                 | 1    | 1    |
| 2                           | 2                                 | 2    | 2    | 2                                 | 2    | 1    | 1                                 | 2    | 1    |
| 1                           | 1                                 | 1    | 1    | 1                                 | 1    | 1    | 1                                 | 1    | 1    |
| K's $\alpha^c$              | 1.000                             |      |      | 0.429                             |      |      | 0.000                             |      |      |

Notes. <sup>a</sup>The correct annotation value expected for the imaginary dataset. <sup>b</sup>Annotations by three groups of experts. The gray cells show the cases of incorrect annotation by an expert. <sup>c</sup>Krippendorff's Alpha.

$$(2) \quad \frac{100}{n} \times \sum_{i=1}^n \frac{\sum_{j=1}^m [item_{i,j} = base_i]}{m}, \quad \text{where}$$

- n – the number of datapoints in the dataset,
- m – the number of annotators,
- item<sub>i,j</sub> – the annotation by jth annotator for ith datapoint,
- base<sub>i</sub> – the annotation in the main dataset corresponding to the ith datapoint.

While this method is unconventional, it provides a rough estimate of how well the experts agreed with the annotations that were used in the main dataset, helping to put other inter-rater agreement matrices into perspective.

As the most commonly used scores for Krippendorff's Alpha and Fleiss' Kappa were rather low, we calculated the inter-rater agreement scores for each initial error category to demonstrate what categories were the most and the least reliable. The scores for separate error categories, as well as for the whole agreement dataset, can be found in Table 2.

All four metrics indicate perfect agreement at 1 (or, in our case, 100, since we use percentages). However, each metric is interpreted slightly differently.

Krippendorff's Alpha can be either negative (indicating higher-than-chance disagreement among annotators) or positive (but not exceeding 1). Typically, inter-annotator agreement above 0.67 is considered high enough to be able to draw cautious conclusions based on the annotated data, while agreement above 0.8 is considered robust enough to consider the data reliable. While for our agreement dataset the metrics are not high enough, one should keep in mind that due to the small size of the dataset this metric is likely to show lower agreement than there actually is. Taking this into account, it can be assumed that the "real" agreement score is at least as high as 0.67.

Fleiss' Kappa and Randolph's Kappa are interpreted in almost the same way as Krippendorff's Alpha, with agreement above 0.6 considered substantial and agreement above

0.8 – almost perfect. For our dataset, we are close to the 0.6 threshold for Fleiss' Kappa and above it for Randolph's Kappa. Keeping in mind the fact that this metric is also sensitive to dataset size, it can be assumed that the real agreement is substantial.

As for the custom metric, while there is no conventional interpretation, we can see that its values are quite high, with an average of 9 out of 10 annotations conforming to those found in the main dataset.

### Evaluation of Context-Sensitive Errors

The evaluation procedure is crucial for context-sensitive GEC, because its standard implementation in the GEC task leads to consistently lower scores for longer texts, independently of their content.

#### Score Calculation

The most common measure used for the evaluation of predictive performance is F<sub>β</sub>-score. In the GEC task, F<sub>0.5</sub>-score is used most often, following Ng et al. (2014). It (arguably) represents the judgments of human experts about the quality of text correction (Grundkiewicz et al., 2015; Napoles et al., 2015; Chollampatt & Ng, 2018).

F<sub>β</sub>-score is a complex measure that takes into account True Positives (TP, cases in which a model made a correct prediction), False Negatives (FN, cases in which a model did not correct an error it was supposed to), and False Positives (FP, cases in which a model changed the text that does not contain errors). Precision = TP / (TP + FP) and Recall = TP / (TP + FN) are calculated as an intermediate step, and the multiplier β = 0.5 weights Precision twice as much as Recall. The straightforward interpretation of this metric is as follows: the higher the score, the better the corresponding model performs. That is, a model with a higher F<sub>0.5</sub>-score accurately corrects more errors and/or does not introduce correction in the fragments of the text that were not annotated as erroneous.

The most recent implementation of an F<sub>0.5</sub>-scorer is in ER-RANT by Bryant et al. (2017). One of the advantages of this

**Table 2***Inter-Annotator Agreement Metrics*

| Error type             | Krippendorff's Alpha | Fleiss' Kappa | Randolf's Kappa | Custom metric <sup>a</sup> | Share of context-sensitive errors in the dataset |
|------------------------|----------------------|---------------|-----------------|----------------------------|--|
| All types              | 55.9                 | 55.9          | 73.6            | 89.5                       | 9/100  |
| DET                    | N/A                  | N/A           | N/A             | 100                        | 0/18   |
| Inappropriate register | 0.0                  | -5.3          | 80.0            | 95.0                       | 0/5  |
| Linking device         | 60.4                 | 58.3          | 60.0            | 90.0                       | 2/5  |
| PRON                   | 35.0                 | 34.0          | 56.9            | 82.4                       | 0/17   |
| PUNCT                  | 53.2                 | 52.6          | 55.0            | 81.2                       | 2/20   |
| Ref_device             | 54.8                 | 52.4          | 60.0            | 90.0                       | 2/5  |
| VERB:TENSE             | 74.0                 | 73.5          | 77.8            | 90.0                       | 3/15   |
| WO                     | -1.7                 | -3.4          | 86.7            | 96.7                       | 0/15   |

Note. <sup>a</sup>Mean percentage of annotations conforming to those found in the main dataset.

tool is that it allows for the calculation of scores for each error type separately. The evaluation by ERRANT includes the following steps.

