Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 832–839
Marseille, 20-25 June 2022

Semi-automatically Annotated Learner Corpus for Russian

Anisia Katinskaia[1,2], Maria Lebedeva[3], Jue Hou[1,2], Roman Yangarber[2]
[1] Department of Computer Science, University of Helsinki, Finland
[2] Department of Digital Humanities, University of Helsinki, Finland
[3] Language and Cognition Laboratory, Pushkin State Russian Language Institute, Russia
first.last@helsinki.fi, m.u.lebedeva@gmail.com

Abstract

We present ReLCo— the Revita Learner Corpus—a new semi-automatically annotated learner corpus for Russian. The corpus was collected while several hundreds L2 learners were performing exercises using the Revita language-learning system. All errors were detected automatically by the system and annotated by type. Part of the corpus was annotated manually—this part was created for further experiments on automatic assessment of grammatical correctness. The Learner Corpus provides valuable data for studying patterns of grammatical errors, experimenting with grammatical error detection and grammatical error correction, and developing new exercises for language learners. Automating the collection and annotation makes the process of building the learner corpus much cheaper and faster, in contrast to the traditional approach of building learner corpora. We make the data publicly available.

Keywords: Learner corpus, L2, ICALL, language learning, grammatical error detection

1. Introduction

This paper presents a new version of Revita Learner Corpus (ReLCo), automatically and manually annotated learner corpus for Russian[1] (Katinskaia et al., 2020). It is created to facilitate research in second language (L2) acquisition and foreign language teaching (Granger, 1996), as well as in various NLP tasks, e.g., grammatical error correction (GEC) and grammatical error detection (GED).

ReLCo is built based on the Revita language-learning platform (Katinskaia et al., 2018; Katinskaia and Yangarber, 2018).[2] It is collected continually and annotated *automatically*, while students perform a variety of exercises in the platform. All collected learner data is used by Revita for updating student models of language proficiency, which in turn are used for generating new personalised exercises. This process creates the "learning feedback loop" helping students to improve their language skills more effectively. Manual annotation was added to the part of released data for further experiments with improving automatic annotation, i.e., detection of grammatical errors in learner answers and tagging of errors by their type.

Most corpora available today contain English as the target learning language (L2). Of the 174 learner corpora in the list prepared by the Centre for English Corpus Linguistics at the Université Catholique de Louvain,[3] 93 are English learner corpora—over 53%. Only a few of the corpora in the list are available for download. This paper partly address both of these problems by focusing on Russian and making our data freely available.

ReLCo is not a "conventional" learner corpus—it does not include sentences that were fully written by L2 students. However, we claim that this is a valuable resource of learner data, which can be used for improving teaching and learning processes and which is *continually growing* as learners perform more exercises. One example of using the corpus is creating distractors for multiple-choice exercises based on common errors.

The data consists of sentences where some of the words in the sentence were automatically selected to be hidden from the learner—to be part of an exercise—and the learners' task was to answer these exercises (see more details in the Section 3), by filling in the correct words or expressions. Revita decides which words to pick for exercises based on the learner's history of previous answers.

In addition to the data, we release a Russian version of ERRANT,[4] a grammatical error annotation toolkit, which automatically extracts edits from parallel original and corrected sentences and classifies them by type, (Felice et al., 2016; Bryant et al., 2017). ERRANT is widely used for evaluating GEC model performance by error type. In addition, ERRANT allows one to transfer existing error correction corpora into a standardized format and to reduce the amount of work for human annotators.

This paper is structured as follows. In Section 2, we review previous work on creating learner corpora for Russian. Section 3 presents the main features of Revita Learner Corpus and how data is collected. In Section 4 we describe the manual and automatic annotation and analysis of collected errors. In Section 5 we present results of testing one of our baseline models for assessing grammatical correctness of learner answers using the

[1] https://github.com/Askinkaty/Russian_learner_corpora
[2] revita.cs.helsinki.fi
[3] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
[4] https://github.com/Askinkaty/errant

new dataset (Katinskaia and Yangarber, 2021). Conclusions and future work are discussed in the final section.

2. Related Work

One definition of a *learner corpus* was provided by (Granger, 2002): “computer learner corpora are electronic collections of authentic textual data collected according to explicit design criteria for a particular SLA/FLT[5] purpose in a standardized format.” SLA and FLT researchers have been collecting learner output since the disciplines emerged.

