UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language

Oleksiy Syvokon
oleksiy.syvokon@gmail.com

Olena Nahorna
Grammarly
olena.nahorna@grammarly.com

Pavlo Kuchmiichuk
pavlo.kuchmiichuk@gmail.com

Nastasiia Osidach
Grammarly
nastasiya.osidach@grammarly.com

Abstract

We present a corpus professionally annotated for grammatical error correction (GEC) and fluency edits in the Ukrainian language. We have built two versions of the corpus – GEC+Fluency and GEC-only – to differentiate the corpus application. We collected texts with errors (33,735 sentences) from a diverse pool of contributors, including both native and non-native speakers. The data cover a wide variety of writing domains, from text chats and essays to formal writing. Professional proofreaders corrected and annotated the corpus for errors relating to fluency, grammar, punctuation, and spelling. This corpus can be used for developing and evaluating GEC systems in Ukrainian. More generally, it can be used for researching multilingual and low-resource NLP, morphologically rich languages, document-level GEC, and fluency correction. To test the effectiveness of our corpus, we trained a basic but reasonable baseline model. The corpus is publicly available at https://github.com/grammarly/ua-gec.

1 Introduction

Grammatical error correction (GEC) is a task of automatically detecting and correcting grammatical errors in written text. GEC is typically limited to making a minimal set of grammar, spelling, and punctuation edits so that the text becomes free of such errors. Fluency correction is an extension of GEC that allows for broader sentence rewrites to make a text more fluent—i.e., sounding natural to a native speaker (Sakaguchi et al., 2016).

Over the past decade, NLP researchers have been primarily focused on English GEC, where they indeed made substantial progress: $F_{0.5}$ score of the best-performing model in the CoNLL-2014 shared task has increased from 37.33 in 2014 to 68.75 in 2022 (Ng et al., 2014; Rothe et al., 2021). Multiple available datasets and shared tasks were a major contributing factor to that success.

However, languages other than English still present a set of challenges for current NLP methods. Mainstream models developed with English in mind are suboptimal for morphologically rich languages as well as languages with differing grammar (Tsarfaty et al., 2020; Ravfogel et al., 2018; Hu et al., 2020; Ahmad et al., 2019). The common issue is a scarcity of data—particularly high-quality annotated data that could be used for evaluation and fine-tuning.

More recently, the NLP community has started to pay more attention to non-English NLP (Ruder, 2020). This positive recent trend manifests itself in the creation of new GEC corpora for mid- and low-resource languages: German, Czech, and Spanish, to name a few (Boyd, 2018; Náplava and Straka, 2019; Davidson et al., 2020). These datasets are important to expand NLP research to new languages and to explore new ways of training models in a low-resource setting.

Furthering that trend, we present a corpus annotated for grammatical errors and fluency in the Ukrainian language: UA-GEC. We first collected texts from a diverse pool of writers, both native and non-native speakers. The corpus covers a wide variety of domains: essays, social media posts, chats, formal writing, and more. We recruited professional proofreaders to correct errors related to grammar, spelling, punctuation, and fluency. Our corpus is open source for the community[2] under the CC-BY 4.0 license.

To summarize, our contributions are as follows:

- For the first time, diverse texts in Ukrainian are collected and annotated for grammatical, punctuation, spelling, and fluency errors.
- The corpus is released for public use under the CC-BY 4.0 licence.
- A baseline model is trained.

[2] https://github.com/grammarly/ua-gec

Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 96–102
May 5, 2023 ©2023 Association for Computational Linguistics

Split	Writers	Texts	Sentences	Tokens	Annotations	Error rate
Train	752	1,706	31,038	457,017	38,383	8.1%
Test	76	166	2,697	43,601	7,865	9.0%
TOTAL	828	1,872	33,735	500,618	46,248	8.2%

Table 1: The GEC+Fluency corpus statistics. Test split is independently annotated by two annotators (*Error rate* is the average of the two in this case)

.

Split	Writers	Texts	Sentences	Tokens	Annotations	Error rate
Train	752	1,706	31,046	457,004	29,390	6.1%
Test	76	166	2,704	43,605	5,931	6.8%
TOTAL	828	1,872	33,750	500,609	35,321	6.3%

Table 2: The GEC-only corpus statistics. Test split is independently annotated by two annotators (*Error rate* is the average of the two in this case)

.

2 Data collection

In this section, we describe the collection of texts with errors in the Ukrainian language. Section 3 will explain the annotation details.

