No Error Left Behind: Multilingual Grammatical Error Correction with Pre-trained Translation Models

Anonymous EACL submission

Abstract

Grammatical Error Correction (GEC) enhances language proficiency and promotes effective communication, but research has primarily centered around English. We propose a simple approach to multilingual and low-resource GEC by exploring the potential of multilingual machine translation (MT) models for error correction. We show that MT models are not only capable of error correction out-of-the-box, but that they can also be fine-tuned to even better correction quality. Results show the effectiveness of this approach, with our multilingual model outperforming similar-sized mT5-based models and even competing favourably with larger models.

1 Introduction

Grammatical Error Correction (GEC) systems are a vital link between expert language use and clear communication, enhancing writing skills and language learning. However, GEC research has primarily focused on the English language with much less coverage for other languages, resulting in English-oriented methodologies and data scarcity for other languages. This highlights the need to diversify GEC research, ensuring that the benefits of these systems extend to all languages for a more inclusive global linguistic landscape.

In the evolving multilingual and non-English Grammar Error Correction (GEC) landscape, two recent notable keywords have risen: the utilization of synthetic data (Náplava and Straka, 2019; Náplava et al., 2022) and the integration of pre-trained models, particularly the mT5 model (Xue et al., 2021; Rothe et al., 2021). The use of mT5 extends to correcting grammar in various specific languages, including Ukrainian, Icelandic, and Lithuanian (Palma Gomez et al., 2023; Ingólfsdóttir et al., 2023; Stankevičius and Lukoševičius, 2022), and serves as an inspiration for other multilingual research (Kementchedjhieva and Søgaard, 2023). However, achieving substantial performance enhancements beyond training basic Transformer models necessitates further adjustments, such as the incorporation of high-quality synthetic data, additional information, or the utilization of significantly larger models.

We demonstrate that building upon similarly sized multilingual machine translation (MT) models is more effective than fine-tuning mT5 (Kementchedjhieva and Søgaard, 2023). Previous studies have shown the value of information obtained through machine translation as data or additional hypothesis (Kementchedjhieva and Søgaard, 2023; Palma Gomez et al., 2023; Lichtarge et al., 2019). We revisit the concept of utilizing zero-shot translation for error correction (Korotkova et al., 2019), developing the idea further.

We demonstrate that massively multilingual MT models can function as multilingual GEC models, and can be substantially improved further via fine-tuning to error correction data. This approach underscores the potential of multilingual MT models as an even simpler yet effective GEC system, allowing for the integration of standard practices in GEC research. In doing so, we highlight that multilingual MT models acquire valuable information for grammatical error correction and it is possible to leverage this knowledge during training.

In our work, we experiment with four languages: English, German, and Czech for comparative purposes with other multilingual studies, plus Estonian, an underexplored language in terms of error correction with a similarly limited publicly available dataset. As a result, our model achieves higher scores than work based on similar-sized mT5 models and performs competitively with even significantly larger models.

Since large language models have recently showed good performance in several NLP tasks via prompting, we also assess GPT-4's performance on the GEC task for the four included languages for comparison. While more sophisticated prompts may lead to improved results, results shown by GPT-4 are worse than state-of-the-art GEC results, and our best results also surpass its performance.

Thus, our main contributions are:

- Demonstrating the applicability of massively multilingual models as multilingual Grammar Error Correction (GEC) systems.
- Experimental results of tuning the multilingual MT models with error correction data, parallel translation data and combinations of both kinds of data.
- Achieving superior results compared to models of similar size based on widely used mT5.
- Presenting the initial $F_{0.5}$-scores for Estonian, German, and Czech using GPT-4 and updating scores for English.

2 Related work

The connection between Grammatical Error Correction (GEC) and Machine Translation (MT) has been significant since Junczys-Dowmunt et al. (2018) demonstrated an innovative approach, treating GEC as a low-resource MT task by translating from erroneous text to corrected text. This work marked the first successful implementation of neural methods in GEC and subsequently led the field to predominantly employ single-direction MT models for GEC, which has spread to other pre-trained models like T5 (Rothe et al., 2021), as mentioned in the introduction.

These methods require a substantial amount of data, leading to the necessity to generate synthetic data and the proposal of various enhancements. Grundkiewicz et al. (2019) introduced a simple reverse spell-checker idea that has been widely used (Flachs et al., 2021; Náplava and Straka, 2019). Other methods include using POS tags (Flachs et al., 2021), Wikipedia edits, or noisy corpora (Lichtarge et al., 2019). Another MT-related approach involves using data translated into a pivot language and back (Palma Gomez et al., 2023; Lichtarge et al., 2019).

In the state-of-the-art English GEC, a different paradigm emerged, with the use of sequence tagging rather than sequence generation. This approach, initially introduced by Omelianchuk et al. (2020), employs various transformer encoders for tagging errors within sentences and then replaces these parts with corrections. While this approach has proven effective for English, attempts to apply it to other languages have yielded less impressive results compared to sequence generation methods (Syvokon and Romanyshyn, 2023).