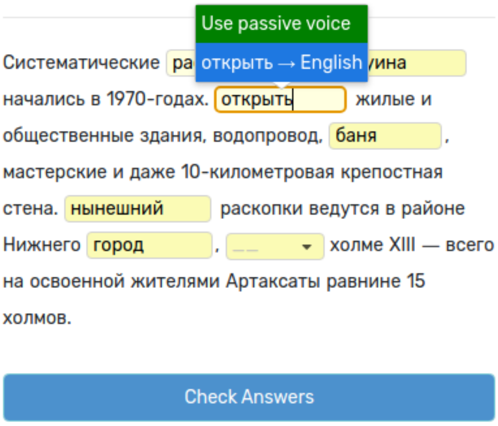
Typology of learner corpora takes into account several dimensions: plain text vs. annotated learner data; written vs. transcribed spoken data; general vs. “Language for Specific Purposes” (LSP) learner corpora; target language; mother tongue; proficiency in target language; synchronic vs. diachronic; global vs. local (collected by a teacher among their students); commercial vs. academic (Granger, 2002; Gilquin and Granger, 2015). Regardless of the type of data, *any* learner corpus is very difficult and expensive to collect.

In the context of this paper, we are interested in written learner corpora of Russian. Several projects focus on Russian as the target learner language:

1. Russian Learner Corpus (RLC) (Rakhilina et al., 2016);
2. Russian Learner Corpus of Academic Writing (RULEC) (Yatsenko et al., 2012) and its error-annotated subset RULEC-GEC (Rozovskaya and Roth, 2019);
3. The Corpus of Russian Student Text (CoRST), which includes academic writing (Zevakhina and Dzhakupova, 2015) produced by *native* speakers of Russian;
4. Narrative collections (Protassova, 2016; Polinsky et al., 2008);
5. Russian Learner Translator Corpus (Kutuzov and Kunilovskaya, 2014), which is a bi-directional parallel corpus of English-Russian translations done by university translation students.

Only the first two are L2 learner corpora with annotated errors. Russian Learner Corpus (RLC) is the first corpus, which makes a clear distinction between “heritage” and L2 speakers. It is a collection of oral and written texts with morphological and error annotation. Morphological annotation was performed by the MyStem morphological analyzer (Segalovich and Titov, 1997), while linguistics students annotated the errors. Along with annotation, RLC provides metadata about the author of each text—sex, L2 or heritage, dominant language, etc.—and about the text itself—written or oral, genre, and a time limit.

RULEC is a subset of RLC—a longitudinal corpus of Academic Writing, collected from students learning Russian as a foreign language and heritage speakers in the US. RULEC-GEC is a subset of RULEC, which was manually annotated: errors were corrected and assigned a type. The error types include syntactic, morpho-syntactic errors, orthographic, and lexical errors.



Систематические ра[открыть → English]уина
начались в 1970-годах. открыть жилые и
общественные здания, водопровод, баня ,
мастерские и даже 10-километровая крепостная
стена. нынешний раскопки ведутся в районе
Нижнего город , __ холме XIII — всего
на освоенной жителями Арктаксаты равнине 15
холмов.

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

3. Revita Learner Corpus for Russian

Revita learner corpus (ReLCo) is collected while students practice with exercises in the language-learning system. Revita is an online L2 learning system for students beyond the beginner level, which allows practicing with arbitrary authentic texts. It covers several languages, most of which are highly inflectional, e.g., Finnish. The same principle works for all languages—the learner can choose an available text from the library or upload any text into the system, which generates exercises automatically. A teacher can also choose texts to be sent to learners who belong to this particular teacher’s groups. The system tries to adapt the level of exercises to each learner individually, depending on her level of proficiency. Continuous assessment of the learners’ answers is performed automatically (Hou et al., 2019).

All exercises in Revita are linked to linguistic “constructs,” which constitute the knowledge representation of the language. Each construct is a “skill” that the learner must learn. For example, one construct may be the usage of various forms of nouns belonging to certain paradigms; preposition or verb government; usage of perfective vs. imperfective aspect, etc. The learner’s performance on all practiced concepts determines which exercises Revita will offer to the learner next. Therefore, the corpus includes not random words in the context replaced by the answers given by the learner, but the ones that should help to learn the skill which the learner is ready to learn. This relates to the *Zone of Proximal Development*—one of the pedagogical frameworks which is used in Revita, and based on the Vygotskian approach (Vygotsky, 1978; Vygotsky, 2012).

