Russian Learner Corpus: Towards Error-Cause Annotation for L2 Russian

D. Kosakin, S. Obiedkov, E. Rakhilina, I. Smirnov, A. Vyrenkova, E. Zalivina

Faculty of Computer Science, HSE University, Moscow, Russia
Faculty of Computer Science / cfaed / ScaDS.AI, TU Dresden, Germany
School of Linguistics, HSE University, Moscow, Russia

Research Goal and Deliverables

Annotation schemes of datasets currently available for Russian as FL are mainly based on grammatical features of individual words

We propose an approach with a stronger focus on causes of errors

- Two open datasets related to GEC tasks that are derived from a large learner corpus of Russian

- Tool for automatic error extraction and classification expected to streamline the process of error annotation for the permanently growing learner dataset

Related Work: Learner Corpora and Learner Datasets

English (Granger, 1998; Dahlmeier, 2013 among others)
German (Boyd et al., 2014)
Japanese (Mizumoto et al., 2011)
Czech (Rosen, 2016) etc.

RULEC-GEC: (Rozovskaya and Roth, 2019)	12,480 sentences	Automatic Error Classification Tool (Rozovskaya, 2022)
RU-Lang8: (Trinh and Rozovskaya, 2021)	48,260 sentences (4,412 annotated)	
ReLCo (Katinskaia et al., 2022)	22,370 sentences	RuERRANT

Russian Learner Corpus (RLC):

- more than 193,180 sentences (2,200,000 tokens)
- essays written by heritage bilinguals
  and students of Russian as a foreign language
- 48 dominant languages
- 36 error tags

RLC: Error annotation challenges

- multiple errors that can be attributed to various patterns of second language acquisition and use
- large dataset
- markup for errors is done manually
- low inter-annotator agreement

What is wrong: grammatical vs lexical vs spelling; grammatical properties

Error Annotation in RLC

- Error markup focuses on the problems that RFL students face in the process of speech production
- Morphological markup is presented in a specifically designed layer and does not form a basis for error classification
- One error tag may cover several tokens
- Several error tags may be put on one token

Error Annotation in RLC

появление специалистов было	**золотое** (A, Nom) **дно** (N, Nom)	для корпораций
poyavlenie specialistov bylo	zolotoe dno	dlya korporaciy
the emergence of specialists was	a gold mine (lit. golden bottom)	for corporations

Error Annotation in RLC

появление специалистов было	**золотым дном** (N,Instr)	для корпораций
poyavlenie specialistov bylo	**zolotym dnom**	dlya korporaciy
the emergence of specialists was	a gold mine (lit. golden bottom)	for corporations

Error Annotation in RLC

появление специалистов было	**золотым** (A, Instr) **дном** (N,Instr)	для корпораций
poyavlenie specialistov bylo	**zolotym** **dnom**	dlya korporaciy
the emergence of specialists was	a gold mine (lit. golden bottom)	for corporations

Error Annotation in RLC

появление специалистов было	**золотое** (A, Nom) **дно** (N, Nom)	для корпораций
poyavlenie specialistov bylo	zolotoe dno	dlya korporaciy
the emergence of specialists was	a gold mine (lit. golden bottom)	for corporations

Error Annotation in RLC

появление специалистов было **золотое** **дно** для корпораций

poyavlenie specialistov bylo **zolotoe** **dno** dlya korporaciy

the emergence of specialists was a gold mine for corporations

A 4 5|||**Прил.:Падеж (Adj:Case)|||золотым**|||REQUIRED|||-NONE-|||0

A 5 6|||**Сущ.:Падеж (Noun:Case)|||дном**|||REQUIRED|||-NONE-|||0

RULEC-GEC (Rozovskaya, 2019)

Error Annotation in RLC

появление специалистов было **золотое** **дно** для корпораций

poyavlenie specialistov bylo **zolotoe** **dno** dlya korporaciy

the emergence of specialists was a gold mine for corporations

Orig: [4, 5, 'золотое'], Cor: [4, 5, 'золотым'], Type: 'R:ADJ:CASE'

Orig: [5, 6, 'дно'], Cor: [5, 6, 'дном'], Type: 'R:NOUN:CASE'

RuERRANT, RelCo (Katinskaya, 2021)

Error Annotation in RLC

появление специалистов было **золотое дно** для корпораций

poyavlenie specialistov bylo **zolotoe** **dno** dlya korporaciy

the emergence of specialists was a gold mine for corporations

золотое дно → золотым дном; Error tag: 'Gov'

- Morphological Markup:

золотое	zolotoe	(A:nom;plen;sg)	дно	dno	(N:acc/nom;sg)
золотым	zolotym	(A:instr;plen;sg)	дном	dnom	(N:instr;sg)

Error Annotation in RLC: out-of-vocabulary words

поэм-ом	→	**поэм-ой**
poem-om (invalid N, Sg, Instr)	→	poem-oj (N, Sg, Instr)

Error Annotation in RLC: out-of-vocabulary words

поэм-ом	→	**поэм-ой**
poem-om (invalid N, Sg, Instr)	→	poem-oj (N, Sg, Instr)

-ом (-om): characteristic of Instr case of masculine nouns

-ой (-oj): characteristic of Instr case of feminine nouns

Error Annotation in RLC: out-of-vocabulary words

поэм-ом → **поэм-ой**

poem-om (invalid N, Sg, Instr) → poem-oj (N, Sg, Instr)

поэма (N, fem., Nom)
poem

-ом (-om): characteristic of Instr case of masculine nouns

-ой (-oj): characteristic of Instr case of feminine nouns

Error Annotation in RLC: out-of-vocabulary words

поэм-ом → **поэм-ой**

poem-om (invalid N, Sg, Instr) → poem-oj (N, Sg, Instr)

поэма (N, fem., Nom)
poem

-ом (-om): characteristic of Instr case of masculine nouns

-ой (-oj): characteristic of Instr case of feminine nouns

Error tag: ‘Gender’

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