REVITA LEARNER CORPUS

Revita Learner Corpus (ReLCo) is an automatically and manually annotated learner corpus for Russian. It is collected **continually** while students perform exercises in Revita—an open platform for language learning and tutoring beyond the beginner level—revita.cs.helsinki.fi.

The corpus grows as students perform more exercises.

- Revita automatically generates exercises.
- Base form—lemma—is given to learner as a hint.
- Answers are checked automatically.
- Revita expects only one answer as correct—the one matching the original text.

We collected learner data to study forms of a word may be grammatically and semantically correct in a given context.

**Alternative answers**—also correct: “I [walk] down the street and [see] a poodle.”

**Expected answer** may be: “I walked down the street and saw a poodle.”

**Alternative answers**—also correct: “I walked down the street and saw a poodle.”

We release the annotated data collected from 531 learners, in two sets:

- Paraphrases with only **grammatical errors**.
- Paraphrases which include at least one orthographic error.

Find the data here: github.com/Askinkaty/Russian_learner_corpora

**MANUAL ANNOTATION**

The first subset was checked by a language expert and manually annotated. Paragraphs with at least one AC answer (1 427 paragraphs) were manually double-checked by 6 native speakers.

We adapted the grammatical error annotation toolkit ERRANT to Russian (https://github.com/Askinkaty/errant) to annotate the data with grammatical error types.

**AUTOMATIC ANNOTATION**

We created a semi-automatically annotated learner corpus for Russian.