The system provides a variety of modes for interaction,

[5] Second Language Learning, Foreign Language Teaching

Subset	Paragraphs	Sentences	Tokens	Errors per paragraph	Errors per sentence
Grammatical errors	6 141	15 568	263 101	1.6	0.64
Non-word errors	2 700	6 802	112 352	2.0	0.75

Table 1: Subsets of the released learner data.

e.g., text-based practice mode, flashcards, crosswords, tests, etc. The learner corpus presented in this paper is built based on the data collected during the practice mode (see Figure 1), which include the following types of exercises:

- "cloze" exercises (fill-in-the-blank) with the lemma of a missing word given as a hint;
- multiple-choice exercises—with distractors generated for many kinds of grammatical constructions;
- listening exercises, where the learner must enter the missing word(s) based on what she hears and the story context.

Sometimes more than one word is removed from the text and replaced by a single lemma as a hint. For example, an infinitive form ("*спасать*", "to save") will be shown as a hint for analytic future tense ("*буду спасать*", "will save") or past passive ("*был спасен*", "was saved") verb forms.

The learner "reads" the text one paragraph at a time with several words or phrases replaced by exercises described above. Learners are given more than one attempt to answer: if the first try was unsuccessful, the learner receives additional feedback and hints and can try the exercises again. For some types of exercises, the system shows additional hints in advance to help if the given context is not sufficient for providing an answer; see Figure 1, where an additional "pre-hint" *"Use passive voice"* is provided in advance. The set of exercises in each paragraph is picked according to the Student Model, which continuously updates estimates of the learner's proficiency, based on the answers given so far. The set of exercises can also be affected by manually-chosen learning settings.

At present, the Revita expects the learner's answer to be the same as the forms found in the original text, otherwise answers are automatically considered to be errors. These errors fall into 3 groups:

1. *Grammatical errors*: the given answer has the same lemma as the expected answer;
2. *Orthographic errors*: the given answer does not correspond to a valid word;
3. *Different lemma*: the given answer has a lemma which is different from the lemma of the expected answer.

This approach to checking answers has some limitations. The main drawback is that learners might insert answers that are different from the expected ones but still *semantically and grammatically suitable* for the given context. Therefore, the system sometimes returns "incorrect" negative feedback to the learner. This problem was formulated as Multiple Admissibility in (Katinskaia et al., 2019). We use the term *alternative-correct answers* (AC) as in (Katinskaia and Yangarber, 2021).

The following examples show answers which have the same lemmas as the expected ones (a) or different lemmas (b):

(a) "*Мне приснилось, как я* ***сдаю*** *экзамены.*"
("*I dreamt that I* ***am taking*** *exams.*")
vs. "*Мне приснилось, как я* ***сдавал*** *экзамены.*"
("*I dreamt that I* ***was taking*** *exams.*")

(b) "*Мне* ***снилось****, как я сдавал экзамены.*"
vs. "*Мне* ***приснилось****, как я сдавал экзамены.*"
("*I* ***dreamt*** *(imperfective vs. perfective) that I was taking exams.*")

In the first example, both highlighted verb forms are valid in the context and have the same lemma "*сдавать*", ("to take"). In the second example, the highlighted verb forms can also be valid, but they belong to different lemmas "*сниться*" ("dream," imperfective) vs. "*присниться*" ("dream," perfective), which differ by the category of verbal aspect.

The released learner corpus consists of paragraphs where each paragraph includes all answers inserted simultaneously by the same learner during one practice session. It is important to note because answers may not be independent, e.g., the learner can put all verb forms in one paragraph in the same tense, and as a result it can affect which answers can be marked as errors.