Lately several massively multilingual machine translation models have been released, including m2m100 (Fan et al., 2021), NLLB (NLLB_Team et al., 2022) and MADLAD-400 (Kudugunta et al., 2023). In our experiments we make heavy use of the NLLB models.

Finally, most recently, large language models have shown capability to correct errors via prompting (Loem et al., 2023; Fang et al., 2023; Coyne et al., 2023). Reported results mostly fall behind GEC-specific approaches.

3 Methodology

Our methodology is centred around exploiting the zero-shot translation capabilities of multilingual translation models applied to the GEC task. We also explore fine-tuning the translation models on parallel data, synthetic error data and human-annotated error correction data yielding improved performance. Finally, we explore the combination of parallel and error correction data, showing that the benefits of both tasks (translation and error correction) can be combined.

3.1 Grammatical Error Correction via Zero-shot Translation

We rely on the multilingual machine translation models' ability to produce zero-shot translation. As exemplified by Johnson et al. (2017), these models can translate between language pairs that have not been seen during training. This quality becomes relevant in the GEC context when we apply the model to monolingual "translation" (for example, English to English), (Korotkova et al., 2019).

Work by Korotkova et al. (2019) underscores the capability of monolingual zero-shot translation to rectify grammatical errors, albeit with unnecessary changes. These adjustments are often attributed to the models having learned to translate, which can cause a lack of preserving the source text's precise linguistic nuances or vocabulary. At the same time, the zero-shot corrections yield a higher recall, as they do not limit themselves with the errors that are present in the directly annotated correction data.

Based on the idea of Korotkova et al. (2019) we avoid training translation models from scratch and use pre-trained multilingual models. Using multilingual MT for GEC inherently gives us a base multilingual GEC system without further modifications. In order to focus on a narrower selection of languages we fine-tune the massively multilingual models with parallel data for the 4 languages of interest and evaluate the effect of fine-tuning. This strategy shows fruitful, especially in combination with error correction data, described in the next subsection.

3.2 Error Correction Data

In our approach, we introduce monolingual error correction data to multilingual Machine Translation (MT) models by fine-tuning the models with new monolingual translation directions. This technique aligns with the initial proposal by Junczys-Dowmunt et al. (2018), which involves training the model to translate from erroneous text to correct text. This can be achieved either through the use of grammatical error correction examples but also allows the incorporation of synthetic data.

However, when fine-tuning multilingual MT models with new data, their performance in other languages or domains often deteriorates due to catastrophic forgetting. This is likely particularly noticeable when fine-tuning large multilingual models exclusively with monolingual translation pairs. In such cases, translation quality, including zero-shot performance, may decrease significantly, leading to the loss of valuable information learned during translation training. To address this, we experiment with combining translation and synthetic error data for fine-tuning the model.

Thus, we introduce monolingual data, including synthetic and error correction data, in three distinct ways to assess the impact of synthetic pre-training and the inclusion of translation data:

1. Solely fine-tuning with GEC corpora.

2. Fine-tuning initially with monolingual synthetic data, followed by GEC corpora.

3. Fine-tuning initially with a mixture of monolingual synthetic and parallel translation examples, followed by GEC corpora.

In addition, we investigate the influence of different monolingual synthetic and parallel translation data ratios, aiming to understand their impact on model performance. This approach allows us to discern the relative benefits of each data type. Simultaneously, we explore how the multilingual aspect of our model affects its performance when trained with synthetic data in a single language or across all 4 languages and how monolingual or multilingual GEC tuning impacts the performance.

4 Experimental Setup

This section presents an overview of our experimental setup, covering data sources, models, and evaluation metrics, providing insights into the technical details of our work.

4.1 Data

We are utilizing three different types of data sources: monolingual text for generating a synthetic corpus, parallel machine translation corpora for mixed pretraining, and grammatical error correction examples for fine-tuning.

Our monolingual text data is primarily derived from NewsCrawl, which consists of text extracted from online newspapers (Kocmi et al., 2022). We randomly sample six million sentences from the latest data available. For synthetic error generation, we are using the same method proposed by Grundkiewicz et al. (2019), with the modifications and frequencies proposed by Náplava and Straka (2019). For Estonian, we use probabilities 0.6 for replacement, 0.15 for insertion and deletion, 0.05 for swap, derived from the training corpus.

For our parallel machine translation data, we merge two distinct sources: the Europarl corpus, which features parallel sentences from European Parliament Proceedings (Tiedemann, 2012), and the OpenSubtitles corpus (Lison and Tiedemann, 2016). This combination yields a dataset of two million sentences for each language pair, maintaining a balance between formal and informal text.