- (1) Preparation. The tool accepts as input a text with errors and a set of versions of this text corrected by experts. It calculates the sets of corrections that must be applied to the original text to obtain the experts' versions of the text, thus yielding the reference corrections. Likewise, the output of a model is compared to the original text, calculating the set of predicted corrections.
- (2) Calculation of  $F_{0.5}$ -scores. For each pair of a reference correction set and the prediction correction set, True Positives (TP), False Negatives (FN), and False Positives (FP) are calculated. Based on them, Precision, Recall, and  $F_{0.5}$ -score for each expert's annotation is found.
- (3) Choice of the closest annotator. Among the annotations provided by all experts, the one that has the highest  $F_{0.5}$ -score is chosen, and TP, FN, and FP of this annotation are selected for this text.
- (4) Iteration over texts. Steps 1–3 are repeated for every text in the dataset, meaning that different texts can be evaluated with respect to different annotators. By default, each sentence in the dataset is treated as a separate text.
- (5) Calculating the final score. TP, FN, and FP received for each text are summarized and used to calculate the  $F_{0.5}$ -score for the whole dataset.

A non-trivial property of the described algorithm is the built-in possibility to have more than one annotator for a dataset and the fact that the scorer relies on the closest possible annotation. This property reflects that language allows various ways of expressing the same thoughts, and that there can be various accurate corrections of the same errors. As a result, evaluating just one annotation (without the possibility for one annotator to suggest various corrections) is insufficient for the decision on the model's efficiency. A more substantial discussion may be found in Bryant and Ng (2015).

### **Relationship between Text Length and $F_{0.5}$ -Score**

The outlined algorithm and the way it solves the problem of variability in accurate corrections highly affects the evaluation of context-sensitive errors.

In L2 texts, the density of errors is relatively high (regularly more than one error per sentence). Due to this and to the fact that each error introduces possible variation, the combinatorics of accurate corrections may become complex. Some correct versions of a text generated by a GEC system may not be found in the reference annotations and would be unfairly claimed to be wrong.

To minimize this effect, the texts are split into sentences: the smaller a text fragment fed to a scorer and the fewer errors it contains, the greater the variation accounted for, meaning that the score is more accurate. If text fragments are small enough, one can expect that a large number of annotators cover all combinations of variable corrections within it.

To demonstrate this, we used ERRANT to evaluate the model BART (Katsumata & Komachi, 2020) on the CoNLL-2014 dataset. We calculated two measurements for the same output provided by the model applied at the full text level. The first measurement followed the regular ERRANT workflow, including splitting the dataset into sentences (note that the model is still applied at the text level). The second measurement differed from the first one in that it was conducted for whole texts. The results of these measurements are presented in Table 3. Additionally, we provide the measurement for the model applied to sentences rather than texts, showing that the model handles longer texts more poorly, which is one of the main focuses in this study.

The two measures calculated for the same prediction differ: as discussed, when longer units are considered, the score becomes lower. Notably, provided that the annotations are totally correct, it is the higher score that characterizes the



performance of the model better, making it reasonable to split texts into sentences for evaluation.

This solution, however, is problematic from the point of view of context-sensitive GEC, which by definition requires processing longer contexts. In some cases, context-sensitive errors are located within one sentence, and even though they require context for correction, it does not really affect the measurement. However, this is not always the case.

Firstly, there are errors located at the edge of sentences. The most straightforward example is a punctuation error such as the replacement of a period with a comma or vice versa, but more complicated cases are also possible.

Secondly, some errors are dependent on each other: it may be the case that two errors must be corrected in agreement with each other – like capitalizing the initial letter in the segment that follows the change of a comma for a period. Another regularly occurring example of this kind is sequences of coordinated verbs used in an incorrect tense with the whole sequence depending on the distant left context. For such examples, treating sentences separately is problematic: one annotator could make use of one form (e.g., Past Simple) throughout the whole sequence, while another annotator might choose a different form (e.g., Present Simple). If the sequence is separated into sentences, switching between Past and Present Simple would be erroneously evaluated by the model as correct.

In order to account for these problems, we perform the evaluation of the concentrated dataset in the following way:

- (1) To balance between the necessity of evaluating the shortest text fragments and the possibility of the incorrect treatment of context-sensitive errors, we split the texts into the smallest spans in which sentences with errors dependent on each other are not separated. That is, if an error occurred at the sentence border or its correction required merging two or more sentences, all sentences involved were taken for evaluation.
- (2) We only evaluate the annotated context-sensitive errors, which allows minimizing the distortions caused by enlarging the accessed contexts.
- (3) We preserve only one annotation for context-sensitive errors, provided that our manual annotation did not

reveal large-scale variation (corrections with variation were found for PRON errors, but there are only few such exceptions).

