Building a learner corpus for Russian*

Ekaterina Rakhilina **Anastasia Vyrenkova** **Elmira Mustakimova**
National Research University Higher School of Economics
21/4, Staraya Basmannaya Ulitsa, 105066, Moscow, Russia
erakhilina,avyrenkova,emustakimova@hse.ru

Alina Ladygina
Eberhard Karls Universität Tübingen
19-23, Wilhelmstrasse,
72074, Tübingen, Germany
aladygina@yahoo.com

Ivan Smirnov
Sholokhov Moscow State
University for the Humanities
3/7a Vladimirskaya Ulitsa,
111123, Moscow, Russia
smirnof.van@gmail.com

Abstract

In this paper we describe an open learner corpus of Russian. The Russian Learner Corpus (RLC) is the first corpus with clear distinction between foreign language learners and heritage speakers. We discuss the structure of the corpus, its development and the annotation principles. This paper describes the platform of the RLC which combines online tools for text uploading, processing, error annotation and corpus search.

1 Introduction

Designing learner corpora has become a rapidly developing branch of corpus linguistics, which is accounted for by obvious reasons — both research and practical. As annotated collections of texts produced by non-native speakers of a certain language, learner corpora open up new horizons in areas, such as quantitative studies in second language acquisition, contrastive interlanguage analysis (Granger, 1996), etc., and the urge for well-organized error classification and frequency based error analysis can hardly be overestimated for language teaching. Moreover, computerized learner data serve as training and test data sets for various NLP tasks, such as native language identification task (Jarvis and Paquot, 2015), automatic error detection (Leacock et al., 2014), etc. The goal of this paper is to present a recently created Russian Learner Corpus (RLC)[1]. The novelty of the RLC is threefold:

1. It is the first open learner corpus for the Russian language enabling search over lemma, grammatical features, and error tags;
2. It is the first learner corpus that draws a clear distinction between HL (heritage language)[2] and L2 (second language) speakers;
3. It is built on an integrated multifunctional platform that provides a single interface for uploading, annotating and search.

Russian Learner Corpus is an international project carried out by the Linguistic Laboratory for Corpus Technologies at the Higher School of Economics in close collaboration with experts from more than 10 countries (see "Our partners" at http://www.web-corpora.net/RLC). The corpus currently comprises more than 730000 tokens. 56 per cent of the data is produced by L2 learners of Russian, 44 per cent - by heritage speakers of Russian, who are college/university-age students at the proficiency

*Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

[1] RLC is available at http://web-corpora.net/RLC.

[2] Heritage speakers are a special type of bilinguals who grew up in a non-native language environment, but use their native language at home or to communicate with their family (see (Valdés, 2000))

Ekaterina Rakhilina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina and Ivan Smirnov 2016. Building a learner corpus for Russian. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016*. Linköping Electronic Conference Proceedings 130: 66–75.

level of intermediate and higher. The first version of the RLC contained only texts from American English-dominant speakers of Russian. The number of dominant languages has by far grown to eight. Three of them are at the moment scarcely presented in the corpus, however, more data on them and two more languages are being prepared for upload. A valuable part of the RLC is a large longitudinal subcorpus of academic writing called RULEC collected by Olessya Kisselev and Anna Alsufieva. All the respondents signed a special consent form and their names are anonymized in the corpus. In the longitudinal RULEC the speakers were assigned fake names so that the user could easily trace the progress of each student. Other respondents are assigned a unique students code.

In Section 2 we give an overview of similar projects developed for the Russian language. Section 3 describes corpus data and metainformation provided to each text. Section 4 presents annotation principles, and Section 5 focuses on characteristics of the corpus platform. In Section 6 we will make some concluding remarks and discuss our future work.

2 Related works

To date there have been several projects focusing on Russian as a target language for learners. Among them are studies based on collections of narratives (Protassova, 2016; Isurin and Ivanova-Sullivan, 2008; Polinsky, 2008), academic writing repositories (e.g. Corpus of Russian Students Texts (Zevakhina and Dzhakupova, 2015), ReBiSlav[3]) and learner translations (Russian Learner Translator Corpus, see (Kutuzov and Kunilovskaya, 2014). Despite their obvious usefulness and prominence, however, none of these projects can be named a full-fledged learner corpus for the reasons that we outline below.