When it comes to grammatical error correction (GEC) examples, for English, we focus on two specific datasets. The first dataset is associated with the BEA Shared Task 2019 (Bryant et al., 2019). This particular dataset's training set comprises language learners' text sourced from the Write & Improve (W&I) corpus (Yannakoudakis et al., 2018). Additionally, for English, we also make use of the FCE corpus (Yannakoudakis et al., 2011).

For Estonian, our source of GEC examples is a language learners' corpus (UT-L2) (Rummo and Praakli, 2017) that Korotkova et al. (2019) used for testing[1]. In the case of German, we rely on the

[1] https://github.com/TartuNLP/estgec/tree/main/Tartu_L2_corpus

Corpus	Lang	Train
W&I+LOCNESS	EN	34,308
FCE	EN	28,350
UT-L2	ET	8,935
FM	DE	19,237
GECCC	CS	66,673

Table 1: Size of grammatical error correction data used for training.

Falko-Merlin (FM) dataset (Boyd, 2018). Lastly, for Czech, we use the recent Grammar Error Correction Corpus for Czech (GECCC) (Náplava et al., 2022) because it is the latest and most diverse. The specifics regarding the number of sentences employed from each dataset can be found in Table 1.

4.2 Models

We fine-tune the No Language Left Behind (NLLB) models (NLLB_Team et al., 2022) in our experiments. These models are among the latest massively multilingual models, encompassing 202 languages and demonstrating strong overall performance. We conduct all our experiments using two variants: NLLB 600M-distilled, the smallest version and NLLB 1.3B-distilled, the slightly larger model. These models are distilled from the 54-billion-parameter Mixture-of-Experts model (NLLB_Team et al., 2022). All data is also preprocessed using the NLLB normaliser and SentencePiece (Kudo and Richardson, 2018).

For fine-tuning, we employ the Fairseq toolkit (Ott et al., 2019). When fine-tuning from the NLLB model, we initialize the process with a starting learning rate of 1×10^{-7} with inverse square root scheduler, perform 4000 warmup updates to the learning rate 5×10^{-4}, using a batch size of 4096 tokens on a single GPU (AMD MI250x), with an update frequency of one. We are using Adam optimizer (Kingma and Ba, 2015). In the case of models already trained with synthetic or mixed data, we continue training with the error examples, maintaining the state of the learning rate scheduler.

We train two sets of models. For exploring the incorporation of synthetic data, we train models involving 1.5M sentences per language for 150k updates. We train the final models with 6M sentences per language and train the models for 600k updates for multilingual synthetic training and 150k for monolingual. We perform all GEC fine-tuning for 25 epochs and pick the best epoch checkpoint based on the development set using GEC scores specified in the next section. However, it has been found that mixing GEC data with synthetic while fine-tuning helps, our initial experiments suggested otherwise. It needs further investigation, but for now, we opted for exclusively fine-tuning with GEC data.

For comparison, we also measure the performance of GPT-4 (OpenAI, 2023) using the prompt by Coyne et al. (2023). See Appendix A for the exact prompts and other details.

4.3 Evaluation

We employ two distinct scorers and evaluate our models using six test sets. For the English language, which offers a multitude of corpora and test sets, we selected two test sets and corresponding scorers officially paired together. We use the not publicly open W&I+LOCNESS test set (Bryant et al., 2019), along with the ERRANT scorer (Bryant et al., 2017). Additionally, we utilize the combination of the CoNLL-2014 dataset (Ng et al., 2014) and the MaxMatch (M2) scorer (Dahlmeier and Ng, 2012) for the same reason.

The evaluation of the Estonian language presents a unique challenge. The only previous work that includes Estonian done by Korotkova et al. (2019) relied on the entire UT-L2 corpus (Rummo and Praakli, 2017) for evaluation. This poses difficulties for direct comparisons since we also intend to use the corpus for training. We opted to use the entire corpus for training and dedicate the annotated Estonian learner language corpus (Est-L2)[2] for evaluation with modified MaxMatch scorer[3], which considers special annotations from the Est-L2 corpus concerning word order mistakes.

For German and Czech, we use standard test sets and the out-of-the-box M2 scorer. Specifically, for German, we use the Falko-Merlin (FM) corpus (Boyd, 2018) and the older AKCES corpus (Náplava and Straka, 2019), which most other works have used. Additionally, we employ the newer and more extensive GECCC test set (Náplava et al., 2022) for Czech.

For evaluation, we tokenized the text using SpaCy[4] in the standard configuration for English and German and Stanza for Estonian and Czech (Qi et al., 2020).

