# GERA: a corpus of Russian school texts annotated for Grammatical Error Correction

Alexey Sorokin[1,2] Regina Nasyrova[1]

[1]MSU AI Institute [2]Yandex

October 17, 2024

## Grammatical error correction

- Grammatical error correction is the task of converting a text into its version without errors.
- Errors may be:
  - Grammatical in a strict sense (wrong choice of word form).
  - Orthographic and typos.
  - Punctuation.
  - Lexical.
  - Discourse.

## M2 Annotation

- M2 is the traditional format for error annotation:

```
S The other advantage of the swim is the less impact the knee .
A 4 5|||U:DET|||||REQUIRED|||-NONE-|||0
A 5 6|||R:NOUN:NUM|||swimming|||REQUIRED|||-NONE-|||0
A 8 8|||M:OTHER|||fact that there is|||REQUIRED|||-NONE-|||0
A 10 10|||M:PREP|||on|||REQUIRED|||-NONE-|||0
A 11 12|||R:NOUN:NUM|||knees|||REQUIRED|||-NONE-|||0
```

## Corpora for Russian

- RULEC (Rozovskaya et al., 2019):
  - 12480 sentences (206258 tokens): 4980 train, 2500 dev, 5000 test.
  - Essays written by 12 second language learners and 5 heritage speakers.
  - Suboptimal quality of annotation, imperfect coverage of errors.
- RU-Lang8 (Trinh, Rozovskaya, 2021):
  - 4412 sentences (54741 tokens): 1968 dev, 2444 test.
  - Sentences collected from Lang8 online learning platform and manually reannotated.
  - No manually annotated train set.

## New corpora: motivation

- There is no native language data for Russian GEC.
- Native and learners written texts differ by several parameters:
  - Native writers use more complex sentences.
  - Foreign learners do not use much punctuation.
  - Native writers make less strictly grammatical errors.
  - Patterns of orthographic and lexical errors also differ.
- This makes existing GEC corpora suboptimal for text quality related applications.

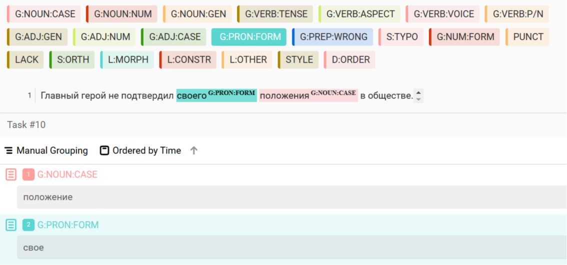
# GERA: data sources

- sources: 456 anonymized essays written in Russian middle school.
- 16 topics: Architecture, Crime and Punishment, Fathers and Sons, Happiness, Oblomov, Petersburg Tales, Song, The Captain's Daughter, The Novice (Mtsyri), The Russian "Miracle" man, The Storm, The ideal ruler, The lyrical hero, War and Peace, Other literary works, Dictations.
- Annotation units are individual sentences.
- Several sentences are grouped in case of errors that require previous sentences for disambiguation:

```
S Песни нас сопровождают всю жизнь и все эпохи . Она способна
влиять на внутренний мир и формировать человека ,
так как связана с культурой страны .
A 9 10|||G:PRON:FORM|||Они|||REQUIRED|||-NONE-|||0
A 10 11|||G:ADJ:NUM|||способны|||REQUIRED|||-NONE-|||0
A 21 22|||G:ADJ:NUM|||связаны|||REQUIRED|||-NONE-|||0
```

# GERA: Annotation process

- Texts were annotated in LabelStudio:



1 Главный герой не подтвердил своего G:PRON:FORM положения G:NOUN:CASE в обществе.

Task #10

1 G:NOUN:CASE

положение

2 G:PRON:FORM

свое

- Annotation was performed by 6 linguistics students.
- It was further verified by the principal annotator who could change the annotation.

| **Annotator** | **A** | **B** | **C** | **D** | **E** | **F** |
|---|---|---|---|---|---|---|
| **Unchanged Sentences (%)** | 80.5 | 67.0 | 73.5 | 78.5 | 73.5 | 60.0 |

## Corpus size

- Subcorpus parameters:

| Sample | Sentences | % of Incorrect Sentences | Tokens |
|---|---|---|---|
| Train | 4592 | 50.30 | 81088 |
| Validation | 775 | 50.71 | 15478 |
| Test | 1314 | 48.48 | 22502 |

- Other corpora:

| Data | Sentences | Tokens | Err. tokens | Errors |
|---|---|---|---|---|
| RULEC-GEC | 12480 | 206258 | 13048 | 11848 |
| RU-Lang8 (Dev&Test) | 4412 | 54741 | 7163 | 6788 |
| GERA | 6681 | 119068 | 5053 | 5988 |

# Error categories

- We annotate error types, not only their presence:

```
S Видя эту композицию исчезает лень ,
появляется желание трудится во благо Отечества .
A 0 1|||L:MULTIMORPH|||Когда видишь|||REQUIRED|||-NONE-|||0
A 3 3|||PUNCT|||,|||REQUIRED|||-NONE-|||0
A 8 9|||G:VERB:FORM|||трудиться|||REQUIRED|||-NONE-|||0
```

- There are 7 main error types:
  - G – grammatical errors,
  - S – spelling errors,
  - L – lexical errors,
  - D – discourse errors
  - STYLE, LACK (missing word), PUNCT.