## Setup of the Experiment

To test the concentrated dataset, we use it to evaluate two SOTA GEC models: BART (large, Katsumata & Komachi, 2020) and T5 (base, Rothe et al., 2021). We chose these two models over more recently released GEC systems (e.g., Zhou et al., 2023) because the latter generally use one or more LLMs from a standard set, adding supplementary pre- or post-processing components. Provided that the overall result is comparable, we select less complex constructions to obtain a more interpretable result.

## RESULTS

### Concentrated Dataset of Context-Sensitive Errors

One practical result of the study is the creation of a concentrated dataset with a higher ratio of context-sensitive errors<sup>1</sup>. The dataset contains 1,014 context-sensitive errors with additional annotation.

Tables 4–6 present general information about the dataset: the distribution of error types, the representation of the original datasets in the concentrated dataset, and the type of context required for correcting an error.

As shown in Table 4, the concentrated dataset contains 5 main types of context-sensitive errors: pronouns (PRON), punctuation (PUNCT), referential device (REF), verb tense (VERB:TENSE), and linking device (LINK). Other types either have a low ratio of context-sensitive errors and were not extensively annotated (e.g., WO – word order) or emerged accidentally as a result of manually correcting inaccurate tag attribution by ERRANT.

Table 6 demonstrates that in most cases it is the left context that determines how the error must be corrected. Most commonly, only one sentence is enough for correction, but a range of larger distances is represented as well. Only the

**Table 3**

*BART Evaluation, Measurements for Sentences and Texts*

| Prediction    | Measurement   | FP  | FN    | TP    | Precision | Recall | F <sub>0.5</sub> |
|---------------|---------------|-----|-------|-------|-----------|--------|------------------|
| for texts     | for sentences | 507 | 1,589 | 1,111 | 68.67     | 41.15  | 60.56            |
|               | for texts     | 620 | 2,131 | 1,012 | 62.01     | 32.2   | 52.32            |
| for sentences | for sentences | 216 | 978   | 1,367 | 86.36     | 58.29  | 78.77            |

<sup>1</sup> <https://huggingface.co/datasets/startc/doc-gec>.



**Table 4***Error Types in the Concentrated Dataset*

| Error type  | Number |
|-------------|--------|
| PRON        | 259    |
| PUNCT       | 202    |
| REF         | 201    |
| VERB:TENSE  | 171    |
| LINK        | 140    |
| DET         | 19     |
| VERB:MODAL  | 8      |
| Other types | 14     |
| Sum         | 1,014  |

**Table 5***Representation of the Four Datasets in the Concentrated Dataset*

| Dataset    | Number of extracted errors |
|------------|----------------------------|
| REALEC     | 633                        |
| BEA-2019   | 218                        |
| FCE        | 135                        |
| CoNLL-2014 | 28                         |
| Sum        | 1,014                      |

**Table 6***Context Required for Correcting an Error*

| Type of context | # of sentences required for detection or correction | Errors |
|-----------------|---|--------|
| Left            | 1   | 810    |
|                 | 2   | 64     |
|                 | 3   | 22     |
|                 | 4   | 17     |
|                 | >4  | 21     |
|                 | Left, sum   | 934    |
| Right           | 1   | 59     |
|                 | 2   | 1      |
|                 | >4  | 1      |
|                 | Right, sum  | 61     |
| Left or right   | 1 to left, 1 to right                               | 16     |
| Left and right  | 1 to left, 1 to right                               | 2      |
|                 | 3 to left, 1 to right                               | 1      |
|                 | Left and right, sum                                 | 3      |
| Sum             |   | 1,014  |

right context is required to correct an error in about 1 out of 20 cases, and it is almost exclusively the neighboring sentence. On rare occasions, either left or right context suffices,<sup>2</sup> and in very few cases, both left and right contexts are required.

## Performance of GEC Systems on the Concentrated Dataset

To test the concentrated dataset, we measure the performance of SOTA GEC models BART and T5. Table 7 presents the results of the evaluation.

Before discussing the patterns in the data, we must comment on a feature of the measurement that highly affects the results. The number of False Positives in the table is zero for every row. Consequently, for every row, the Precision equals 100. This directly follows from the measurement

procedure described in the Materials and Methods section: we only evaluate context-sensitive errors and therefore do not access other types of errors, to which False Positives are automatically assigned. To evaluate the number of context-sensitive False Positives properly, one would have to manually process all the False Positives in the output of every model and annotate whether they are dependent on distant context.

As a result,  $F_{0.5}$ -scores presented in the table must be treated as the upper bound estimate. The total absence of False Positives in the output of even the best-performing model is outstandingly unlikely, so the actual  $F_{0.5}$ -scores are lower. Provided that the 0.5 coefficient of the  $F_{0.5}$ -score weighs Precision twice as much as Recall, the  $F_{0.5}$ -estimates in Table 7 are significantly more optimistic than they should be, and considering the raw True Positives, False Negatives, and Recall is more relevant.