Narrative collections present a huge interest for teachers and linguists studying non-native speech (see (Pavlenko, 2008) for more detail), yet those that are listed above are relatively small, closed and used for specific purposes of a researcher, which inevitably entails heterogeneity in annotation principles. The Corpus of Russian Student Texts (CoRST) is an open annotated resource, however, it consists of the samples of academic writing produced by native speakers of Russian, and thus the understanding of the term *learner* for this project is not at all common, as it implies the process of mastering a new register of Russian by native Russian speakers. Translation corpora are traditionally granted a special status primarily because they contain constrained language production and should be particularly designed.

Another important feature of the RLC is that it allows for differentiating between heritage and L2 production. Contrasting heritage speakers and L2 learners has attracted much attention from both pedagogical and theoretical researchers in the recent decades. At the same time learner corpora do not normally incorporate data on heritage production or probably do not make any clear distinction between heritage and L2 texts. There are also several collections solely devoted to Heritage Russian data (Corpus for Heritage Language Variation and Change[4], (Polinsky, 2008). However, to our knowledge, by far there have been no open annotated resources covering both L2 and heritage data.

Thus, the Russian Learner Corpus (RLC) that we wish to present here is the first collection of oral and written texts by Heritage and L2 speakers of Russian integrated under a single interface and annotated according to single principles.

3 Data

The corpus consists of two data subsets: the first subset is composed of texts produced by second language learners of Russian (about 2000 texts), and the second one contains written texts and transcripts of speech by Russian heritage speakers (about 1500 texts). The texts were collected by our colleagues who are teaching Russian as a second or heritage language and/or making their research in SLA and heritage linguistics. The students filled in the form of consent and a sociolinguistic questionnaire. RLC represents the texts of Russian language learners who have 5 different dominant languages (American English, French, Korean, Kazakh, and German,

[3] http://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/rund-ums-institut/korpora/rebislav

[4] http://projects.chass.utoronto.ca/ngn/HLVC/0_0_home.php

including Swiss German).

Furthermore, the data from Italian, Serbian, Japanese, Swedish, Norwegian and Dutch students are going to be released soon. The sampling of dominant languages is explained by two reasons: we were initially aiming at presenting typologically different L1 in the sample, and presenting texts coming from both foreign and Post-Soviet countries. Currently, however, there is a bias in the data that is spelled out by its accessibility. The corpus also represents various genres: essays and summaries of the articles on social, cultural, historical, political and ecological topics, abstracts of term-papers, biographical stories, blogs and narratives (cartoon and pictures description).

3.1 Metadata

In order to successfully use a corpus a researcher should be provided with relevant information about the origins and specific features of the data, i.e. metadata. Well-organized metadata enables setting various options for individual subcorpora thus providing efficient search and broader opportunities for data analysis. It also gives a clear picture of overall corpus statistics.

According to Tono (2003), there are three major categories in learner corpora design: (a) language-related criteria (e.g. mode, medium, genre, topic), (b) task-related criteria (e.g. longitudinal vs. cross-sectional; spontaneous vs. prepared), and (c) learner-related criteria (e.g. EFL or ESL, age, gender, mother tongue, overseas experience). This classification served as a starting point for developing metatextual markup for RLC and led us to determining a set of 8 metadata items grouped into 2 categories: author-related and text-related.

Among author-related items are author's unique code, gender, language background (HL vs L2), dominant language, proficiency level and educational type. Proficiency level is ascribed according to the Common European Framework of Reference for Languages (CEFR) and American Council on the Teaching of Foreign Languages (ACTFL). These scales are most commonly used by language teachers in the USA and Europe, and the majority of texts in the corpus are authored by students with the proficiency level thus assessed. For search unification, we have introduced three general tags for student proficiency ("Beginner", "Intermediate" or "Advanced"), they also allow to specify the level of students attested in line with other principles. Proficiency level is assigned by the teacher against the scale they work with.