# Error categories

- Errors are further divided into subtypes.
- For grammatical errors we indicate word part-of-speech and incorrect grammatical features:

| | | |
|---|---|---|
| G:NOUN:NUM | Incorrect noun number | Н. В. Гоголь использует гротеск в этих __*произведении__<br>*‘N. V. Gogol uses grotesque in these __*work__.’* |
| G:NOUN:CASE | Incorrect noun case | Графиня была в очень глубоком __*отчаяние__<br>*‘The countess was in very deep __despair__.’*<br>*Accusative Case/Locative Case |

# Error categories

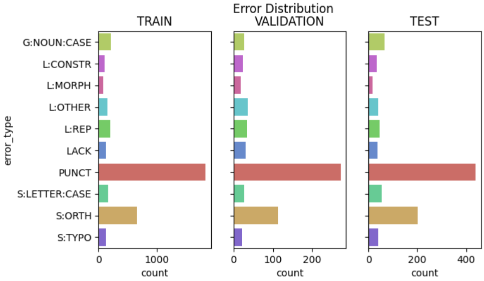
- For grammatical errors we indicate word part-of-speech and incorrect grammatical features:
  - Nouns: NUM, CASE;
  - Verbs: TENSE, ASPECT, P/N/G (person/number/gender), FORM;
  - Adjectives: GEN, NUM, CASE, Deg (degree), Sh/L (short/long);
  - Pronouns: FORM, WRONG (его/свой);
  - Numerals: FORM;
  - Prepositions, conjunctions: WRONG.

## Error categories

- Spelling errors: S:ORTH, S:LETTER:CASE, S:TYPO.
- Lexical: L:MORPH (same root), L:MULTIMORPH (word combination with the same root), L:CONSTR (word combination errors), L:OTHER, L:REP (repetition).
- Discourse: D:ORDER, D:REF (referential error or ambiguity).

# Error frequencies

| RULEC-GEC (learners) | | RULEC-GEC (heritage) | | RU-Lang8 | | GERA | |
|---|---|---|---|---|---|---|---|
| Spell | (18.6) | Spell | (42.4) | Spell | (19.2) | Punct | (42.5) |
| Noun:Case | (14.0) | Punct | (22.9) | Noun:Case | (12.6) | Spell | (23.6) |
| Lex. Choice | (13.3) | Noun:Case | (7.8) | Lex. Choice | (11.6) | Lex. Choice | (13.6) |
| Lack | (8.9) | Lex. Choice | (5.5) | Punct | (10.3) | Noun:Case | (5.1) |



# Model results

| model | P | R | $F_{0.5}$ |
|---|---|---|---|
| GPT2-large finetuned[1] | 73.4 | 23.4 | 51.4 |
| GPT2-large finetuned generator + ranker[1] | 78.4 | 44.4 | 68.0 |
| rule-based generator + ranker[1] | **86.1** | 42.9 | 71.6 |
| Yandex GPT zeroshot | 65.3 | 56.9 | 63.4 |
| Yandex GPT finetuned | 77.8 | **58.3** | **73.0** |

[1] – pipeline from Sorokin (2022)

- YandexGPT finetuning provides the best results.
- The highest precision is achieved by the combination of rule-based generation and reranking.

# Results by categories

- Results of the two best models by subsets:

| Label | YandexGPT | | | generator-ranker | | |
|---|---|---|---|---|---|---|
| G:ADJ:CASE | 80.0 | 42.1 | **67.8** | 100.0 | 26.3 | 64.1 |
| G:NOUN:CASE | 70.9 | 58.2 | 67.9 | 83.3 | 52.2 | **74.5** |
| G:NOUN:NUM | 62.5 | 50.0 | **59.5** | 100.0 | 25.0 | 55.6 |
| G:PRON:FORM | 57.1 | 23.5 | **44.4** | 100.0 | 11.8 | 40.0 |
| G:VERB:P/N/G | 84.6 | 68.8 | **80.9** | 77.8 | 43.8 | 67.3 |
| L:MORPH | 100.0 | 27.8 | **65.8** | 100.0 | 5.6 | 22.8 |
| L:OTHER | 2.9 | 5.1 | 3.2 | 6.7 | 2.6 | **5.1** |
| PUNCT | 75.8 | 67.7 | **74.0** | 80.8 | 54.7 | 73.8 |
| S:LETTER:CASE | 84.2 | 59.3 | **77.7** | 71.4 | 18.5 | 45.5 |
| S:ORTH | 87.6 | 73.3 | **84.3** | 78.8 | 51.5 | 71.2 |

# Results by categories

- LLMs clearly outperform earlier approaches for almost all categories.
- The only exceptions are noun case errors and punctuation.
- Even LLMs struggle with lexical errors.

# Results for other corpora

- Pretraining on GERA improves quality for other corpora (for generator-ranker pipeline):

| Training data | RULEC-GEC | | | RU-Lang8 | | |
|---|---|---|---|---|---|---|
| single corpus | **68.1** | 24.2 | 49.9 | 66.0 | **30.2** | 53.3 |
| RULEC-GEC+RU-Lang8 | 64.4 | **29.8** | 52.2 | 67.6 | 30.0 | 54.4 |
| all corpora | 66.5 | 28.6 | **52.6** | **70.5** | 29.1 | **54.8** |

# Conclusions

- We released a corpus of native russian texts annotated for grammatical errors.
- Finetuning Yandex GPT yields the best results on this corpus.
- Lexical errors are the hardest with almost zero correction quality.
- Probably, multi-reference corpora should be collected to better deal with such mistakes.

Спасибо за внимание!

Thank you!

Рахмат!