*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 386–391
Marseille, 11–16 May 2020

# Toward a Paradigm Shift in Collection of Learner Corpora

**Anisia Katinskaia,[1,2] Sardana Ivanova,[1,2] Roman Yangarber[2]**
[1] Department of Computer Science, University of Helsinki, Finland
[2] Department of Digital Humanities, University of Helsinki, Finland
first.last@helsinki.fi

## Abstract

We present the first version of the longitudinal Revita Learner Corpus (ReLCo), for Russian. In contrast to traditional learner corpora, ReLCo is collected and annotated fully automatically, while students perform exercises using the Revita language-learning platform. The corpus currently contains 8 422 sentences exhibiting several types of errors—grammatical, lexical, orthographic, etc.—which were committed by learners during practice and were automatically annotated by Revita. The corpus provides valuable information about patterns of learner errors and can be used as a language resource for a number of research tasks, while its creation is much cheaper and faster than for traditional learner corpora. A crucial advantage of ReLCo that it grows continually while learners practice with Revita, which opens the possibility of creating an unlimited learner resource with longitudinal data collected over time. We make the pilot version of the Russian ReLCo publicly available.

**Keywords:** L2 learner corpus, grammatical error detection, grammatical error correction, CALL, ICALL, Revita, low-resource languages

## 1. Introduction

The most widely used 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[1] purpose in a standardized format." Granger (2002) notes that in the context of foreign/second language learning the notion of *authenticity* is problematic: language is learnt in classrooms and learner data is rarely fully natural. Even when learners are free to write what they like, they are usually constrained by topic and time limits.

Language teachers and researchers in language teaching and learning have been collecting learner corpora for decades. Most of the corpora available today contain English as the 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,[2] 93 are English learner corpora (over 53%). We can observe that most languages, except English, Spanish, French, and German, are low-resourced in relation to available learner data. For example, only two Russian leaner corpora appear in the list.[3]

Creation of learner corpora is an extremely labor-intensive task, which involves collecting text samples, transcribing, annotating/classifying errors, and providing corrections for all errors that the learners made, as well as collecting metadata. Nevertheless, this massive effort is justified, because learner corpora are useful for a wide range of critical tasks. These tasks include discovering and studying:

- common errors and patterns of errors that learners (or certain groups of learners) make,
- common learning paths—how and in what order learners acquire linguistic skills,
- how learner characteristics—mother tongue, age, level of education, etc.—affect patterns of learning.

All of these tasks ultimately improve the teaching and learning process. Data from learner corpora can be directly used in the classroom, e.g., for creating new exercises. Text produced by different learners  can clarify certain phenomena that are not mentioned in grammar books (Granath, 2009).

The prevalent tendency in studies of learner language is to collect corpora for specific experiments, and then either discard them or not make them fully available—from the same list of 174 corpora, only several can be freely downloaded. This makes the learner data not usable for research and creates obstacles to collaboration in developing better language resources (Nesselhauf, 2004).

In this work we propose a *new paradigm* for creating learner corpora—building them automatically based on an existing language learning platform, in particular, the Revita platform, (Katinskaia et al., 2017; Katinskaia et al., 2018; Katinskaia and Yangarber, 2018).[4] The learner corpus is collected continuously and annotated *automatically* while students practice the language by performing a variety of exercises. All collected learner data is used in Revita for generating new exercises—it creates the "learning feedback loop" helping students to improve their language skills more effectively. We make available the first version of the Russian ReLCo [5] and hope to encourage collaborators to join in improving the learner datasets and in performing research experiments based on collected data.

The remainder of this paper is structured as follows. In Section 2, we review several learner corpora created for languages that are available in Revita. The Section 3 presents the main features of Revita Learner Corpora, how data is collected, and automatically annotated. In Section 4 we describe the Russian ReLCo, its annotation, the problem

[1] Second Language Learning, Foreign Language Teaching

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

[3] http://www.web-corpora.net/RLC/, http://rus-ltc.org/search

[4] revita.cs.helsinki.fi

[5] https://github.com/Askinkaty/Russian_learner_corpora

of Multiple Admissibility, and analysis of collected errors. Conclusions and future work are discussed in the final section.

## 2. Related work

The typology of learner corpora takes into account several dimensions: written vs. spoken transcribed data; general vs. Language for Specific Purposes (LSP) learner corpora; target language; learner's mother tongue; synchronic vs. diachronic; global vs. local (collected by a teacher among their students); commercial vs. academic (Gilquin and Granger, 2015).