2.1 Statistics

Parameter		Writers	Sent.
Native	Yes	600	27,646
	No	238	6,072
Gender	Female	537	18,520
	Male	288	14,212
	Other	9	986
Background	Technical	291	13,654
	Humanities	356	12,819
	Natural sci.	39	1,389
	Other	168	5,856

Table 3: Profile of respondents

We have collected 1,872 texts (33,735 sentences) written by 492 unique contributors. The average length of a text snippet is 18 sentences.

We partition the corpus into training and test sets. Each split consists of texts written by a randomly chosen disjoint set of people: all of a particular person's writing goes to exactly one of the splits. To better account for alternative corrections, we annotated the test set two times (Bryant and Ng, 2015). The resulting statistics are shown in Table 2.

In order to collect the data, we created an online form for text submission. All respondents who contributed to the data collection were volunteers. To attract a socially diverse pool of authors, we shared the form on social media. It contained a list of questions related to gender, native tongue, region of birth, and occupation, making it possible to further balance subcorpora and tailor them so they meet the purpose of various NLP tasks. Table 3 illustrates the profile of respondents based on some of these parameters.

2.2 Collection tasks

The online form offered a choice of three tasks: 1) writing an essay; 2) translating a fictional text fragment into Ukrainian; 3) submitting a personal text. Our goal was to collect a corpus of texts that would reflect errors typically made by native and non-native speakers of Ukrainian. Therefore, before performing a task, the respondents were asked not to proofread their texts as well as to refrain from making intentional errors. Each task varied in the number of requirements.

Write an essay on the topic "What's your favorite animal?" Genre: fictional. In the essay, state: what your favorite animal is; what it looks like; why you like this particular animal; whether you would like to keep it at home. Volume: about 15 sentences.
Write a letter of complaint. Recipient: a restaurant administrator. Genre: formal. In the letter, state: the date of your visit to the restaurant; the reason for your complaint; your suggestions about how the restaurant could improve its service. Volume: about 15 sentences.

Table 4: Examples of the essay prompts. In total, there were 20 prompts in the *Essay* task.

Essays. Respondents were offered one of twenty essay topics, each stipulating the genre, length, and structure of the essay. We chose from among the most common topics for essays (e.g., "What was your childhood dream?") not requiring a profound knowledge of a certain subject, which made it easy for the respondents to produce texts. Each essay was supposed to be written in accordance with one of four genres: formal, informal, fictional, or journalistic. The scientific genre was excluded as a potential writing blocker due to its inherent complexity. Specification of the genre allowed us to moderate the heterogeneity of the corpus. Besides topic and genre requirements, each task description contained prompts—i.e., prearranged points to cover in the text that facilitated text production. Refer Table 4 for essay prompts examples.

Translation of fictional texts. Fictional text fragments were taken from public domain books written by classic authors in five languages: English, French, German, Polish, and Russian. The rationale behind suggesting translation from a range of foreign languages was to diversify the errors made by respondents as a result of L1 interference.

Personal texts. Unlike the aforementioned tasks, personal text submission was not explicitly regulated: respondents could submit texts of any genre, length, or structure. However, no more than 300 sentences submitted by a unique person were added to the corpus. This was done to balance the corpus from an idiolect perspective.

UA-GEC is mostly composed of personal texts (62%); fictional texts translations rank second (35%), and essays are the least numerous (3%).

3 Data annotation

We enrolled two annotators on the project, both native speakers of Ukrainian with a degree in Ukrainian linguistics. One of them was a freelance editor, and the other was a teacher of Ukrainian.

In order to diversify the type of tasks one can perform using the corpus, we released two versions of UA-GEC: GEC+Fluency and GEC-only. The former surfaces spelling, punctuation, grammar errors as well as errors associated with unnatural-sounding sentence elements. The latter captures only GEC errors, which makes it possible to perform tasks that are narrower and more objective in scope.

GEC+Fluency. The annotation process encompassed two sequential subtasks: error correction followed by error labeling. We found that the given annotation design was more efficient than performing error correction and labeling in a combined mode as it would increase the cognitive load of the task.

GEC-only. After having the data fully edited and labeled, we programmatically removed edits labeled as Fluency and had annotators review the remaining annotations to make sure Fluency-dependent edits were still valid and correct suggestions that no longer made sense.

3.1 Annotation format

The categorized errors in the processed data are marked by the following in-text notations: `{error=>edit:::Tag}`, where `error` and `edit` stand for the text item before and after correction, respectively, and `Tag` denotes an error category. Table 5 lists example sentences annotated for each high-level category.