[2] https://github.com/tlu-dt-nlp/m2-corpus/

[3] https://github.com/TartuNLP/estgec/tree/main/M2_scorer_est

[4] https://spacy.io/api/tokenizer

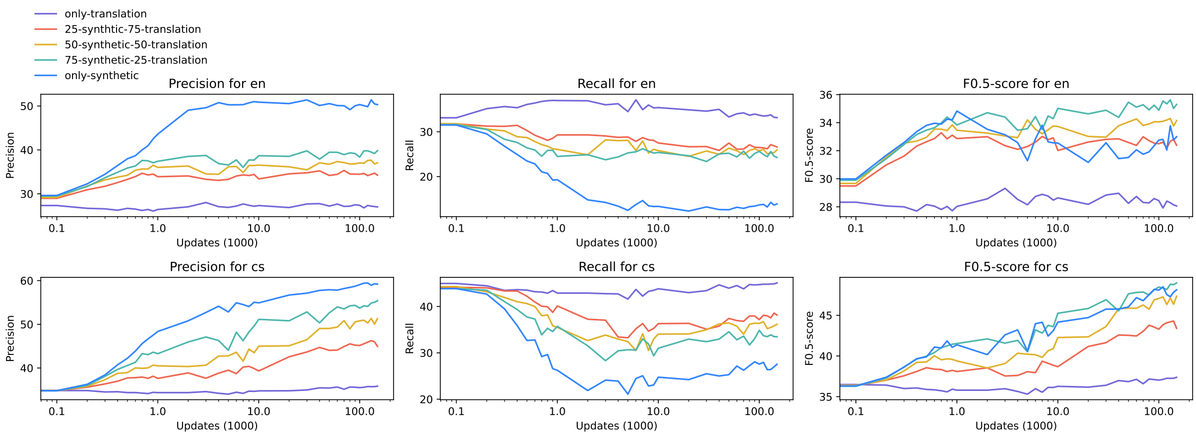

Figure 1: Precision, recall and $F_{0.5}$-score for models trained with only synthetic, only translation or mixed data evaluated on English W&I+OCNESS and Czech GECCC development sets. Models are trained with 1.5M sentences per language.

	EN	ET	DE	CS
NLLB (zero-shot)	39.82	39.72	51.6	44.04
NLLB + 1-lang GEC	64.78	51.24	70.9	64.44
NLLB + 4-lang GEC	66.29	51.89	70.01	63.19
NLLB + 1-lang synthetic + 1-lang GEC	66.12	60.58	72.63	**68.08**
NLLb + 4-lang synthetic + 1-lang GEC	66.60	58.76	72.89	67.35
NLLb + 4-lang synthetic + 4-lang GEC	66.81	59.64	73.32	66.63
NLLB + 4-lang mixed + 1-lang GEC	66.70	60.05	73.72	67.14
NLLB + 4-lang mixed + 4-lang GEC	**67.35**	**60.69**	**73.94**	66.32

Table 2: Comparison of $F_{0.5}$-scores for models trained using various synthetic and GEC training strategies. The test sets are W&I+LOCNESS for English, Est-L2 for Estonian, FM for German, and GECCC for Czech. Models are trained with 6M sentences per language for around 2.5 epochs

5 Results

We first describe the results of our experiments related to mixing data during pre-training, then show how different data and pre-training affect the model's behaviour and, lastly, we benchmark our models with comparable and state-of-the-art research solutions and GPT-4 performance.

5.1 Pre-training Scenarios

When training the NLLB model using only synthetic monolingual data in four different languages, we observe a significant increase in precision. However, this improvement in precision comes at the cost of reduced recall, which rapidly drops (see Figure 1). Interestingly, the recall starts to slowly recover after the initial drop.

Continuing training with translation data exclusively results in relatively stable precision and recall. There is a slight increase in recall for Czech but a decrease for English. This could be due to the balanced nature of the data, with proportionally less English and more Czech compared to NLLB training.

When we combine translation data and monolingual synthetic examples, we achieve precision and recall values that fall between the two previous scenarios. While precision is not as high as in the monolingual synthetic scenario, recall remains higher. Based on $F_{0.5}$-scores, for these languages, a ratio of 75% monolingual synthetic data and 25% parallel data seems to yield the best results out of the three mixed, only synthetic and only parallel translation data, except for Estonian, where using more parallel data leads to better results (see the Appendix B for more details).

Moreover, it seems that overall Estonian and Czech benefit more from longer training, while German and especially English improve at a slower pace after rather short training, which indicates that the languages have different optimal pre-training durations.