The text related data include mode (oral or written), genre, and time limit. The list of genres available for the corpus metadata was developed in collaboration with the teachers of Russian as a Foreign Language from our partner universities and represents the most common tasks that students of Russian complete to train free production skills (listed in Table 1 - 2). A more elaborated system of genres is presented in RULEC. We ask our partners to provide only free production data, however we dont have any exact information on whether students use any reference materials. In some cases the reliance on extra sources can be inferred from the task (cf. paraphrase or book description).

Category	Description
Authors id	
Gender	male vs. female
Language background	L2 learner vs. heritage speaker
Dominant language	American English, French, German (including Swiss German), Korean, Kazakh, Norwegian, Italian, Serbian
Proficiency level	Beginner / Intermediate / Advanced
Scale	CEFR: A1-C2 ACTFL: Beginner Novice - Advanced High
Educational program type	intensive vs. regular course, course for heritage speakers, etc.

Table 1: Author-related metadata

Category	Description
Mode	Written / Oral
Genre	Answers to questions, academic essay, non-academic essay, blog, letter, story, paraphrase, definition, biography, description, summary
Time limit	limited / unlimited

Table 2: Text-related metadata

4 Annotation

The texts in RLC are provided with morphological and error annotation.

Morphological markup is carried out automatically with help of the morphological analyzer MyStem (Segalovich and Titov, 1997). The tag set of 52 morphological labels[5] meets the standard established by Russian National Corpus (ruscorpora.ru). However, morphological ambiguity is not resolved automatically: every ambiguous word is provided with all possible grammatical analyses, so the texts need to be manually disambiguated.

While designing the error annotation scheme for RLC, we took into account annotation schemes used in other learner corpora, such as ICLE (Granger, 1998), Cambridge Learner Corpus (Nicholls, 2003), FALKO (Reznicek et al., 2012) and (Štindlová et al., 2013). Although these tagsets differ in granularity and error categories, it was necessary to compare various approaches. The annotation scheme of the Czech Learner Corpus was particularly relevant for our project, since Czech and Russian, both belonging to Slavic languages, share certain error types, e.g. inflectional and aspectual errors.

Furthermore, we examined error classifications created for Russian which describe common error types in the speech of Russian monolingual children (Tsejtlin, 1982; Rusakova, 2013), Russian first-generation emigrants (Zemskaya, 2001), Russian L2 learners and heritage speakers (Polinsky, 2006; Polinsky, 2010; Ovchinnikova and Pavlova, 2016). Having compared these classifications, we identified common errors, typical for all categories of Russian language learners, and error particularly frequent for heritage speakers and second language learners. Such error types were included in our error tagset. Furthermore, the error taxonomy was discussed with foreign language teachers of Russian and SLA researchers, collaborating with our project, and was additionally refined according to their suggestions.

The resulting error tagset consists of two tag classes: linguistic error classification and target modification taxonomy. According to Tono (2003), at least these two aspects should be included into error annotation scheme. The first group of tags defines an error in terms of linguistic types, e.g. derivational errors, agreement errors etc. Our classification includes broad categories corresponding to different levels of linguistic description, such as spelling, morphological, syntactic, lexical errors and errors in the use of constructions[6]

Each of these classes contains more specific error types. For instance, morphological errors comprise non-word errors, such as incorrect stem alternation, inflectional and derivational errors, as well as incorrect derivation of plural/singular for pluralia and singularia tantum nouns. We tried to avoid inclusion of infrequent error types in our tagset, in order to make it manageable for annotators. Therefore, the errors which do not correspond to more specific error types present in the tagset, are marked with a more general tag (e.g. "Morph" for morphological mistakes, "Syntax" for other syntactic errors etc.)