Besides error correction and labeling, the annotators were asked to identify sensitive content—i.e., sentences containing pejorative lexis or perpetuating bias related to race, gender, age, etc. Such sentences are marked in the metadata, which enables simple data filtering to debias it by the stated criteria. The GitHub repository contains a detailed description of the annotation scheme along with a Python library to process the corpora.

3.2 Error categories

Our label set includes four high-level categories: punctuation, spelling, grammar and fluency. Additionally, grammar and fluency suggestions are further divided into fine-grained categories. Table 6 provides a detailed description of error categories and Table 7 demonstrates the error distribution by category.

`Spelling` accounts for 19% of all corrections. This is similar to RULEC-GEC (Rozovskaya and Roth, 2019), where the portion of spelling errors is 21.7%. `Punctuation` edits (43%) are more frequent than in other corpora (for example, in the W&I corpus (Bryant et al., 2019), `Punctuation` is 17%). We explain this by the fact that in the Ukrainian language, punctuation rules are sharply defined; thus, a lot of punctuation marks are frequently misused, especially commas. Also, there were a large number of typographical fixes, like replacing a dash ("-") with an em-dash ("—") where appropriate. Grammatical errors (`G/`) accounts for 14.4% of all errors.

Fluency. The fluency category (`F/`) embraces error types that have to do with the inaccurate use of lexical or structural units. Specifically, such edits relate to the correction of miscollocations and

Error type	Example
Grammar	Він {ходимо=>ходить:::G/Number} до школи. `He  {go=>goes:::Grammar} to school.`
Spelling	Він {хотв=>хотів:::Spelling} поговорити. `He {wnted=>wanted:::Spelling} to talk.`

Punctuation	Ти будеш завтра вдома {=>?:::Punctuation}
	`Are you going to be home tomorrow {=>?:::Punctuation}`

Fluency	{Існуючі =>Теперішні:::F/Style} ціни дуже високі. `{Existing=>Current:::Fluency} prices are very high.`

Table 5: Examples of annotation in each error category

Error type	Description
Grammar-related errors	
G/Case	incorrect usage of case of any notional part of speech
G/Gender	incorrect usage of gender of any notional part of speech
G/Number	incorrect usage of number of any notional part of speech
G/Aspect	incorrect usage of verb aspect
G/Tense	incorrect usage of verb tense
G/VerbVoice	incorrect usage of verb voice
G/PartVoice	incorrect usage of participle voice
G/VerbAForm	incorrect usage of an analytical verb form
G/Prep	incorrect preposition usage
G/Participle	incorrect usage of participles
G/UngrammaticalStructure	digression from syntactic norms
G/Comparison	incorrect formation of comparison degrees of adj. and adverbs
G/Conjunction	incorrect usage of conjunctions
G/Other	other grammatical errors
Fluency-related errors	
F/Style	style errors
F/Calque	word-for-word translation from other languages
F/Collocation	unnatural collocations
F/PoorFlow	unnatural sentence flow
F/Repetition	repetition of words
F/Other	other fluency errors

Table 6: Description of Grammar and Fluency fine-grained categories

calques, words inappropriate from a style perspective, rewriting syntactic structures that contain dysfluencies (repetitions, redundancies, etc.) or simply sound awkward to a native speaker.

`Fluency` accounts for 23.6% of all errors. This may be attributed to the fact that around 30% of respondents were not native Ukrainian speakers and therefore used a lot of calques, both lexical and structural, from other languages. Another reason is style correction: annotators corrected non-standard language into standard one to make the text sound more fluent and natural.

3.3 Inter-annotator agreement

Pass 1	Pass 2	Error rate	Unchanged
Ann. A	Ann. B	2.9%	64%
Ann. B	Ann. A	1.2%	75%

Table 8: Inter-annotator agreement based on the second-pass proofreading. *Error rate* is the density of annotations made on the already corrected text. *Unchanged* is the percentage of sentences that have not been changed on the second pass.