<sup>2</sup> Note that if the material necessary for correction can be found in both left and right contexts, but is closer on one side, only the closest context is used. For example, if a clause can be corrected based on the previous sentence or the one located in four sentences to the right, we only take into account the left context.

**Table 7***Evaluation of the Concentrated Dataset by BART and T5*

| Error type | BART |     |     |      |       |                  | T5 |     |    |      |      |                  |
|------------|------|-----|-----|------|-------|------------------|----|-----|----|------|------|------------------|
|            | FP   | FN  | TP  | Prec | Rec   | F <sub>0.5</sub> | FP | FN  | TP | Prec | Rec  | F <sub>0.5</sub> |
| All types  | 0    | 893 | 121 | 100  | 11.93 | 40.39            | 0  | 986 | 24 | 100  | 2.38 | 10.85            |
| PUNCT      | 0    | 147 | 55  | 100  | 27.23 | 65.17            | 0  | 192 | 10 | 100  | 4.95 | 20.66            |
| VERB:TENSE | 0    | 142 | 29  | 100  | 16.96 | 50.52            | 0  | 165 | 4  | 100  | 2.37 | 10.81            |
| PRON       | 0    | 239 | 20  | 100  | 7.72  | 29.5             | 0  | 197 | 4  | 100  | 1.99 | 9.22             |
| REF        | 0    | 192 | 9   | 100  | 4.48  | 18.99            | 0  | 256 | 3  | 100  | 1.16 | 5.54             |
| LINK       | 0    | 138 | 2   | 100  | 1.43  | 6.76             | 0  | 140 | 0  | 100  | 0    | 0                |

Even with these caveats, the scores in Table 7 are definitely lower than the scores for errors that are not sensitive to distant context. The overall scores of BART are 40.39 for context-sensitive errors vs. 78.04 for non-context-sensitive errors (the latter measured on the CoNLL-2014 dataset); for T5, the measurements are 10.85 and 74.38, respectively.

If we look at the Recall, the poor performance of the models on context-sensitive errors becomes even more noticeable. For PUNCT as the best-handled error, and with BART as the best-performing model, only 27.23% of errors are properly corrected. The other results are even lower, and most values (all of them for T5; REF and LINK for BART) are close to noise. At the same time, the dataset provides distinctive power to observe the difference between the quality of the models' performance: BART consistently shows higher results than T5. This fact confirms that the poor results obtained on the concentrated dataset do not boil down to its inner properties, but reveal imperfections of GEC systems with respect to selected types of errors.

Lastly, it is interesting to point out the pattern in the difference of metrics for different error types. The highest scores are attributed to punctuation, which presents an artificially regulated construction above the writing system, and tense, which is the only purely grammatical type in the sample. Pronouns as anaphoric means represent a more discourse-oriented language domain, yet they are usually discussed as a part of the grammar, while the type Referential\_device contains lexically encoded (and less grammar-related) anaphoric means. Lastly, linking phrases are purely in the discourse realm. Thus, one could claim that the quality of error correction decreases with the shift of the error type from grammar to discourse.

## DISCUSSION

Studying the evaluation scheme in distant-context-sensitive GEC tasks, we have been able to make several observations. First, we have proved that the dataset with a distribution bias of error types helps to realistically assess the model performance. In fighting inflation in evaluation metrics

obtained on conventional GEC datasets, the concentrated datasets may serve as additional indicators of the models' failures, along with breakdown per-type evaluation reports (Bryant et al., 2017).

Second, we have noticed that the evaluation metrics differ significantly across four subsets stratified according to the data source. As shown in Figure 2, the F<sub>0.5</sub>-score ranges from 0.61 in BEA-2019 to 0.26 in REALEC, while Recall ranges from 0.24 in BEA-2019 to 0.07 in REALEC. This is in line with other comparative GEC studies based on full test or evaluation datasets (Zhang et al., 2023; Volodina et al., 2023, among others), confirming that the observed variance can be attributed to many factors such as L2 proficiency level, text register, writing task type, text length, sentence length, annotation strategy, and associated differences in the distribution of error tags. Yet, it is necessary to note that the drop in performance across subsets in the concentrated data is clearly more pronounced compared to results observed on non-concentrated datasets.

Third, even though the concentrated dataset is relatively small to be able to draw decisive conclusions, we have observed that error types are associated with the amount of context needed to detect and correct errors. The latter information can be extracted from the dataset annotations as the number of context units (sentences or clauses). For instance, the vast majority of VERB:TENSE errors require no more than one clause (see example (3)), whereas errors tagged as LINK tend to be associated with one or more sentences in the left or right window (see example (4)). Obviously, this affects the overall metrics.

(3) *When I was little I had → [∅] tried a lot of sports...*

(4) *From 2000 the percentage of elderly people in Sweden began to rise to 20 per cent. Moreover → [Contrary to that], from 2000 the percentage in the USA was at the same level of 14 per cent.*

Further applications of the received results will involve more experiments with different GEC architectures and methods to understand the metric variability across datasets and the role of the available context in models' performance.

