SEMI-AUTOMATICALLY ANNOTATED LEARNER CORPUS FOR RUSSIAN

Anisia Katinskaia
Maria Lebedeva
Roman Yangarber

first.last@helsinki.fi,
m.u.lebedeva@gmail.com

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
FACULTY OF ARTS

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.

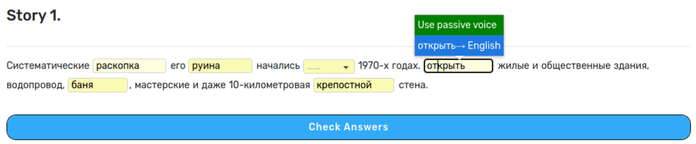

Figure 1: Practice mode in the Revita language-learning system.

Revita expects only one answer (= the original form) and can tag potentially correct answers as errors.

Example text with [cloze exercises]: "*I [walk] down the street and [see] a poodle.*"

Expected answer may be: "*I **walk** down the street and **see** a poodle.*"

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

We collected learner data to study ***alternative-correct (AC) answers***—alternative forms of a word may be grammatically and semantically correct in a given context.

AC category	%	# Examples
Number: pl/sg	25.9	436
Tense: present/past	11.4	191
Number: sg/pl	9.8	165
Aspect: imperf/perf	8.7	146
Tense: past/present	8.3	140
Aspect: perf/imperf	5.6	94
Adj: short/full	3.8	64
Verb form: transgressive/past	2.6	44
Other: word form	2.4	41
Preposition	2.1	36
Tense: future/past	1.6	27
Case: accusative/locative	1.3	22
Tense: past/future	1.2	20

Table 1: Types of the most frequent alternative-correct (AC) answers in the annotated dataset.

Subset	Paragraphs	Sent.	Tokens	Errors per paragraph	Errors per sent.
Gramm. errors	6 141	15 568	263 101	1.6	0.64
Non-word errors	2 700	6 802	112 352	2.0	0.75

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

- Paragraphs with only *grammatical errors* errors.
- Paragraphs 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.

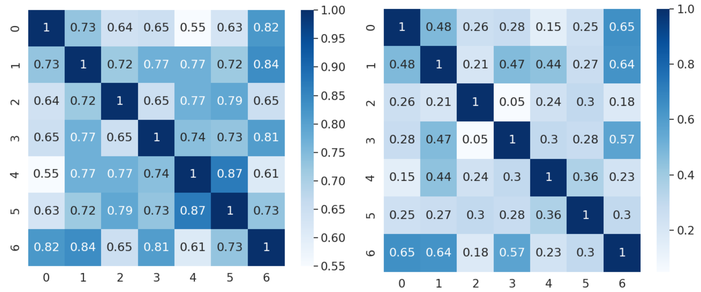

Figure 2: Agreement between 6 annotators, shown as percentage (left) and Cohen's kappa (right)

All conflicting annotations were resolved by a language expect. 34% of all answers include at least one annotator who disagrees with other annotators. These answers are tagged as "hard", see Table 2.

Category	%	# Examples
Number: pl/sg	15.5	90
Aspect: perf/imperf	12.3	71
Aspect: imperf/perf	10.0	58
Tense: present/past	10.0	58
Number: sg/pl	8.4	49
Tense: past/pres	7.2	42
Adj: short/full	4.6	27
Verb form: transgressive/past	3.7	22
Tense: future/past	2.9	17
Tense: past/future	2.4	14
Case: accusative/locative	1.7	10

Table 2: Grammatical types of answers for which the annotators disagree. Percentage shows the fraction of the type among all answers tagged as "hard".

AUTOMATIC ANNOTATION

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

Error type	%	#
R:SPELL	25.9	3251
R:NOUN:CASE	11.2	1403
R:ADP	5.0	630
R:NOUN:NUM:CASE	4.7	585
R:PRON	4.4	553
R:VERB:NUM	3.9	491
R:VERB:FORM	3.8	474
R:NOUN:NUM	3.6	452
R:OTHER	4.2	451
R:VERB:TENSE	3.4	430

Error type	%	#
R:DET	3.1	395
R:ADJ:CASE	2.0	254
R:VERB:GENDER	1.9	240
R:MORPH	1.8	228
R:VERB:ASPECT	1.7	215
R:VERB:INFL	1.7	210
R:ADJ:NUM	1.6	205
R:ADJ:NUM:CASE	1.6	199
R:ADJ:FULL/SHORT	1.1	140
R:ADJ:GENDER	1.0	128
R:AUX	1.0	125

Table 3: Statistics on types of grammatical errors assigned automatically by RuERRANT.

ANNOTATORS VS. A NEURAL MODEL

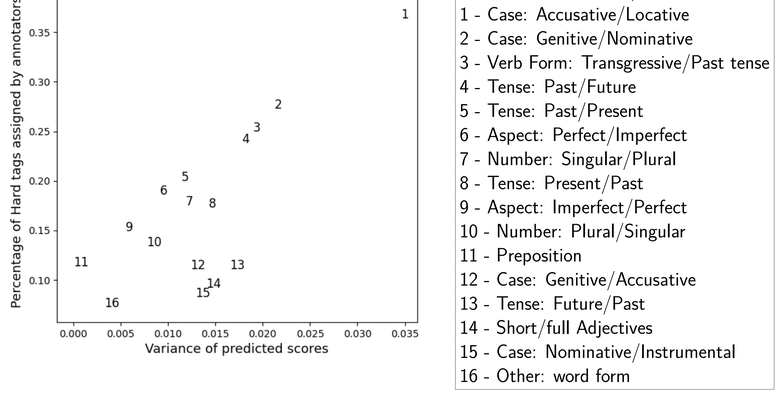

Figure 3: Uncertainty of human annotators vs. uncertainty of a neural model, both assessing grammatical correctness of learners answers. On the right: 17 types of errors annotated.

CONCLUSIONS AND FUTURE WORK

- Released leaner corpus with automatically and manually annotated answers
- Released RuERRANT for automatic annotation of error types
- Plan to improve RuERRANT with better models for morphological analysis
- Improve automatic assessment of grammatical correctness of answers
- Process full essays written by learners