In the context of this paper and Revita ReLCo, we are interested in written learner corpora of Finnish and Russian. The Centre for English Corpus Linguistics lists five learner corpora for Finnish:

1. "Linguistic Basis of the Common European Framework for L2 English and L2 Finnish" (CEFLING) is a collection of writing samples from adults taking the National Proficiency Certificate exams and from young learners of 7–9 grades (Martin et al., 2012).
2. "Paths in Second Language Acquisition" (TOPLING). It contains 1194 samples for L2 Finnish on various tasks.
3. The Advanced Finnish Learner Corpus (LAS2). It is a collection of three academic genres: exam essays, theses and article manuscripts, and texts for studying purposes (631 402 tokens in 640 texts). LAS2 has morphological and syntactic annotation (Ivaska, 2014).
4. The Finnish National Foreign Language Certificate Corpus (YKI) is compiled from the National Certificates of Language Proficiency Examinations. The YKI corpus covers learners at the basic, intermediate, and advanced levels.
5. The International Corpus of Learner Finnish (ICLFI) contains approximately 1 million tokens. Errors are annotated in only 5% of the corpus (Brunni et al., 2015).

Several projects focus on Russian learner language:

1. Russian Learner Corpus (RLC) (Rakhilina et al., 2016);
2. The Corpus of Russian Student Texts, which includes academic writing (Zevakhina and Dzhakupova, 2015) produced by native speakers of Russian;
3. Narrative collections (Protassova, 2016; Polinsky et al., 2008);
4. Russian Learner Translator Corpus (Kutuzov and Kunilovskaya, 2014) which is a bi-directional multiple corpus of English-Russian translations done by university translation students.

Russian Learner Corpus (RLC) contains around 1.6 millions words. It is a collection of oral and written texts by "heritage" and L2 speakers, which includes morphological and error annotation. Morphological annotation was marked automatically by the MyStem (Segalovich and Titov, 1997) morphological analyser, 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 text itself (written or oral, genre, and a time limit).

## 3. Revita Learner Corpora

As mentioned in the Introduction, ReLCo is collected while students practice with a variety of exercises on the Revita language-learning platform. Revita is an online L2 learning system for learners beyond the beginner level. It covers several languages, most of which are highly inflectional, with rich morphology. Revita allows learners to practice with authentic texts, which can be chosen and uploaded to the platform by the learner herself, or by a teacher for a group of learners. The system creates a variety of exercises automatically trying to adapt the level of exercises to every user depending on her level of proficiency. A continuous assessment of the users' answers is also performed automatically (Hou et al., 2019).

At present, Revita provides no mode for submitting essays, so the data collected is based on pre-existing texts.

Despite this limitation, ReLCo presents:

- authentic learners errors in context;
- the time when the errors were made;
- unique internal identifiers (IDs) of the learner;
- the types of exercises which were practiced.

This makes ReLCo a valuable resource of learner data, which can be used for improving teaching and learning processes.

Our main thesis in this paper is that although ReLCo is not a "conventional" learner corpus, it provides many of the same benefits as "proper" learner corpora, which make it equally valuable in many situations where learner corpora are used—while *collecting* it is much cheaper and faster, since annotation is fully automated and costs nothing. A key advantage of ReLCo is that it *grows constantly* as learners do more exercises. We need no extra effort to collect learner data—it happens as part of the learners' regular coursework. Manual work, which always accompanies the creation of traditional learner corpora, is drastically reduced.

Further, since Revita tracks all learner interactions with the system, we can extend ReLCo with valuable information that is usually not present in classic learner corpora, such as:

- for which words in text the user requested translations;
- how many attempts the learner needed to solve an exercise;
- did the learner repeat the same error, or made new errors in the given context.

Since our main goal is ICALL—building Intelligent Computer-Assisted Language Learning systems—all features mentioned above make ReLCo especially attractive in the context of the following tasks:

1. Detecting patterns of L2 errors through time, which can be leveraged for generating new exercises relevant for a particular learner;

2. Common errors can be used for creating *distractors*—option answers for multiple-choice exercises; the distractors can be offered to *new* learners, to check whether they will also commit errors found to be common.
3. Detecting and correcting grammatical errors;
4. Detecting the effect of different exercise types on learner responses;
5. Developing models of learner knowledge states.