While the difference in the  $F_{0.5}$ -scores for the concentrated and non-concentrated datasets are evident, the suitability of this metric for the GEC task remains an open question. With recent advances in generative models prompting, Recall is reported to be equal to, or even greater than, Precision. In this regard, Zeng et al. (2024) suggest using  $F_1$  and  $F_2$  scores as representative metrics in GEC results. As we have shown,  $F_{0.5}$ , Precision, and Recall calculated for the same model applied to texts vs. separate sentences and measured in text-based vs. sentence-based conditions (see Table 3 above) do not directly correspond to each other. The harmonization of metrics is necessary to establish a consistent benchmark for distant-context-sensitive GEC in various settings.

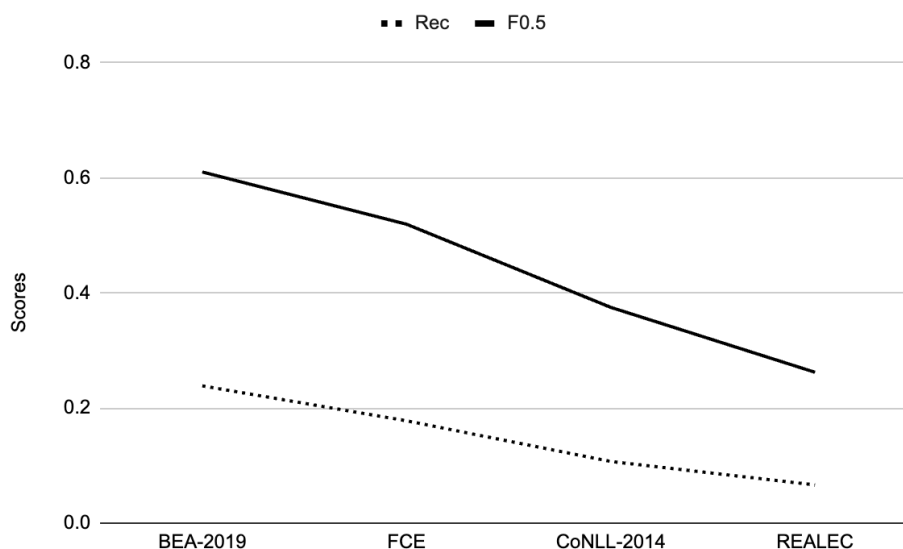
## LIMITATIONS

The nature and key properties of the corpora have to be assessed in the task of compiling the concentrated dataset. Future work may focus on increasing the size of the dataset, balancing the examples with regard to the proficiency level of the authors and to error types, and involving more experts to ensure the robustness of the annotations.

Another limitation of our approach is that the dataset presented in this article is just a preliminary step towards detailed surveys in data curation, evaluation techniques, and model training in the field. We only used off-the-shelf models for evaluation. It is clear that future experiments with training models using concentrated (training) datasets are needed to improve the overall understanding of the role of the error-type bias methods in distant-context-sensitive GEC.

### Figure 2

Evaluation Results for BEA-2019, FCE, CoNLL-2014, and REALEC Subsets of the Concentrated Dataset (BART Model)



<sup>3</sup> <https://huggingface.co/datasets/startc/doc-gec>

## CONCLUSION

In this study, we propose using a concentrated dataset with a high ratio of context-sensitive errors as a way to solve the resolution limit problem in GEC. This problem arises because the metrics commonly used for evaluating GEC systems may overestimate the model performance, even though certain types of errors are frequently overlooked by these models. By manually annotating examples of various error types (those related to punctuation, verb tense, determiners, pronouns, referential tools, and linking constructions), we have created a dataset containing 1,014 errors that require distant context for identification and/or correction. We have evaluated two GEC models on this dataset and demonstrated that their performance is significantly lower on a concentrated dataset compared to a non-concentrated one. This finding confirms that GEC systems still require substantial improvement and highlights the potential of concentrated datasets as a tool for both training and evaluation.

Based on the performance of the two models across different error types, we hypothesize that error correction becomes more challenging as the error type shifts from the realm of grammar to discourse. For instance, errors in punctuation and verb tense are corrected more successfully than those related to referential and linking devices.

Overall, this article demonstrates the potential of using concentrated datasets with a high ratio of context-sensitive errors to further enhance GEC systems and improve their applicability to real-world tasks. As a practical contribution, we publish the dataset<sup>3</sup>.

## ACKNOWLEDGMENTS

This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

## CONFLICT OF INTERESTS

None declared.

## AUTHOR CONTRIBUTIONS

**Vladimir Starchenko:** Conceptualization; Data mining; Data curation (automatization); Investigation; Methodology; Model testing; Project administration; Statistics; Supervision; Writing – original draft; Writing – review & editing.

**Elizaveta Klykova:** Conceptualization; Data curation; Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing.

**Anastasia Shavrina:** Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

**Olga Vinogradova:** Conceptualization; Data curation; Investigation; Methodology; Writing – original draft.

**Olga Lyashevskaya:** Conceptualization; Data curation; Investigation; Supervision; Writing – original draft; Methodology; Project administration; Writing – review & editing.

**Darya Kharlamova:** Conceptualization; Data curation; Investigation; Methodology; Resources; Writing – original draft, Writing – review & editing.