For example, the words in the brackets are the hints (*lemmas*) of the generated exercises:
"*Я [идти] по улице и [увидеть] пуделя.*"
("*I [walk] down the street and [see] a poodle.*")

Revita might expect the following answers:
"***Я иду*** *по улице и* ***вижу*** *пуделя.*"
("***I walk*** *down the street and* ***see*** *a poodle.*")

But the learner may provide different answers, which alternatively correct only if inserted jointly and annotated together:
"***Я шёл*** *по улице и* ***увидел*** *пуделя.*"
("*I* ***walked*** *down the street and* ***saw*** *a poodle.*")

We release the data in two subsets (see Table 1). One subset includes grammatical errors, and only those answers with different lemma that are either verbs or prepositions. We set this constraint because many

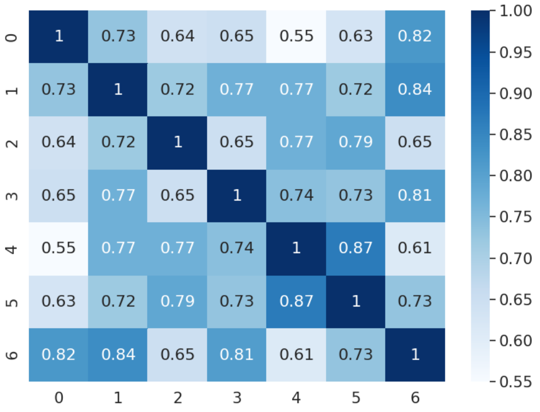


Figure 2: Agreement between 6 annotators, shown as percentage.

learner answers with different lemmas are verb forms that differ by aspect (as presented in the example above), and the grammatical category of verbal aspect is of crucial importance for learning the language. Thus, aspect errors are also considered grammatical errors. Errors with prepositions can occur only in multiple-choice exercises (at present). A wrong choice of preposition can affect the case of the following noun phrase, which is usually governed by the preposition.

This version of the corpus was manually checked and annotated by multiple annotators, in contrast to the versions of ReLCo released previously. The annotation procedure is described in Section 4.

The second subset was build later. It is smaller and partially overlap with the first one. Every paragraph in it includes at least one orthographic error. Other types of errors can also occur in the paragraph if they were inserted at the same time and by the same learner. This data can be used for investigating the patterns of orthographic errors and developing models for spelling correction. In total, both subsets include answers from 531 learners.

Revita automatically identifies orthographic errors. Detection of this type of error is based on the output of a morphological analyzer. The analyzer was built to have very wide coverage, recognizing over 5M valid word forms, therefore if the analyzer returns no analysis for a word, we conclude that it does not exist in the vocabulary, and we consider it as a word with orthographic errors.

4. Corpus Annotation

Revita automatically checks all answers given by the user. Any answer that differs from the *expected* answer—the forms found in the original text—is considered as an error.[6] In cases where grammatically and semantically valid alternative answers are possible in the given context, the system should be able to recognize the alternatives as "alternative-correct" answers, rather than "wrong", to give correct guidance to the learner. This is especially relevant for advanced learners, who, our research found, insert alternative-correct answers more frequently (Katinskaia and Yangarber, 2021). To investigate the problem of alternative-correct answers, we manually annotated a subset with grammatical errors—first row in Table 1.

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

4.1. Manual Annotation

First, the subset with grammatical errors was manually checked by a Russian language expert. All answers which are actual errors were tagged as errors, other answers which could be potentially correct in the paragraph were marked as alternative-correct (AC). Paragraphs which include at least one AC answer (1 427 paragraphs) were manually double-checked by 6 native speakers. The annotators were not informed whether these paragraphs include AC answers or errors, and were asked to tag the highlighted answers as "correct", "incorrect", or "uncertain".

Annotation was performed using the Tagtog platform,[7] where all annotators were given identical guidelines. Tagtog allows one to annotate not only spans of text with the given set of labels, but also to add comments, which can be used to resolve annotation conflicts and uncertainties. The documents were automatically distributed among the annotators so that we have no less than two annotations for each paragraph.

Pairwise agreement calculated as the percentage of identical tags assigned to the same answers is shown in Figure2. The agreement between annotators is not high. One reason for the low agreement is that the annotators have different opinions regarding what can be considered as grammatically acceptable. For example, annotators who work with language professionally are more strict. Another possible reason is that some anno-

[6] This rule has a some special, language-specific exceptions, where multiple alternative forms may be linguistically equivalent.

[7] tagtog.net

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

tators have a tendency to tag every answer which they do not consider as perfectly fitting the context as "incorrect", while others tend to mark them as "uncertain" and often add some comments.

The annotators also manually assigned a *type* to each tagged answer. For example, if the learner used nominative case instead of genitive, the assigned type will be "Case: genitive/nominative". In the case of verbs, the learner might use a form which is different from the expected one, i.e., transgressive participle rather than present tense 3rd person form ("Verb form: 3 present/transgressive"). If the learner changed multiple features in one word form, they will all be marked.