Analysis of learner corpora and collecting learner data are important for creating gold-standard data for NLP models for learner language. For example, predicting learner mistake patterns was at the focus of the Duolingo Shared Task, 2018.[6] This is a crucial research problem for any language learning system, which attempts to make the process of L2 acquisition more effective.

Considering the application of ReLCo for grammatical error detection/correction, we should stress that this problem is currently in the research focus primarily for English. Very few researchers focus on other languages, e.g., German (Boyd, 2018).

The ReLCo approach to collecting learner data, which is available and growing, will help strengthen the link between learner corpora and intelligent tutoring systems—seen as a key future direction in learner corpora (Meurers, 2015).

## 4. Learner Corpora for Russian

We next describe an initial, pilot version of ReLCo for Russian. This data includes answers to exercises which were practiced during 2017–2019 by 150 learners. The corpus is still rather small: around 1.35 million tokens, 8 422 sentences in total. Nevertheless, it is comparable in size with the released Write & Improve+LOCNESS corpus (Bryant et al., 2019; Granger, 1998) for the Low-resourced track at the BEA Shared Task on Grammatical Error Correction (Bryant et al., 2019).[7] The W&I+LOCNESS corpus contains 801 361 tokens, including correct sentences.

Every sentence in the Russian ReLCo includes answers given by students to the following types of exercises:

- "cloze" exercises (fill-in-the-blank) with the lemma of the missing word given as a hint;
- multiple-choice exercises—with distractors generated for many kinds of exercises;
- listening exercises.

The learner receives the text one "snippet" at a time (about 1 paragraph), with several words replaced by exercises of the types listed above. Learners are given more than one attempt to answer: if the first try was unsuccessful, the learner receives additional hints, and can try the exercises again.

Revita expects the learner's answer to be the same as the form used in the base text. Errors fall into 3 groups:

- *Grammatical errors*: the given answer has the same lemma as the expected answer;
- *Non-word error*: the given answer does not correspond to a valid word;
- *Different lemma*: the given answer has a lemma which is different from the lemma of the expected answer. If the learner changes the lemma, the answer is considered to be wrong.

The number of sentences including different error types in ReLCo is presented in Table 1. Sentences have approximately 2 errors on average.

The data format is simple for processing: CSV files, where each line has a user id (a randomized key); the timestamp of the answer; the sentence with the errors, which were made simultaneously at this time; the number of attempts; the correct corresponding sentence—in a separate csv file.[8]

| Subset | #Sentences | #Tokens | #Errors per sentence |
|---|---|---|---|
| Grammar | 5 361 | 879 154 | 1.9 |
| Non-word | 2 166 | 332 591 | 1.9 |
| Diff. lemma | 895 | 138 597 | 1.9 |

Table 1: The Russian ReLCo. "Grammar" is the subset of the corpus with grammatical errors: answers that have the same lemma as the correct answer. "Non-word" is the subset with orthographic errors, which do not correspond to any real word form. "Diff. lemma" is the subset with answers whose lemma is different from the lemma of the correct answer.

Additional exercise types, which will allow the learner to make certain types of errors—word order, omission/deletion/insertion—are planned for future releases of the system. Currently, the corpus mostly contains grammatical errors, which are difficult to annotate in traditional learner corpora. For example, the Koko corpus of German (Abel et al., 2014) mainly has non-grammatical errors annotated, with grammatical error annotation left for future work.

### 4.1. Automatic Annotation

Texts are tokenized and analyzed using language-specific analyzers when they are uploaded to Revita.[9]

All orthographic errors are annotated automatically. This annotation is based on the output of a morphological analyzer: if the analyzer return no analysis for a word, it does not exist in the vocabulary, and we consider it as a word with orthographic errors (i.e., non-word errors).

We are mostly interested in grammatical errors, which are also annotated automatically by Revita. If a given answer is not accepted by the system but it has the same lemma as the expected answer, the system marks it as a grammatical error. This approach has limitations because some grammatical errors could have a lemma which differs from the lemma of the expected answer. At present, all of these grammatical

[6] http://sharedtask.duolingo.com/

[7] https://www.cl.cam.ac.uk/research/nl/bea2019st/

[8] The corpus provides no metadata, because we do not ask the learners' age, sex, nationality, mother tongue, proficiency levels, etc. For some learners we know the native tongue and CEFR levels from their teachers; this information is anonymized and used for developing the Revita platform, but is not shared publicly.