	W&I+LOCNESS			CoNLL-2014		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
GPT-4 (zero-shot)	56.68	71.57	59.14	61.96	59.82	61.52
Coyne et al. (2023) GPT-4 2-shot	-	-	52.79	-	-	-
Loem et al. (2023) GPT-3 16-shot	-	-	57.41	-	-	57.06
Náplava and Straka (2019)	-	-	69.00	-	-	63.40
Rothe et al. (2021) xxl+cLANG8	-	-	75.88	-	-	68.75
Omelianchuk et al. (2020)	79.4	57.2	73.7	78.2	41.5	66.5
Qorib et al. (2022)	86.6	60.9	79.9	81.48	43.78	69.51
Rothe et al. (2021) base	-	-	60.2	-	-	54.10
Rothe et al. (2021) xxl	-	-	69.83	-	-	65.65
NLLB 600M-distilled	37.05	56.82	39.82	48.7	49.15	48.79
NLLB 1.3B-distilled	40.28	57.68	42.87	51.8	49.04	51.22
NLLB 600M-distilled + 4-lang GEC	66.99	63.66	66.29	66.29	50.68	62.45
NLLB 1.3B-distilled + 4-lang GEC	67.41	66.89	67.31	66.07	54.28	63.32
NLLB 600M-distilled + mixed + 4-lang GEC	67.84	65.43	67.35	67.14	51.8	63.39
NLLB 1.3B-distilled + mixed + 4-lang GEC	70.04	67.09	69.43	68.8	54.08	65.25

Table 3: Main results for the English language calculated with ERRANT scorer for W&I+LOCNESS and MaxMatch for CoNLL. Work by Rothe et al. (2021) is multilingual, except for the version trained with cLANG8. Works by Omelianchuk et al. (2020); Qorib et al. (2022) represent other top methods, and Náplava and Straka (2019) uses Transformer pre-trained with synthetic and fine-tuned with GEC data. GPT-4 scores are calculated in mid-October.

5.2 Fine-tuning with error correction examples

When analysing the $F_{0.5}$-scores of our NLLB 600M-distilled models, it becomes evident that pre-training with synthetic data substantially enhances performance, and the choice of training data type exerts a notable impact on the model's effectiveness across various languages (refer to Table 2). A consistent trend emerges: for all languages except Czech, the most favorable results are achieved when the initial training phase combines monolingual synthetic data with parallel translation examples, followed by subsequent multilingual fine-tuning with GEC data.

The results further highlight the distinct behavior of the Czech language under multilingual training conditions. Despite having the largest and most diverse training corpus, Czech tends to experience adverse effects from multilingual training across all scenarios. In contrast, English, with a training corpus of comparable size, consistently benefits from multilingual training. The case of German, which possesses a smaller GEC corpus, also reveals improved performance with multilingual training. However, Estonian, despite a smaller corpus, does not display a clear preference for multilingual training. Interestingly, languages that lean less towards multilinguality, such as Estonian and Czech, exhibit more substantial performance gains from synthetic data compared to using only GEC examples. This suggests that high-resource languages in the context of MT derive substantial benefits from multilinguality, while the size of the GEC corpus appears to have a lesser influence on the overall outcome. Additionally, languages less prominently represented in the MT model require additional support from synthetic data, though this may be negatively impacted by the inclusion of multilingual data.

5.3 Final results

In this section, we will give the final results for all languages in the context of other works.

For English, when we compare our best models to the mT5-based model, which has received similar training in error correction, is multilingual and has a comparable number of parameters, we outperform it simply by fine-tuning our NLLB 600M-distilled model with GEC data in four languages, as highlighted in Table 3. Additional training with synthetic data increases the performance further. Our 1.3B-distilled model achieves results nearly as high as the model based on mT5 xxl, which has ten times more parameters.

We also recalculated scores for English with GPT-4 (OpenAI, 2023), utilizing the same prompt that Coyne et al. (2023) employed, albeit without

	Est-L2		
	P	R	$F_{0.5}$
GPT-4 (zero-shot)	72.74	44.72	64.64
NLLB 600M-distilled (zero-shot)	40.44	37.09	39.72
NLLB 1.3B-distilled (zero-shot)	43.55	41.32	43.09
NLLB 600M-distilled + 4-lang GEC	57.83	36.77	51.89
NLLB 1.3B-distilled + 4-lang GEC	60.85	44.73	56.75
NLLB 600M-distilled + mixed + 4-lang GEC	66.57	44.85	60.69
NLLB 1.3B-distilled + mixed + 4-lang GEC	69.78	50.58	64.85

Table 4: Main results for the Estonian language calculated using MaxMatch scorer. GPT-4 scores were calculated in mid-October.

	Falko-Merlin		
	P	R	$F_{0.5}$
GPT-4 (zero-shot)	67.75	68.46	67.89
Náplava and Straka (2019) (210 param)	78.21	59.94	73.71
Rothe et al. (2021) base	-	-	69.21
Rothe et al. (2021) xxl	-	-	75.96
Kementchedjhieva and Søgaard (2023) base	76.0	61.5	72.6
Kementchedjhieva and Søgaard (2023) large	76.4	64.3	73.6
NLLB 600M-distilled (zero-shot)	40.44	37.09	39.72
NLLB 1.3B-distilled (zero-shot)	43.66	41.52	43.22
NLLB 600M-distilled + 4-lang GEC	72.3	62.12	70.01
NLLB 1.3B-distilled + 4-lang GEC	74.05	65.74	72.22
NLLB 600M-distilled + mixed + 4-lang GEC	76.76	64.46	73.94
NLLB 1.3B-distilled + mixed + 4-lang GEC	77.65	67.0	75.26

Table 5: Main results for the German language calculated using MaxMatch scorer. Work by Náplava and Straka (2019) uses a Transformer model with synthetic pre-training and fine-tuning with GEC corpus. Rothe et al. (2021); Kementchedjhieva and Søgaard (2023) models are multilingual and based on mT5 model. GPT-4 scores are calculated in mid-October.

presenting examples, which they noted enhances performance. Our results show a substantial improvement.