We follow the Rozovskaya and Roth (2010) setup for computing the inter-annotator agreement. A

Error type	Total	%	Per 1000 tokens
Grammar (all)	6,682	14.4	11.9
Fluency (all)	10,924	23.6	19.4
Spelling	8,771	19.0	15.6
Punctuation	19,871	43.0	35.3
F/Calque	2,397	5.2	4.3
F/Collocation	459	1.0	0.8
F/Other	245	0.5	0.4
F/PoorFlow	3,477	7.5	6.2
F/Repetition	621	1.3	1.1
F/Style	3,725	8.1	6.6
G/Aspect	92	0.2	0.2
G/Case	2,536	5.5	4.5
G/Comparison	135	0.3	0.2
G/Conjunction	417	0.9	0.7
G/Gender	539	1.2	1.0
G/Number	409	0.9	0.7
G/Other	236	0.5	0.4
G/PartVoice	99	0.2	0.2
G/Participle	2	0.0	0.0
G/Particle	60	0.1	0.1
G/Prep	542	1.2	1.0
G/Tense	223	0.5	0.4
G/Ungrammatical Structure	1,046	2.3	1.9
G/VerbAForm	52	0.1	0.1
G/VerbVoice	294	0.6	0.5
TOTAL	46,248	100.0	82.1

Table 7: Error distribution by category

text that was corrected by one annotator is passed to the other annotator. *Agreement* then is the percentage of sentences that did not require any changes during the second pass. This metric is important, given that our goal is to make a sentence well-formed, no matter whether the annotators propose the same changes (Rozovskaya and Roth, 2019). We run this evaluation on a set of 200 sentences. Table 8 shows that 64% of sentences corrected by Annotator A remained unchanged after the Annotator B's pass. The error rate has dropped from 7.1% to 2.9% errors. Similarly, Annotator A that proofreads after Annotator B leaves 75% of sentences unchanged.

This inter-annotator agreement (64%/75% of unchanged sentences) is in line with other GEC corpora: for English the reported numbers are 37%/59%, for Russian they are 69%/91% (Rozovskaya and Roth, 2010, 2019).

3.4 Comparison to other GEC datasets

Table 9 lists statistics of our corpus in relation to some similar GEC corpora in other languages.

Language	Corpus	Sent.	Er.
English	Lang-8	1,147,451	14.1
	NUCLE	57,151	6.6
	FCE	33,236	11.5
	W&I+L	43,169	11.8
	JFLEG	1,511	
	CWEB	13,574	1.74
Czech	AKCES-GEC	47,371	21.4
German	Falko-MERLIN	24,077	16.8
Romanian	RONACC	10,119	
Russian	RULEC-GEC	12,480	6.4
Spanish	COWS-L2H [3]	12,336	
Ukrainian	**UA-GEC**	**33,735**	**8.2**

Table 9: Statistics of related GEC corpora. *Er.* is the error rate, in percent. This work is highlighted in bold.

4 Model

To prove the utility of our dataset, we trained a simple baseline model. We fine-tuned mBART-50-large (Tang et al., 2021) on the UA-GEC train data without any preprocessing or data augmentation, similarly to (Katsumata and Komachi, 2020).

The model was fine-tuned for 3 epochs using Adam optimizer with a learning rate of 5e-5 and batch size of 8. We used greedy decoding. The full training cycle takes around 3 hours on a single Nvidia P100 GPU.

4.1 Results

Table 10 shows the results of our baseline model on the test set.

Task	Precision	Recall	$F_{0.5}$
GEC only	0.7706	0.5004	0.6955
GEC+Fluency	0.6996	0.4159	0.6156

Table 10: Results of the baseline model on the test set.

5 Conclusion

We release the first professionally annotated corpus. We hope it will facilitate further development of grammatical error correction in the Ukrainian language. The corpus is made publicly available at `https://github.com/grammarly/ua-gec` under the CC-BY 4.0 license.

[3] COWS-L2H statistics is for March 2021

Limitations

UA-GEC has some limitations that must be taken into account.

First, the dataset has been annotated with only two annotators, so their linguistic biases and preferences may affect the annotation of the dataset.

Second, despite our best efforts, it is not guaranteed that the accuracy of the corrected text will be perfect. It is possible that some errors may be overlooked by the annotators or that unnecessary corrections may be made.

Finally, a part of the dataset consists of translations from other languages. This could induce specific types of errors which are not generalizable across different types of text.

Acknowledgments

This work is supported by Grammarly. We thank Ira Kotkalova, Anna Vesnii, Halyna Kolodkevych, and everyone else who participated in the corpus creation.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

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 and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China. Association for Computational Linguistics.

Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. 2020. Developing a new classifier for automated identification of incivility in social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 95–101, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Satoru Katsumata and Mamoru Komachi. 2020. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 827–832, Suzhou, China. 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.

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.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. 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.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36, Los Angeles, California. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. http://ruder.io/nlp-beyond-english.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online. Association for Computational Linguistics.