[9] For Russian we use Crosslator (Klyshinsky et al., 2011).

errors are considered to be errors of type “different lemma” errors (i.e., wrong word choice). In addition to this problem, the current approach of annotating errors suffers from several limitations discussed in the following subsection.

At the moment, Revita resolves ambiguity by disambiguation rules, e.g., by leveraging agreement of adjectives and nouns in a noun phrase, etc. Therefore, many errors in the Russian ReLCo can be additionally automatically annotated as errors in agreement or government, because these relations were identified during disambiguation.

## 4.2. Multiple admissibility

Since the answer is checked automatically, the system should be able to accept more than one expected answer, if there are grammatically and semantically valid alternatives in the given context. Otherwise, it returns negative (actually incorrect) feedback to the learner. This problem, when more than one answer is valid in the given context, was formulated as Multiple Admissibility (MA) in (Katinskaia et al., 2019).

This means that Russian ReLCo includes some answers which are actually correct but automatically annotated as grammatical errors.[10] In the experiments, we took a subset of 2 884 errors from the database and manually annotated all errors which were marked by Revita as grammatical. A subset of errors made by advanced learners was annotated separately. We found that 7.5% of errors are possible alternative answers, for advanced learners this number is twice as high. Therefore, around 8% of errors in the Russian ReLCo are not true grammatical errors.

We have begun work on detecting correct answers in the context automatically: a baseline neural model is able to detect valid alternative answers with an accuracy of 85.9% on a modest test set (Katinskaia et al., 2019). We continue these experiments and plan to update Russian ReLCo with improved error annotations as well as to enhance Revita with a new error detection tool.

## 4.3. Error Types

To investigate the Russian ReLCo, we have randomly sampled for manual annotation 714 errors marked by Revita as grammatical errors. All selected errors were annotated by two native Russian speakers with an agreement rate of 91%. Cases where annotators did not agree were resolved by consensus. The manually annotated part is made available.

Errors were annotated in the following way: first, an annotator marks which grammatical category was mistaken and what exactly was the mistake. For example, if the learner made a mistake in the category of case by using nominative instead of genitive, the error is marked as “Case: nom/gen”. In case of verbs, the learner can use a form which is different from the expected one, i.e., transgressive participle instead of a 3rd person present tense verb form (“Verb form: transgr/3 pres”). If the learner made several errors in one word form, they all will be marked. In case an error was, in fact, a correct answer (an instance of MA), annotators specified which category is different from the expected answer. Statistics across the most frequent categories for grammatical errors and MA instances is presented in Table 2.

| Error category | Count | MA category | Count |
|---|---|---|---|
| Case | 240 | Number | 61 |
| Number | 131 | Tense | 27 |
| Verb form | 67 | Verb form | 16 |
| Gender | 51 | Case | 11 |
| Tense | 22 | Short/full adj | 5 |
| Short/full adj | 11 | Gender | 1 |
| Person | 10 | Compar | 1 |

Table 2: Categories of the most frequent grammatical errors and multi-admissible answers made by non-native learners

Since we observed a difference between advanced and other learners, we asked 30 native speakers to perform “cloze” exercises. We assume that the higher the level of language skills of an advanced learner, the closer her answers will be to the answers of native speakers. We also think that studying variations of native answers could help in our experiments with identifying MA. We manually annotated 526 answers from native speakers, all of which were marked by Revita as “grammatical errors.” Answers of natives have 4 times more alternative answers (MA) than those of other users, because they know the language better and see more possible options in the context.

If we compare the most frequent grammatical errors and multi-admissible answers of native and non-native learners, we notice that the natives make fewer grammatical errors in general. For example, non-native learners often mistake forms in the locative case for forms in the accusative case. It is 3.5% of errors for non-native learners in the annotated dataset. For natives it is only 0.1% of all errors. MA answers of natives also differ: e.g., natives often replace verb forms with transgressive forms or use instrumental case instead of nominative case where it is possible. These variations characterise advanced level of grammatical skills.