**Aleksey Starchenko:** Conceptualization; Investigation; Methodology; Project administration; Statistics; Supervision; Writing – original draft; Writing – review & editing.

## REFERENCES

- Bentley, J. (1985). Programming pearls: A spelling checker. *Communications of the ACM*, 28(5), 456–462. <https://doi.org/10.1145/3532.315102>
- Brockett, C., Dolan, B., & Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. *21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* (pp. 249–256). Association for Computational Linguistics. <http://dx.doi.org/10.3115/1220175.1220207>
- Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications* (pp. 52–75). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/W19-4406>
- Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 793–805). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/P17-1074>
- Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3), 643–701. [http://dx.doi.org/10.1162/coli\\_a\\_00478](http://dx.doi.org/10.1162/coli_a_00478)
- Bryant, C., & Ng, H. T. (2015). How far are we from fully automatic high quality grammatical error correction? *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (vol. 1: Long Papers, pp. 697–707). Association for Computational Linguistics. <http://dx.doi.org/10.3115/v1/P15-1068>
- Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion<sup>SM</sup> online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Conference on Innovative Applications of Artificial Intelligence* (pp. 3–10). American Association for Artificial Intelligence.
- Cargill, T. A. (1980). The design of a spelling checker's user interface. *ACM SIGOA Newsletter*, 1(3), 3–4. <https://doi.org/10.1145/1017923.1017924>
- Chollampatt, S., & Ng, H. T. (2018). A multilayer convolutional encoder-decoder neural network for grammatical error correction. *Proceedings of the AAAI conference on artificial intelligence* (vol. 32(1), pp. 5755–5762). Association for the Advancement of Artificial Intelligence. <http://dx.doi.org/10.1609/aaai.v32i1.12069>
- Chollampatt, S., Wang, W., & Ng, H. T. (2019). Cross-sentence grammatical error correction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 435–445). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/P19-1042>

- Dahlmeier, D., & Ng, H. T. (2011). Grammatical error correction with alternating structure optimization. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 915–923). Association for Computational Linguistics.
- Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 22–31). Association for Computational Linguistics.
- Dale, R., Anisimoff, I., & Narroway, G. (2012). HOO 2012: A report on the preposition and determiner error correction shared task. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 54–62). Association for Computational Linguistics.
- Du, Z., & Hashimoto, K. (2023). Sentence-level revision with neural reinforcement learning. *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)* (pp. 202–209). Association for Computational Linguistics.
- Grundkiewicz, R., Junczys-Dowmunt, M., & Gillian, E. (2015). Human evaluation of grammatical error correction systems. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 461–470). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/D15-1052>
- Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 489–500). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/D18-1045>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Jensen, K., Heidorn, G., Miller, L., & Ravin, Y. (1993). Parse fitting and prose fixing. *Natural Language Processing: The PLNLP Approach* (pp. 53–64). Springer. [https://doi.org/10.1007/978-1-4615-3170-8\\_5](https://doi.org/10.1007/978-1-4615-3170-8_5)
- Katsumata, S., & Komachi, M. (2020). Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. *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* (pp. 827–832). Association for Computational Linguistics.
- Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. [https://repository.upenn.edu/asc\\_papers/43](https://repository.upenn.edu/asc_papers/43)
- Kwasny, S. C., & Sondheimer, N. K. (1981). Relaxation techniques for parsing grammatically ill-formed input in natural language understanding systems. *American Journal of Computational Linguistics*, 7(2), 99–108.
- Lee, J. S. (2004). Automatic article restoration. *Proceedings of the Student Research Workshop at HLT-NAACL 2004* (pp. 31–36). Association for Computational Linguistics.
- Nangia, N., Vania, C., Bhlerao, R., & Bowman, S. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1953–1967). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.emnlp-main.154>
- Omelianchuk, K., Atrasevych, V., Chernodub, A., & Skurzshanskiy, O. (2020). GECToR—Grammatical Error Correction: Tag, not Rewrite. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–170). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.bea-1.16>
- Leacock, C., Gamon, M., & Brockett, C. (2009). User input and interactions on Microsoft Research ESL assistant. *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 73–81). Association for Computational Linguistics.
- Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners* (2nd ed.). Morgan & Claypool Publishers. <https://doi.org/10.1007/978-3-031-02153-4>
- Li, W., & Wang, H. (2024). Detection-correction structure via general language model for grammatical error correction. *arXiv preprint arXiv:2405.17804*. <http://dx.doi.org/10.48550/arXiv.2405.17804>
- Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator—Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *Methods X*, 12, 102545. <https://doi.org/10.1016/j.mex.2023.102545>
- Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2015). Ground truth for grammatical error correction metrics. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 2: Short Papers, pp. 588–593). Association for Computational Linguistics. <http://dx.doi.org/10.3115/v1/P15-2097>
- Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the eighteenth conference on computational natural language learning: Shared task* (pp. 1–14). Association for Computational Linguistics. <http://dx.doi.org/10.3115/v1/W14-1701>