The target modification tags denote alternations of learner errors comparing to correct target element, such as deletion, insertion, substitution, transposition. These are used only in combination with the linguistic tags. Also, we included an additional tag marking cases of language transfer. As the influence of L1 can occur on different levels (spelling, morphology, syntax, lexical use), this tag should be combined with a linguistic error type, similar to target modification tags.

Along with the tagset, we needed to formulate error annotation principles in order to reduce subjectivity in annotation process and assure reliable inter-annotator agreement. The focus of error annotation in RLC is on severe spelling, grammatical and lexical errors which result in anomalous production. These errors should be corrected with minimal changes of the initial sentence, following the principle of the so-called first target hypothesis (Reznicek et al., 2013; Meurers, 2015). Hence, stylistic, dis-

[5]The tags contain information about parts of speech and all grammatical categories of the Russian language: gender, number, case, animacy, aspect, tense, mood, person, transitivity, voice, degree, full/short form.

[6]The term construction is broadly used within the framework of Construction Grammar (see (Fillmore et al., 1988; Goldberg, 1995; Goldberg, 2006; Tomasello, 2003; Ellis, 2013) and others). We understand constructions in a more neutral sense as lexical and grammatical patterns paired with particular meanings, cf. Russian possessive construction *u menya est'* (lit. 'at me is'), which is translated into English as 'I have'

course and pragmatic errors are not taken into account, since correction and annotation of such errors might require deeper interpretation of a learners utterance and this might lead to high variation in annotators decisions.

4.1 Inter-annotator agreement experiment

The error annotation is performed by students of linguistics and supervised by our team. Currently, the RLC annotation tool does not allow two annotators to work on the same texts without seeing the annotation decisions of each other. Therefore, we designed an offline inter-annotator agreement experiment in order to evaluate the consistency of annotation and to reveal ambiguous tags and/or inconsistencies in annotation guidelines.

The experiment was conducted on the sample consisted of 50 texts (8547 tokens in total) written by English and German L2 students. The annotation was made in files retrieved from the corpus. These have the following format: every word was presented on a separate line consisting of 6 columns — sentence number in the database, word, number of words in sentence, error tags, error correction, and annotator code.

Each text was annotated by two annotators (6 pairs in total). Before tagging each participant received 5 trial texts which were checked by supervisors. The most common mistakes were discussed with the annotators and outlined in the annotation guidelines. Afterwards the experimental sample was annotated, the tag mismatches were counted and Cohens kappa coefficient (Cohen, 1960) was calculated.

We assume that relatively low agreement (the highest score was obtained for syntactic (0.317) and spelling (0.249) errors, while the lowest coefficient of 0.185 was achieved for errors in constructions) was primarily caused by the lack of more detailed annotation guidelines. Although the current guidelines list the definition of all the tags and illustrate them with corresponding examples, difficult or ambiguous cases have not been outlined yet. Thus, the annotators made typical errors and did not distinguish between close error types, such as lexical errors and errors in constructions or spelling and inflectional errors. Moreover, since the experiment was performed outside the corpus platform, the annotators had to accommodate to a new data format and workflow, which might also serve as a source for inconsistent annotation. Therefore, an extensive annotation training might help to increase the inter-coder agreement score.

Having analyzed discrepancies in annotation, we decided to elaborate new annotation guidelines in order to improve the annotator agreement rate. We believe that this will lead to better results in the next session of our inter-annotator agreement experiments.

5 Corpus platform and tools

The corpus platform is a powerful and complex tool which enables various search options for researchers.

5.1 Development

The previous version[7] of the platform included only texts written by American learners of Russian; it also had no integrated annotation tool. Corpus users had only access to search interface and they could not upload their own texts or annotate them. The corpus workflow during that time was extremely time-consuming and ineffective. First, the contributors needed to send the texts to the corpus chief, who then sent these texts to annotators. The latter ran plain texts through morphological analyzer, which transformed them into XML files, and then annotated these XMLs using "Les Crocodiles"[8] annotation software which works only on Windows. In the next stage, the annotated texts were collected by the chief and sent to the database manager, who uploaded the texts to the corpus server and converted them into a special format, required to run the texts through the database indexator. As a result, we decided to automate the routine steps of the workflow and to enable the access to annotation for any OS (Windows, Mac, Linux).