Table 4 presents more detailed statistics on the distribution of the annotated grammatical errors for non-native learners. A detailed analysis of learner errors is beyond the scope of this paper.

| Error category | Count | MA category | Count |
|---|---|---|---|
| Case | 60 | Number | 117 |
| Number | 50 | Tense | 78 |
| Verb form | 40 | Verb form | 33 |
| Gender | 13 | Short/full | 32 |
| Tense | 10 | Case | 27 |
| Other | 4 | Spelling variant | 18 |
| Person | 2 | Other | 7 |

Table 3: Categories of most frequent grammatical errors and multi-admissible answers made by native speakers

# 5. Conclusions and Future Work

In the current work we presented the first version of a longitudinal Russian Revita Learner Corpus (ReLCo), which is

[10] “Different lemma” errors could also be valid, but we have not yet manually annotated this part of the corpus to study the percentage of MA cases in this group.

| Error type | Count |
|---|---|
| Number: Singular/Plural | 95 |
| Number: Plural/Singular | 63 |
| Case: Nominative/Genitive | 46 |
| Case: Locative/Accusative | 29 |
| Gender: Masculine/Feminine | 20 |
| Case: Nominative, Accusative/Genitive | 14 |
| Case: Locative/Genitive | 14 |
| Case: Dative/Genitive | 13 |
| Gender: Feminine/Masculine | 13 |
| Case: Genitive/Nominative | 13 |
| Full/short adjective | 12 |
| Case: Nominative/Locative | 11 |
| Case: Nominative/Dative | 10 |

Table 4: Most frequent error types made by non-native learners

collected while learners practice with exercises on the language learning platform Revita. All errors in ReLCo are annotated automatically. We claim that, since creating a traditional learner corpora is a highly labor-intensive and expensive task, useful learner data can be collected automatically with no intervention into the learning process. This approach allows us to collect potentially unlimited amounts of data representing the learning process through time, rather than making "snapshots" of learner knowledge states.

As future work, we investigate neural approaches for resolving all ambiguous morphological annotations in the Russian ReLCo and correcting answers which were mistakenly annotated as errors. Since Revita includes many other languages, we plan to release learner corpora for these languages as well.

Although the current ReLCo is limited by answers provided in the given context, we are working on adding a new mode, where learners can write essays and receive an instant feedback about errors.

## 6. Acknowledgements

Work was supported in part by the Academy of Finland, Project FinUgRevita, Grant No. 267097, and by HIIT—Helsinki Institute for Information Technology.

## 7. Bibliographical References

Abel, A., Glaznieks, A., Nicolas, L., and Stemle, E. (2014). Koko: an l1 learner corpus for German. In *LREC*, pages 2414–2421.

Boyd, A. (2018). Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 4th Workshop on Noisy User-generated Text*. Association for Computational Linguistics.

Brunni, S., Lehto, L.-M., Jantunen, J. H., and Airaksinen, V. (2015). How to annotate morphologically rich learner language. principles, problems and solutions. *Bergen Language and Linguistics Studies*, 6.

Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75.

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

Granath, S. (2009). Who benefits from learning how to use corpora? *Corpora and language teaching*, 33:47.

Granger, S. (1998). *The computer learner corpus: a versatile new source of data for SLA research.*

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.

Ivaska, I. (2014). The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish. *Apples: journal of applied language studies*.

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., Nouri, J., and Yangarber, R. (2017). Revita: a system for language learning and supporting endangered languages. In *6th Workshop on NLP for CALL and 2nd Workshop on NLP for Research on Language Acquisition, at NoDaLiDa*, Gothenburg, Sweden. Linköping University Electronic Press.

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.

Klyshinsky, E., Kochetkova, N., Litvinov, M., and Maximov, V. Y. (2011). Method of POS-disambiguation using information about words co-occurrence (for Russian). *Proc. of GSCL*, pages 191–195.

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

Martin, M., Alanen, R., Huhta, A., Kalaja, P., Mäntylä, K., Tarnanen, M., and Palviainen, Å. (2012). Cefling: Combining second language acquisition and testing approaches to writing. *Learning To Write Effectively: Current Trends In European Research. Edited by Torrance, M., Alamargot, D., Castelló, M., Ganier, F., Kruse, O., Mangen, A., Tolchinsky, L., & Van Waes, L. ISBN 978-1-78052-928-8.*

Meurers, D. (2015). Learner corpora and natural language processing. *The Cambridge handbook of learner corpus research*, pages 537–566.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. *How to use corpora in language teaching*, 12:125–156.

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.

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

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*.