- Qorib, M. R., & Ng, H. T. (2022). Grammatical error correction: Are we there yet? In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2794–2800). International Committee on Computational Linguistics.
- Randolph, J. J. (2005). Free-marginal multirater kappa (multirater K[free]): An alternative to fleiss' fixed-marginal multirater kappa. *Presented at the Joensuu Learning and Instruction Symposium 2005* (October 14–15, 2005). <http://files.eric.ed.gov/fulltext/ED490661.pdf>
- Rothe, S., Mallinson, J., Malmi, E., Krause, S., & Severyn, A. (2021). A simple recipe for multilingual grammatical error correction. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (vol. 2: Short Papers, pp. 702–707). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2021.acl-short.89>
- Rozovskaya, A., & Roth, D. (2010). Training paradigms for correcting errors in grammar and usage. *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 154–162). Association for Computational Linguistics.
- Rozovskaya, A. & Roth, D., (2021). How good (really) are grammatical error correction systems? *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2686–2698). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.eacl-main.231>
- Sakaguchi, K., Napoles, C., Post, M., & Tetreault, J. (2016). Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4, 169–182. <http://dx.doi.org/10.18653/v1/P18-1020>
- Starchenko, V. M., & Starchenko, A. M. (2023). Here we go again: modern GEC models need help with spelling. *Proceedings of ISP RAS*, 35(5), 215–228. [http://dx.doi.org/10.15514/ISPRAS-2022-35\(5\)-14](http://dx.doi.org/10.15514/ISPRAS-2022-35(5)-14)
- Starchenko, V. M. (2024). No need to get wasteful: The way to train a lightweight competitive spelling checker. *Computación y Sistemas*, 28(3), 1–12. <https://doi.org/10.13053/CyS-28-4-5068>
- Vinogradova, O., & Lyashevskaya, O. (2022). Review of practices of collecting and annotating texts in the learner corpus REALEC. *International Conference on Text, Speech, and Dialogue* (pp. 77–88). Springer International Publishing. [http://dx.doi.org/10.1007/978-3-031-16270-1\\_7](http://dx.doi.org/10.1007/978-3-031-16270-1_7)
- Volodina, E., Bryant, C., Caines, A., De Clercq, O., Frey, J., Ershova, E., Rosen, A., & Vinogradova, O. (2023). MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. *Linköping Electronic Conference Proceedings* (pp. 1–16). LiU Electronic Press. <https://doi.org/10.3384/ecp197001>
- Wang, C., Li, R., & Lin, H. (2017). Deep context model for grammatical error correction. *SLaTE* (pp. 167–171). International Speech Communication Association. <http://dx.doi.org/10.21437/SLaTE.2017-29>
- Wang, Y., Xia, Y., He, T., Tian, F., Qin, T., Zhai, C., & Liu, T. Y. (2019). Multi-agent dual learning. *Proceedings of the International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations.
- Wang, Y., Wang, Y., Dang, K., Liu, J., & Liu, Z. (2021). A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology*, 12(5), 1–51. <http://dx.doi.org/10.1145/3474840>
- Warrens, M. J. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, 4(4), 271–286. <https://doi.org/10.1007/s11634-010-0073-4>
- Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., & Ng, A. Y. (2016). Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*. <http://dx.doi.org/10.48550/arXiv.1603.09727>
- Yannakoudakis, H., Briscoe, T., & Medlock, B. (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* (pp. 180–189). Association for Computational Linguistics.
- Yuan, Z., & Bryant, C. (2021). Document-level grammatical error correction. *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 75–84). Association for Computational Linguistics.
- Yuan, Z., & Felice, M. (2013). Constrained grammatical error correction using statistical machine translation. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task* (pp. 52–61). Association for Computational Linguistics.
- Yuan, Z., Briscoe, T., & Felice, M. (2016). Candidate re-ranking for SMT-based grammatical error correction. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 256–266). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/W16-0530>
- Zeng, M., Kuang, J., Qiu, M., Song, J. and Park, J. (2024). Evaluating prompting strategies for grammatical error correction based on language proficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 6426–6430). ELRA and ICCL. <https://doi.org/10.48550/arXiv.2402.15930>

- Zhang, Y., Zhang, B., Li, Z., Bao, Z., Li, C., & Zhang, M. (2022). SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2518–2531). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2022.emnlp-main.162>
- Zhao, J., Fang, M., Pan, S., Yin, W., & Pechenizkiy, M. (2023). GPTBIAS: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*. <http://dx.doi.org/10.48550/arXiv.2312.06315>
- Zhou, H., Liu, Y., Li, Z., Zhang, M., Zhang, B., Li, C., Zhang, J., & Huang, F. (2023). Improving Seq2Seq grammatical error correction via decoding interventions. *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7393–7405). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2023.findings-emnlp.495>



## APPENDIX A

### GENERAL INFORMATION ABOUT THE SOURCE DATASETS USED FOR THE COLLECTION OF THE CONCENTRATED DATASET

| Dataset                   | Size, tokens | # of annotations per document <sup>a</sup> | Error types | Language proficiency |
|---------------------------|--------------|--|-------------|----------------------|
| FCE, evaluation part      | 41.9k        | 1  | 71          | B1–B2                |
| CoNLL-2014                | 30.1k        | 2–18                                       | 28          | C1                   |
| BEA-2019, evaluation part | 85.7k        | 5  | 55          | A1–Native            |
| REALEC                    | 1550.6k      | 1  | 48          | B1–B2                |

Note. <sup>a</sup> The number of annotation sets (by different annotators) provided for each document.