The new platform is powered by Django, a web-framework written in Python programming language. The texts are kept in a MySQL database,

[7] The first version of the platform is available at `http://web-corpora.net/RussianLearnerCorpus/search/`

[8] The tool "Les Crocodiles" 2.7. was developed by Timofey Arkhangelsky.

which has dedicated tables for each of the text layers: metadata layer, sentence layer, morphological and error annotation layers. Such structure and the general Django framework allow to automate the majority of text processing. We also solved the OS problem by creating web-application for annotation, i.e. the annotation tool only needs a browser and Internet access and does not depend on the annotator's OS. In the following sections we describe the new features of the corpus.

5.2 Online data management

First of all, corpus users can upload new texts to the system and add metadata (see Figure 1), so each user can contribute their own collections of L2 or heritage texts to the project. When the text is uploaded, it is in the first place automatically processed by MyStem which includes sentence splitting and morphological analysis. Then the text is available for online annotation.

5.3 Annotation tool

The annotation tool is based on open-source JavaScript library Annotator.js[9]. The annotation is performed at three tiers. The first tier represents the original sentence: this is the tier where annotators mark errors. The second tier shows the original sentence with corrected spelling and morphological mistakes. The final corrected version of the sentence - with all syntactic and lexical changes that were added by annotators (see Figure 2) - is displayed by the third tier.

Some words in the first tier are underlined: this is done automatically when the corpus system detects a word which was analyzed by MyStem as "bastard". Such feature was added to help annotators find errors; this is based on the idea that the word which is not present in MyStem dictionaries or does not link to any template is likely to have an error.

To add a new tag, the annotator selects a fragment of text in the first tier and clicks the "Add annotation" button . After that a small dialog window appears, it has three fields: for error tags, for correction and for adding a comment if necessary. The selected fragment might be a word or several consecutive words within one sentence. It is possible to assign several tags to a single fragment if it contains multiple errors. The annotated spans might intersect: for example, one can annotate one word and then annotate a larger fragment including that word and several others. The comment section is meant to contain information about alternative target hypotheses (in case of competing target hypotheses) and possible sources of errors (e.g. examples of calques).

Each text in the corpus is classified into one of three groups: not annotated texts, annotated texts, and texts that were annotated and checked. The last group includes texts that were first annotated by corpus contributors or students of linguistic departments and later that annotation was reviewed by the corpus staff. Annotators and staff members can change the texts category by clicking corresponding buttons in the annotators workspace: "Mark as annotated" or "Mark as checked". The corpus staff aims at having all texts annotated and checked. As for now, almost 20,000 errors are annotated in about 35% of texts.

5.4 Corpus search

As in many corpora, one can execute search queries online. The corpus search engine allows to search texts for exact quotes or perform lexico-grammatical search: by lemma, part of speech and other grammatical features (like gender, number, voice, tense etc). These search queries can be also expanded with error tags. It is worth mentioning that the errors are searchable as soon as they are tagged in the annotation tool. For example, such queries can be executed in RLC:

- Find all code-switching errors tagged as CS;
- Find all examples of incorrect usage of passive voice;
- Find lemma *ja* – 'I, me' in dative or instrumental case tagged with any mistake;
- Find lemma *ja* – 'I, me' followed by a verb and/or a preposition.

Moreover, it is possible to define subcorpora: texts can be filtered by its mode (oral or written), native language of the author, gender, year of creation, language background of the author or level of proficiency in Russian.

[9]Official website of Annotator.js: `http://annotatorjs.org/`.