As we can see from the Table2, most of the alternative-correct answers differ by number (singular vs. plural), tense (present vs. past), or belong to different aspect forms. The corpus contains 1 704 AC answers in total. For a number of answers the annotators could not agree, so all conflicting annotations were resolved by a language expect. There are 34% of all double-checked answers which include at least one annotator who disagrees with all other annotators. In the released corpus these answers are tagged as "hard." The types of these answers are presented in Table3.

4.2. Automatic Annotation

To perform automatic annotation of grammatical error types in both subsets of the learner corpus, we adapted the grammatical error annotation toolkit ERRANT (Bryant et al., 2017). The tool adapted to Russian is called RuERRANT.[8]

ERRANT automatically extracts edit operations from parallel original and corrected sentences and assigns error types to them. Error classification is rule-based that makes it fully independent of the dataset. Its performance depends only on the number and quality of rules and the quality of the parser, which is the main component of the tool. For English, human raters considered 95% of the error types predicted by ERRANT to be either "Good" or "Acceptable". In the future, such evaluations should be performed for RuERRANT as well.

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

We use the SnowballStemmer from the NLTK,[9] and spaCy 3.1.[10] For detecting spelling errors, we use the Hunspell word list and a word list based on the Taiga Corpus.[11]

The tags in RuERRANT are more fine-grained than in the English ERRANT, due to the vast differences in morphology. We have added tags for case, gender, aspect, mood, voice, comparative, and short/full forms of adjectives (i.e., predicative/attributive). The most frequent tags assigned by RuERRANT, their explanations and examples of errors are presented in Table 5. The statistics on the error types in both of the released subsets are in Table 4. Still, many errors—e.g., spelling errors, inflection errors of different parts of speech, etc.—are assigned the generic tag "OTHER". To address this problem and improve annotation quality, we plan to add more language-specific rules.

To evaluate the performance of RuERRANT, we compare the error types that it assigns to answers against error types that were manually assigned by the annotators. Among 1 431 assigned types, 4.9% were incorrect, which matches the performance of the English ERRANT. We did not consider a type tag as incorrect, if it was a general one and includes only a part of speech of a target word—e.g., if an expert annotation was "Tense: past/present" and RuERRANT type

[8] https://github.com/Askinkaty/errant

[9] https://www.nltk.org/

[10] https://spacy.io/

[11] https://tatianashavrina.github.io/taiga_site

Error type	Meaning	Examples
R:Spell	Spelling error	подзимелям → подземельям "*dungeons*"
R:Noun:Case	Case of nouns	дому "*house*" (sg., dat.) → домом "*house*" (sg., instr.)
R:Adp	Preposition	к "*to*" → в "*in*"
R:Noun:Num:Case	Case and number of nouns	саду "*garden*" (sg., dat.) → садами "*gardens*" (pl., instr.)
R:Pron	Pronouns	наш "*ours*" (sg.) → наши "*ours*" (pl.)
R:Verb:Num	Number of verbs	прочитал "*read*" (sg.) → прочитали "*read*" (pl.)
R:Verb:Form	Verb form	прочитать "*to read*" (inf.) → прочитан "*read*" (participle)
R:Noun:Num	Number of nouns	дереву "*tree*" (sg.) → деревьям "*trees*" (pl.)
R:Other	Other	два "*two*" (cardinal) → двое "*two*" (collective)
R:Verb:Tense	Tense of verbs	увидел "*saw*" (past) → увижу "*will see*" (fut.)
R:Det	Demonstrative pronouns	этот "*this*" (masc.) → эта "*this*" (fem.)
R:Adj:Case	Case of adjectives	быстрому "*fast*" (dat.) → быстрый "*fast*" (nom.)
R:Verb:Gender	Gender category of verbs	сделал "*did*" (masc.) → сделала "*did*" (fem.)
R:Morph	Morphology	ошибочный "*mistaken*" (adj.) → ошибочно "*mistakenly*" (adv.)
R:Verb:Aspect	Aspect of verbs	бежал "*was running*" (imperf.) → прибежал "*run*" (perf.)
R:Verb:Infl	Inflection of verbs	бросишь "*throw*" (2nd person) → брошу "*throw*" (1st person)
R:Adj:Num	Number of adjectives	красивые "*beautiful*" (pl.) → красивый "*beautiful*" (sg.)
R:Adj:Num:Case	Case and number of adjectives	тонком "*thin*" (sg., loc.) → тонкими "*thin*" (pl., instr.)
R:Adj:Full/Short	Full and short adjective forms	непоседливый "*restless*" (full) → непоседлив "*restless*" (short)
R:Adj:Gender	Gender of adjectives	доброе "*kind*" (neut.) → добрая "*kind*" (fem.)
R:Aux	Auxiliary verb forms	был "*was*" (past) → буду "*will be*" (fut.)