## APPENDIX B

## ERROR TAGS USED IN THE DATASET AND THE RATIO OF CONTEXT-SENSITIVE ERRORS

| Original tag <sup>a</sup> | New tag <sup>b</sup>   | Ratio of distant-context-sensitive errors <sup>c</sup> | Description  | Example <sup>d</sup>   |
|---------------------------|------------------------|--|--|--|
| Linking_device            | LINK                   | 59,05%   | The linking device is either wrong or erroneously absent   | Secondly, the majority of the population will use other kinds of public transport, for example, trains, cars, or ships. <b>So</b> → <b>However</b> , we cannot say that these types of transport harm our environment less than planes do.   |
| Ref_device                | REF                    | 50,83%   | The wrong referential device is used                       | We should not create barriers for ambitious people and accept <b>persons</b> → <b>those</b> who don't have interest in education just because of sex equality.   |
| VERB:TENSE                | VERB:TENSE             | 45,35%   | The wrong verb tense is chosen                             | When I was small, we lived in the country. I <b>remembered</b> → <b>remember</b> , we used to have oil lamps which used a cotton string dipping in the oil in the small bottle and made it burn the tip of the cotton string to give us light during the night.  |
| PUNCT                     | PUNCT                  | 37,61%   | The wrong punctuation mark is used                         | In Sweden the level fell from 84% to 15%, a similar situation was in France. <b>The</b> → <b>:</b> <b>the</b> level changed from 90% to 50%.   |
| PRON                      | PRON                   | 36,72%   | The personal pronoun is either wrong or erroneously absent | Also, he is very funny and I laugh a lot with him. <b>Both</b> → <b>We both</b> like to travel around the world and to do some sports, for example, tennis, running or trekking.   |
| Inappropriate_register    | REF, PRON <sup>4</sup> | 15,50%   | Errors related to style and appropriateness                | When a child begins learning, for example, English in primary school, <b>he</b> → <b>they</b> get the necessary basis for further studying. (Tagged as PRON)<br><br>Unfortunately, watching sports doesn't teach <b>us</b> → <b>viewers</b> anything and people don't get any information about the surrounding world from it. (Tagged as REF) |
| DET                       | DET                    | 9,45%  | The determiner is either wrong or erroneously absent       | This situation creates a lot of pollution for <b>∅</b> → <b>the</b> environment, so we have to be more concerned about the planet's health.  |

Notes. <sup>a</sup> Original tag is the tag used in the original dataset. <sup>b</sup> New tag is the tag used in the concentrated dataset.

<sup>c</sup> Ratio of distant-context-sensitive errors denotes the percentage of such errors among all annotated errors marked with the original tag.

<sup>d</sup> For clarity purposes, all the other mistakes present in the example sentences were corrected in accordance with corrections suggested by the annotators of the source datasets.

<sup>4</sup> During the annotation process, we concluded that other tags (such as PRON or REF) were suitable for the context-sensitive examples tagged as Inappropriate\_register in REALEC.

## APPENDIX C

### OTHER TAGS USED IN THE CONCENTRATED DATASET

| Tag        | Description   | Example   |
|------------|---|---|
| LEX        | Lexical choice error  | Also, it is a good way to get some positive emotions. <b>All of this</b> → <b>Watching sports</b> can even promote future productivity at work.   |
| NOUN:NUM   | The noun is used in the wrong number  | By the way, there is an opposite tendency with young people, their <b>num-ber</b> → <b>numbers</b> are the largest at the science courses and the smallest in the sports and health courses. Additionally, students of the health and sports <b>course</b> → <b>courses</b> are mostly middle-aged. |
| SPELL      | Spelling error  | To sum up, both characteristics are important in our life. We need to know how to operate with <b>once</b> → <b>ones</b> we were born with and know how to develop knowledge gained from our experience to have a successful life and reach goals we set for ourselves.                             |
| SYN        | Wrong choice or erroneous change of syntactic structure                                     | Although the grandparents are in most cases ready to help, they can not transfer the values of the new world to the kids, and <b>their</b> → <b>this</b> results in the wrong choice of paths of life for the grown-up adults in future.  |
| VERB:MODAL | The modal verb is erroneously absent, unnecessarily present, or used incorrectly            | In addition, to decrease the risk of negative comments or posts, Facebook and Twitter <b>would</b> → <b>should</b> improve their futures by solving the personal privacy problem.   |
| VERB:SVA   | Errors related to subject-verb agreement  | Today, public transport still <b>play</b> → <b>plays</b> an important role in the transport system and it will keep on doing so in the future.  |
| WO         | Errors in word order, e.g., the subject and verb are not inverted in the necessary contexts | But when I was a teenager, I began to experience situations that I did not like, for instance, girls said <b>to me bad things</b> → <b>bad things to me</b> or they talked unkindly about me.   |