For Estonian, the only other work we can compare us to is GPT-4. GPT-4 shows a similar $F_{0.5}$-score to our best model but exhibits notably lower recall and higher precision. However, it outperforms our model when compared to zero-shot translation, as illustrated in Table 4.

For German, we achieve near state-of-the-art results. Only an mT5-based model that is ten times larger than our model manages to achieve a slightly higher $F_{0.5}$-score, as indicated in Table 5.

When comparing our NLLB 600M-distilled model, fine-tuned exclusively with GEC data, to the base model from Rothe et al. (2021), our only GEC fine-tuned model surpasses their work, similar to English. However, Kementchedjhieva and Søgaard (2023) utilized pre-training with cleaned Lang-8 data, containing 114K sentence pairs (Rothe et al., 2021), and gained an additional performance boost from roundtrip translation. Although their work achieved higher scores compared to our model fine-tuned with GEC data alone, when we incorporate pre-training, our 600M-distilled model outperforms theirs. The same trend is observed in the comparison between mT5 large and our 1.3B-distilled model. Our model even surpasses their XL model, which is almost 3 times larger.

For Czech, we lack directly comparable multilingual models. Our approach uses the latest and slightly larger corpus GECCC, which is more diverse and includes more data, particularly in the informal web domain. This makes it challenging to assess how it affects performance on the AKCES test set. Nevertheless, our best models outperform similarly-sized multilingual models from previous studies (see Table 6).

	GECCC			AKCES		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
GPT-4 (zero-shot)	72.74	44.72	64.64	76.73	71.9	75.72
Náplava and Straka (2019) (210M param)	-	-	-	83.75	68.48	80.17
Náplava et al. (2022) (210M param)	-	-	72.96	-	-	-
Rothe et al. (2021) base	-	-	-	-	-	71.88
Rothe et al. (2021) xxl	-	-	-	-	-	83.15
Kementchedjhieva and Søgaard (2023) base	-	-	-	79.4	65.0	76.0
Kementchedjhieva and Søgaard (2023) large	-	-	-	81.9	70.5	79.3
Kementchedjhieva and Søgaard (2023) xl	-	-	-	82.0	70.8	79.5
NLLB 600M-distilled (zero-shot)	43.7	45.43	44.04	39.54	51.76	41.5
NLLB 1.3B-distilled (zero-shot)	45.79	49.25	46.44	42.6	56.2	44.76
NLLB 600M-distilled + 4-lang GEC	65.33	55.88	63.19	77.02	69.17	75.31
NLLB 1.3B-distilled + 4-lang GEC	68.45	58.33	66.16	77.92	72.32	76.73
NLLB 600M-distilled + mixed + 4-lang GEC	68.9	57.67	66.32	79.94	70.94	77.96
NLLB 1.3B-distilled + mixed + 4-lang GEC	71.19	60.71	68.81	81.69	74.8	80.21

Table 6: Main results for the Czech language calculated using MaxMatch, works by Náplava et al. (2022); Náplava and Straka (2019) are Czech-specific Transformer models pre-trained with synthetic data and fine-tuned with GEC corpus, models by Rothe et al. (2021); Kementchedjhieva and Søgaard (2023) are multilingual and based on mT5 model. GPT-4 scores are calculated in mid-October.

It is worth noting that our models struggled with the GECCC test set, primarily due to difficulties with web text, such as issues related to repeated punctuation marks. This data might not have been adequately represented during translation training or fine-tuning. We did not add any specific length penalty other than default settings but it could be useful to stop models from over-repeating symbols.

6 Discussion

Our tuned multilingual MT models consistently outperform mT5-based approaches. In addition to mT5-based works, our approach outperforms GPT-4 in a zero-shot setting for all the languages we tested, with a larger margin for English, German, and Czech and more comparable performance for Estonian. However, GPT-4, being a large general-purpose model, is not practical for real-time GEC due to its current quality, availability, and speed. Therefore, we have not explored few-shot prompts or fine-tuning options for ChatGPT at this time.