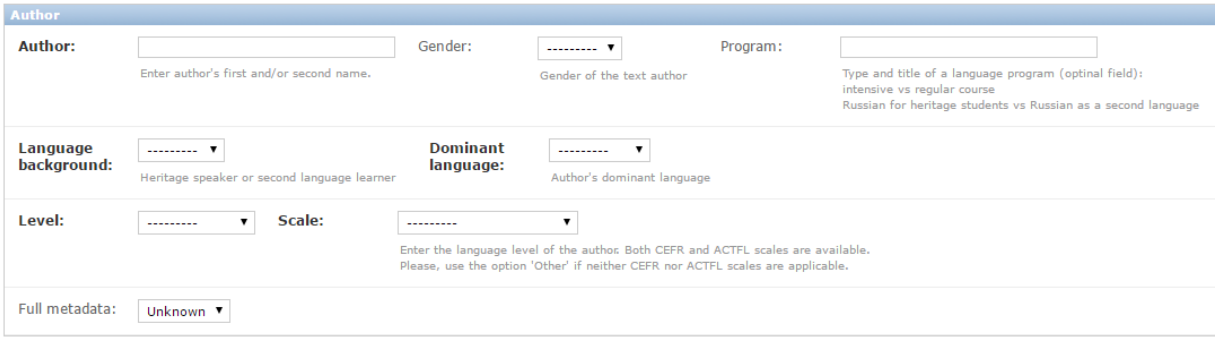

Figure 1: Data upload. Before uploading texts to the corpus, the user fills in metadata fields. The picture shows the form dedicated to author-related data.

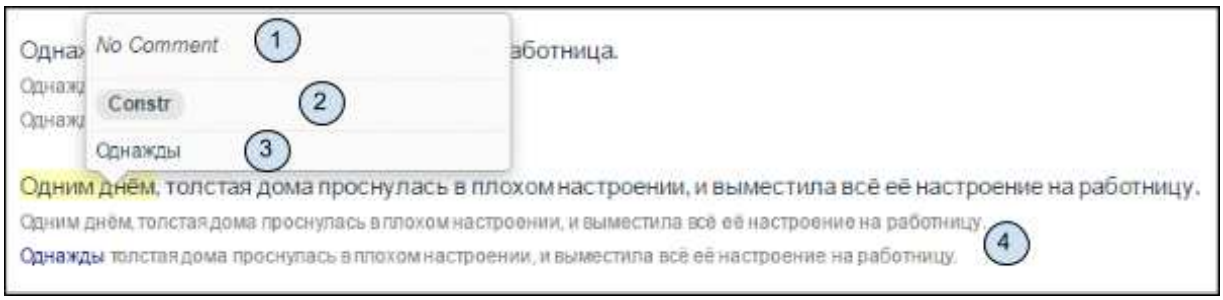

Figure 2: Annotators workplace. (1) Comment field. (2) Field for error tags. (3) Correction field. (4) The two layers of corrections are displayed under the original sentence. All the changes are highlighted.

Figure 3: Search results. (1) The text title, author's L1, and the sentence code. The code in only visible to authenticated users. (2) The expand context button. (3) The string matching the search query is shown in bold. The annotations are highlighted. (4) Search results show the original sentence and its final corrected version.

Search results include the number of sentences and number of texts that match the query and the list of all the sentences (Figure 3). Each entry in the results page contains the title of the text, the native language of the author, the original sentence with all allocated annotations and the corrected variant.

There is a possibility to view the context of the passage, but the context is restricted to the maximum of 3 sentences. Authorized users may have more access to the data. The corpus system was developed with user hierarchy, where each group of users has different permissions. Guest access gives permission only for searching the corpus, annotator access permits full text view and annotation. Annotators can also edit or add annotations directly in the search results page and also can view larger context for each result entry. Contributor access licenses not only annotation, but also adding new texts and editing their metadata. At the time of writing, RLC has around 100 users with different access permissions.

The corpus platform also creates a statistics page on the go: it is updated whenever anything is added to the corpus. It allows to see the whole perspective of available data at every moment.

6 Conclusion

In this paper we presented a pioneering Russian Learner Corpus which introduces a clear distinction between HL and L2. This resource has a unique platform with combined tools for corpus search and annotation.

