Russian Learner Corpus:

Towards Error-Cause Annotation for L2 Russian

Daniil Kosakin[1], Sergei Obiedkov[2], Ekaterina Rakhilina[1],
Ivan Smirnov[1], Anastasia Vyrenkova[1], Ekaterina Zalivina[1]

[1]HSE University [2]TU Dresden

Outline

- RLC: Russian Learner Corpus
- RLC-GEC: Annotated subset of RLC
- RLC-Crowd: Crowdsourced corrections
- RLC-ERRANT: Automatic error annotation

Russian Learner Datasets

RULEC-GEC (Rozovskaya and Roth, 2019)	12,480 sentences Essays written by English-speaking learners of Russian	Automatic error classification tool (Rozovskaya, 2022)
RU-Lang8 (Trinh and Rozovskaya, 2021)	48,260 sentences (4,412 re-annotated) Data from Lang8 language-learning website	
ReLCo (Katinskaia *et al.*, 2022)	22,370 sentences Data from exercises performed using the Revita language-learning system	RuERRANT

The annotation systems are mainly based on grammatical features of individual words.

Russian Learner Corpus (RLC)
(Rakhilina *et al.* 2016)

- Essays written by heritage speakers and L2 learners of Russian
- 48 dominant languages
- Over 190,000 sentences (2,200,000 tokens)
- Half of RLC is manually corrected and annotated
- Available through a search interface at http://web-corpora.net/RLC/

Error Annotation in RLC

- Highlight what caused the error rather than simply indicate where it occurred
- 36 error tags: grammatical vs lexical vs derivational vs spelling errors
- Morphological markup is present in a separate layer
- One error tag may cover several tokens
- Several error tags may be attached to one token

Error Annotation in RLC

Example: Error Boundaries

Ремонт делает этим (instr) *великолепным* (instr) *зданием* (instr) *идеальным для жилья.*
Ремонт делает это (acc) *великолепное* (acc) *здание* (acc) *идеальным для жилья.*
Renovation makes this gorgeous building perfect for living.

- The noun is in the wrong case.
- The determiner and adjective are also in the wrong case, but they agree with the noun.
- This is a single error in government (not three errors).

Error Annotation in RLC

Example: Noun Endings

- (*быть*) *друг* → *другом* (instr) — `Gov`
  (be a) friend
  “Friend” is in the nominative case: error in government
- (*является*) *поэмом* → *поэмой* (instr) — `Gender`
  (is a) poem
  “Poem” is treated as a masculine word, although it’s feminine
- (*хлеб с*) *моцарелли* → *моцареллой* (instr) — `Infl` / `Gov + Infl`
  (bread with) mozzarella
  Using an existing inflection results in a non-existing word

Error-Cause Annotation Challenges

- Errors can be attributed to different patterns of second language acquisition.
- Errors may be caused by patterns transferred from the dominant language.

- This leads to low inter-annotator agreement
- and makes annotation hard to automate.

RLC-GEC

An annotated subset of RLC

- 2,004 texts
- 31,519 sentences
- 41,410 error annotations
- Meta-information: dominant language, L2/heritage, language proficiency level

- RLC-Test
  - 204 sentences
  - 519 error annotations

Dominant language	Texts
English	760
Chinese	304
French	214
Kazakh	157
Spanish	123
Turkmen	98
Italian	72
+21 other languages	276

Error Tag	%
`Lex`	19.7
`Ortho`	15.8
`Syntax`	13.8
`Gov`	8.3
`Constr`	6.9
`Miss`	5.7
`Prep`	5.3
…	…

RLC-Crowd

34,150 sentences from RLC, most have no annotations in RLC

Toloka platform (https://toloka.ai) was used to obtain at least five corrections for each sentence

213,683 corrected sentences

- The quality of corrections varies greatly.
- Aggregation methods are needed to obtain reliable corrections.
- Five corrections per sentence may not be enough.

- May be good as is for training or fine-tuning machine-learning GEC models.
- A valuable resource for studying users' correction strategies, the visibility of errors across various types, etc.

RLC-ERRANT

Error-annotation tool following the rule-based approach of ERRANT (Bryant *et al.* 2017).

Input: A sentence and its correction

Output: A list of edits classified into RLC types

Можно увлечься чем-то более полезней и при том отдохнуть.

Можно увлечься чем-то более полезным и притом отдохнуть.

```
Orig: [4, 5, 'полезней'], Cor: [4, 5, 'полезным'], Type: 'Com'

Orig: [6, 8, 'при том'], Cor: [6, 7, 'притом'], Type: 'Space+Ins'
```

RLC-ERRANT

Error Extraction

- Alignment based on Damerau-Levenshtein distance,
- followed by rule-based merging of some adjacent edits

Example

If adjacent words in the original sentence share the number and case different from those in the corrected sentence, this is a single error.

Ремонт делает этим (sg instr.) *великолепным* (sg instr) *зданием* (sg instr) *идеальным для жилья.*
Ремонт делает это (sg acc) *великолепное* (sg acc) *здание* (sg acc) *идеальным для жилья.*

```
Orig: [2, 5, 'этим великолепным зданием’],
 Cor: [2, 5, 'это великолепное здание'], Type: 'Gov'
```

RLC-ERRANT

Error Classification

- A simplified version of the RLC tagset is used.
- Each edit is assigned a single tag.
- Tag assignment is rule-based.
- Rules are applied sequentially.
- The first rule that fires defines the tag.

```
WO, CS, Brev, Tense, Passive, Num, Gender, Nominative/Gov/AgrCase,
AgrNum, AgrPers, AgrGender, Refl, Asp, Impers, Com, Mode, Hyphen+Ins,
Hyphen+Del, Space+Ins, Space+Del, Conj, Ref, Prep, Graph, Infl, Lex,
Constr, Ortho, Morph, Ortho, Misspell
```

A Classification Rule

Nominative/Gov/AgrCase

этим (Det sg instr) *великолепным* (Adj sg instr) *зданием* (Adj sg instr)
→ *это* (Det sg acc) *великолепное* (Adj sg acc) *здание* (Adj sg acc)

- The sequences contain the same number of tokens.
- All tokens within each sequence agree in number and case.
- The cases are different for the two sequences, but the numbers are the same.
- The corresponding tokens have the same lemmas.

`AgrCase` if none of the tokens is a noun or a pronoun.

`Nominative` if the correct case is nominative.

`Gov` otherwise.

RLC-ERRANT

Experimental Evaluation

We tested RLC-ERRANT on RLC-Test.

- Overall accuracy: 0.58
- Many classification errors are due to incorrectly determined morphological categories, especially for non-existing words.
- Orthographic errors are often hard to differentiate from morphological errors.
- Training a machine-learning classifier can help here.

Tag	Precision	Recall
`Lex`	0.70	0.77
`Ortho`	0.73	0.10
`Gov`	0.91	0.75
`Constr`	0.62	0.38
`Prep`	0.97	0.78
`Ref`	0.76	0.81
`Asp`	0.71	0.71
`Conj`	0.77	0.87

Conclusion

We released https://github.com/Russian-Learner-Corpus

- Two L2 Russian datasets with over 30,000 sentences each
  - RLC-GEC is linguistically annotated
  - RLC-Crowd contains 200K+ crowdsourced corrections, at least five per sentence
- RLC-ERRANT, an error annotation tool for RLC error tagging system

Plans

- Make other parts of RLC publicly available
- Analyze the crowdsourced data for users' correction strategies
- Use the data to train or fine-tune machine-learning models
- Improve the performance of RLC-ERRANT using machine learning