Multilingual training presents both advantages and complexities. It demonstrates effectiveness for languages that are well-represented in the translation model, while languages with limited representation may not experience such clear benefits. This disparity may be attributed to their weaker zero-shot performance, indicating that they have more to learn from synthetic data. To address this, a potential solution could involve more extensive pre-training or initial training with select translation data. This approach may negatively impact other languages, as indicated by decreasing English and German scores for zero-shot translation with balanced translation training.

Our work focused on one MT system covering approximately 200 languages as a starting point for building a GEC system. Future research can explore different models and sizes, improve data balance during pre-training, use better synthetic data, and refine fine-tuning strategies. A recent study, MADLAD-400 (Kudugunta et al., 2023), has already covered twice as many languages, indicating a promising direction for further investigation and language coverage.

7 Conclusion

We propose a simple approach for a multilingual GEC system, simplifying the creation of non-English GEC solutions. Through the use of multilingual machine translation models supplemented with synthetic and error correction data, we have presented an effective approach to enhancing GEC performance. Our results reveal the superiority of this method, with our multilingual model consistently outperforming similar-sized models and even competing with larger counterparts.

8 Limitations

While our research sheds light on the effectiveness of a single multilingual machine translation model for error correction across four languages and two model sizes, several limitations should be acknowledged. First, our findings primarily apply to the model configurations tested, and we can reasonably infer that larger models may yield enhanced performance. However, a comprehensive validation of this assumption is beyond the scope of our work and computational capacity.

Furthermore, our study prioritizes specific languages and settings, leaving room for expanded inclusivity and validating the method with other languages. Testing the model across a broader range of languages and fine-tuning configurations would provide a more comprehensive understanding of its utility and potential limitations.

Additionally, our investigation does not encompass an exhaustive hyperparameter search and each experiment was executed only once. Conducting multiple runs could provide more robust and reliable results. Also, our work does not encompass a detailed exploration of the impact of retaining a portion of pre-training data during GEC fine-tuning. These aspects present avenues for future research and further refinement of the model's performance.

References

Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 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 (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation.

Simon Flachs, Felix Stahlberg, and Shankar Kumar. 2021. Data strategies for low-resource grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 117–122, Online. Association for Computational Linguistics.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Svanhvít Lilja Ingólfsdóttir, Petur Ragnarsson, Haukur Jónsson, Haukur Simonarson, Vilhjalmur Thorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-level grammatical error correction using synthetic and curated corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316, Toronto, Canada. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

Yova Kementchedjhieva and Anders Søgaard. 2023. Grammatical error correction through round-trip machine translation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2208–2215, Dubrovnik, Croatia. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elizaveta Korotkova, Agnes Luhtaru, Maksym Del, Krista Liin, Daiga Deksne, and Mark Fishel. 2019. Grammatical error correction and style transfer via zero-shot monolingual translation. *CoRR*, abs/1903.11283.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.

Jakub Náplava and Milan Straka. 2019. Grammatical error correction in low-resource scenarios. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 346–356, Hong Kong, China. Association for Computational Linguistics.

Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics*, 10:452–467.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

NLLB_Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Frank Palma Gomez, Alla Rozovskaya, and Dan Roth. 2023. A low-resource approach to the grammatical error correction of Ukrainian. In *Proceedings* of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 114–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Muhammad Reza Qorib, Seung-Hoon Na, and Hwee Tou Ng. 2022. Frustratingly easy system combination for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974, Seattle, United States. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Ingrid Rummo and Kristiina Praakli. 2017. TÜ eesti keele (võõrkeelena) osakonna õppijakeele tekstikorpus [the language learner's corpus of the department of estonian language of the university of tartu]. In *EAAL 2017: 16th annual conference Language as an ecosystem, 20-21 April 2017, Tallinn, Estonia: abstracts, 2017, p. 12-13*.

Lukas Stankevičius and Mantas Lukoševičius. 2022. Towards lithuanian grammatical error correction.

Oleksiy Syvokon and Mariana Romanyshyn. 2023. The UNLP 2023 shared task on grammatical error correction for Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 132–137, Dubrovnik, Croatia. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Helen Yannakoudakis, Øistein Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31.

Helen Yannakoudakis, Ted Briscoe, and Ben 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.

A GPT-4 Prompts

We used the prompts found to be the best by (Coyne et al., 2023) and added the non-English language for clarification. The exact prompts used are the following:

```
Reply with a corrected version of the input sentence with all grammatical and spelling errors fixed. If there are no errors, reply with a copy of the original sentence.

Input sentence: {sentence}
Corrected sentence:
```

```
Reply with a corrected version of the input sentence in Estonian with all grammatical and spelling errors fixed. If there are no errors, reply with a copy of the original sentence.

Estonian input sentence: {sentence}
Corrected Estonian sentence:
```

```
Reply with a corrected version of the input sentence in German with all grammatical and spelling errors fixed. If there are no errors, reply with a copy of the original sentence.

German input sentence: {sentence}
Corrected German sentence:
```

```
Reply with a corrected version of the input sentence in Czech with all grammatical and spelling errors fixed. If there are no errors, reply with a copy of the original sentence.

Czech input sentence: {sentence}
Corrected Czech sentence:
```

We added the unchanged sentence when the API responded with a content filter. It did not happen excessively but is still a notable disadvantage for the system reducing the quality of error correction.