Table 5: Types of grammatical errors assigned automatically by RuERRANT.

was "VERB". Most errors in the automatic annotation were caused by the inability of spaCy to correctly disambiguate some forms. For example, the nominative plural vs. genitive singular forms of a noun "*письма*" ("letters" or "of a letter").

5. Annotators vs. a Neural Model

An interesting part of the annotated data is where the annotators cannot agree whether the learner's answer is correct in the given context. To investigate how a neural model will perform on this data, we took one baseline model which was trained to do the same task as the task given to human annotators: to tag a word in context as grammatically correct or incorrect. The model is a pre-trained BERT-base, which was fine-tuned on synthetic training data. This data was generated based on error-free text by inserting random grammatical errors. The process of training data generation and the model specification is described in (Katinskaia and Yangarber, 2021).

Our goal was to estimate the model's confidence on those learner answers where annotators could not agree with each other and assigned different tags (including the tag "uncertain"). We call these answers "hard". We applied the method of Monte Carlo dropout (Gal and Ghahramani, 2016) to estimate the uncertainty: keeping the dropout activated at test time, we repeatedly sampled 20 predictions for every input, and estimated the predictive uncertainty by measuring the variance and entropy of the scores.

Figure3 shows a plot of the variance of the predicted scored against the percentage of answers which were "hard" for annotators, for each manually assigned type of error. As we can see from the figure, the uncertainty of annotators tends to correlate with the uncertainty of the model, with Pearson correlation coefficient 0.69, with a p-value of 0.002, 95% CI 0.32 to 0.88.

It is an interesting observation and application of the annotated learner corpus. It shows that we can trust the estimations of model uncertainty in experiments with automatic assessment of grammatical correctness.

6. Conclusion and Future Work

In this paper we presented a new version of a longitudinal Revita Learner Corpus for Russian. This version contains paragraphs with multiple answers to exercises that were given by learners *simultaneously*, within a given snippet. The exercises are generated automatically by the Revita language-learning system. All correct answers and errors and were automatically and manually annotated. The data also include information about which learner answers were hard for annotators to agree upon.

Together with the learner data, we release a version of ERRANT, an error annotation toolkit, adapted to Russian. It can be improved further by using better models for morphological analysis. By releasing RuERRANT, we hope to contribute to research on various problems related to grammatical correctness and error correction for Russian.

One of the main advantages of ReLCo is that it is constantly growing, as learners continue practicing with the system. As future work, we plan to improve the automatic assessment of grammatical correctness of learner answers and to introduce new types of exercises to the system, which will allow insertion of longer chunks of text, including full-scale essays.