B Pre-training Experiment Extended

Figure 2 provides a visual representation of the pre-training process for models across all four languages. It highlights how the model's performance changes when using different types of data: solely synthetic data, translation training with selected languages, or a combination of these data sources while maintaining consistent sentence quantities for each language.

The graph illustrates that, as pre-training progresses, English and German exhibit a plateau in performance improvement, indicating that they do not continue to advance rapidly. However, for Estonian and Czech, there is a clear and continued upward trajectory, indicating rapid improvement in these languages.

Additionally, a noticeable spike in the $F_{0.5}$-score is observed for models trained with synthetic data in German and English. This spike is marked by a significant increase in precision, with recall not yet showing a corresponding decrease.

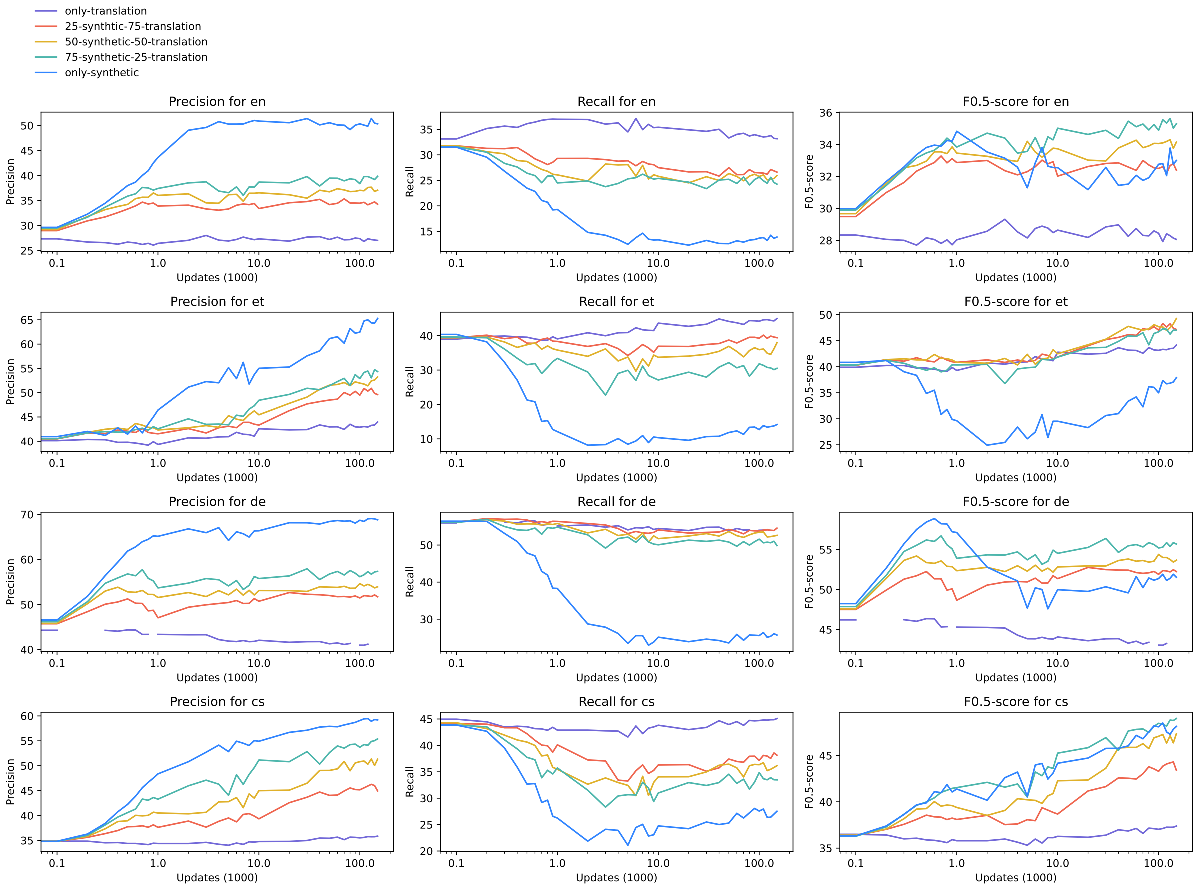

Figure 2: Precision, recall and $F_{0.5}$-score for only synthetic, only parallel and mixed data with different ratios for English W&I+OCNESS, Estonian Est-L2, German FM and Czech GECCC development sets measured with ERRANT scorer for English and MaxMatch scorer for other languages. Models are trained with 1.5M sentences per language for 150k updates with batch size 4096 tokens.