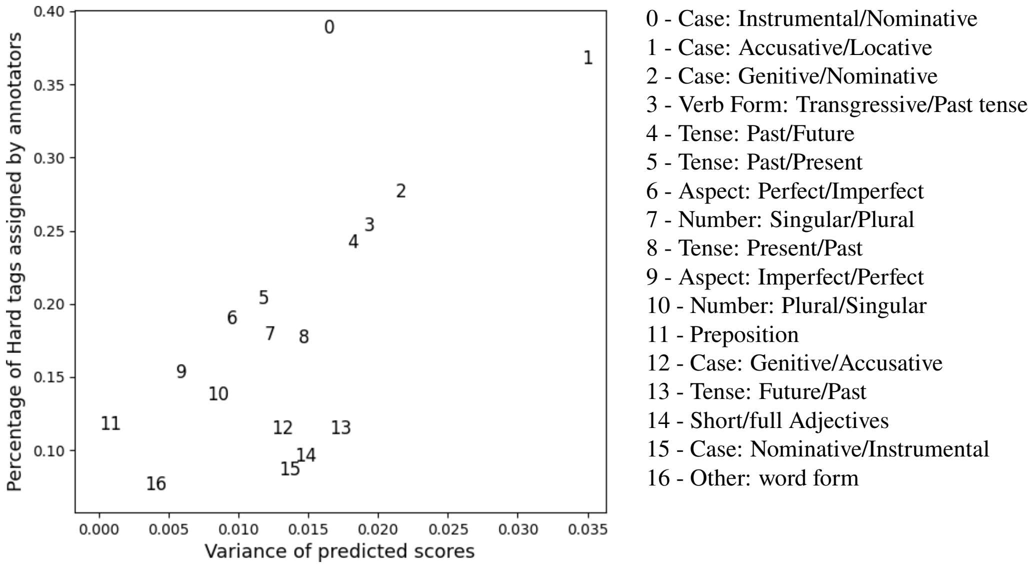


Figure 3: Uncertainty of human annotators vs. uncertainty of a neural model, both assessing grammatical correctness of learners answers. On the right: 17 types of errors annotated.

Acknowledgements

We thank Master's students from Pushkin State Russian Language Institute and bachelor students from the Institute of Linguistics, Russian State University for the Humanities, for their generous help with annotation. This work was supported in part by the Academy of Finland, HIIT—Helsinki Institute for Information Technology, BusinessFinland (Grant Revita, 42560/31/2020), and Tulevaisuus Rahasto (Future Development Fund), University of Helsinki.

7. Bibliographical References

Bryant, C., Felice, M., and Briscoe, E. (2017). Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics.

Felice, M., Bryant, C., and Briscoe, E. (2016). Automatic extraction of learner errors in esl sentences using linguistically enhanced alignments. Association for Computational Linguistics.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Gilquin, G. and Granger, S. (2015). From design to collection of learner corpora. *The Cambridge handbook of learner corpus research*, 3(1):9–34.

Granger, S. (1996). From ca to cia and back: An integrated approach to computerized bilingual and learner corpora.

Granger, S. (2002). A bird's-eye view of learner corpus research. *Computer learner corpora, second language acquisition and foreign language teaching*, 6:3–33.

Hou, J., Koppatz, M. W., Quecedo, J. M. H., Stoyanova, N., Kopotev, M., and Yangarber, R. (2019). Modeling language learning using specialized Elo ratings. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, ACL: 56th annual meeting of the Association for Computational Linguistics*, pages 494–506.

Katinskaia, A. and Yangarber, R. (2018). Digital cultural heritage and revitalization of endangered Finno-Ugric languages. In *Proceedings of the 3rd Conference on Digital Humanities in the Nordic Countries*, Helsinki, Finland.

Katinskaia, A. and Yangarber, R. (2021). Assessing grammatical correctness in language learning. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 135–146, Online, April. Association for Computational Linguistics.

Katinskaia, A., Nouri, J., and Yangarber, R. (2018). Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Katinskaia, A., Ivanova, S., and Yangarber, R. (2019). Multiple admissibility: Judging grammaticality using unlabeled data in language learning. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 12–22, Florence, Italy, August. Association for Computational Linguistics.

Katinskaia, A., Ivanova, S., and Yangarber, R. (2020). Toward a paradigm shift in collection of learner corpora. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 386–391.

Kutuzov, A. and Kunilovskaya, M. (2014). Russian learner translator corpus. In *International Conference on Text, Speech, and Dialogue*, pages 315–323.

Springer.

Polinsky, M., Brinton, D., Kagan, O., and Bauckus, S. (2008). Heritage language education: A new field emerging.

Protassova, E. (2016). Narrative. frog stories in Russian: 41 transcripts–ages 5, 6, 7, 8, 9, 10, and adult.

Rakhilina, E., Vyrenkova, A., Mustakimova, E., Ladygina, A., and Smirnov, I. (2016). Building a learner corpus for Russian. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 66–75.

Rozovskaya, A. and Roth, D. (2019). Grammar error correction in morphologically rich languages: The case of russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Segalovich, I. and Titov, V. (1997). Mystem.

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes*. Harvard University press.

Vygotsky, L. S. (2012). *Thought and language*. MIT press.

Yatsenko, A. A. A., Kisselev, O. V., and Freels, S. G. (2012). Results 2012: Using flagship data to develop a russian learner corpus of academic writing. *Russian Language Journal/*, 62:79–105.

Zevakhina, N. and Dzhakupova, S. (2015). Corpus of Russian student texts: design and prospects. In *Proceedings of the 21st International Conference on Computational Linguistics Dialog, Moscow*.