The future development of the RLC is connected with the following tasks. First, we intend to annotate the remaining texts. In order to assure annotation quality, we are planning to improve the annotation guidelines and create an online tool for carrying out inter-annotator agreement experiments. Second, we will add more texts with different L1s and balance dominant languages in the corpus. Third, our team is going to improve the corpus search tool, for example, by including an option to save selected search results to the users directory.

Although not all texts have been annotated yet, the corpus still enables to retrieve interesting patterns in over- and under-using of certain constructions, some of them have been already described in linguistic research (Vyrenkova et al., 2014; Rakhilina, 2015; Polinsky et al., 2016). The annotated corpus data can also have numerous NLP applications, e.g. automatic error correction for language learners, automatic error tagging, author's native language identification. For example, the RLC data served as training and test data for tools for automatic error detection (Klyachko et al., 2013; Ramsajtseva et al., 2016). Therefore, we believe that the further corpus development will open new opportunities for SLA and heritage linguistics research, teaching Russian and creating tools for analyzing Russian learner interlanguage.

Acknowledgments

We would like to express our deep gratitude to everyone who has contributed to the project: Maria Polinsky (University of Maryland), Olessya Kisselev (Penn State University), Anna Alsufieva, Evgeny Dengub (Middlebury Langugage Schools), Irina Dubinina (Brandeis University), Anna Mikhaylova (University of Oregon), Alla Smyslova (Columbia University), Ekaterina Protassova (University of Helsinki), Anna Pavlova (University of Mainz), Anna Möhl (Johannes Gutenberg University of Zurich), Anka Bergmann (Humboldt University of Berlin), Irina Kor Chahine (Aix-Marseille University), Suhyoun Lee (Seoul National University), Svetlana Slavkova, Francesca Biagini and Monica Perotto (Bologna University), Svetlana Sokolova (Tromse University), Natalia Ringblom (University of Stockholm), Hayashida Rie (Osaka University), Margarita Kazakevich (Osaka Universty), Nazija Zhanpeisova (Aktubinsk University), Timofey Arkhangelsky, Olga Eremina, Ekaterina Shnittke and Evgenia Smolovskaya (Higher School of Economics), as well as BA and MA students ("Fundamental and Computational Linguistics" program at the Higher School of Economics) who at different years participated in annotation process.

We would like to thank the anonymous reviewers for their valuable and insightful comments.

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

N. C. Ellis, 2013. *Oxford Handbook of Construction Grammar*, chapter Second language acquisition, pages 365–378. Oxford University Press, Oxford.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language*, 64(3):501–538.

Adele Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

Adele Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Grammar*. Oxford University Press, Oxford.

Sylviane Granger. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. *Lund Studies in English*, 88:37–51.

Sylviane. Granger, 1998. *The computer learner corpus: a versatile new source of data for SLA research*, pages 191–202. Longman, London.

Ludmila Isurin and Tanya Ivanova-Sullivan. 2008. Lost in between: The case of Russian heritage speakers. *Heritage Language Journal*, 6(1):72–104.

S. Jarvis and M. Paquot, 2015. *Native language identification*. Cambridge University Press.

Elena Klyachko, Timofey Arkhangelskiy, Olesya Kisselev, and Ekaterina Rakhilina. 2013. Automatic error detection in Russian learner language. In *Proceedings of the First workshop Corpus Analysis with Noise in the Signal (CANS 2013)*, Lancaster, United Kingdom.

Andrey Kutuzov and Maria Kunilovskaya, 2014. *Russian Learner Translator Corpus*, pages 315–323. Springer International Publishing, Cham.

C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners: Second Edition*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Detmar Meurers, 2015. *Learner Corpora and Natural Language Processing*. Cambridge University Press.

Diane Nicholls. 2003. The Cambridge learner corpus - error coding and analysis for lexicography and ELT. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster University, UK.

I.G. Ovchinnikova and A.V. Pavlova. 2016. *Perevodcheskij bilingvizm. Po materialam oshibok pis'mennogo perevoda*. FLINTA: Nauka, Moscow.

A. Pavlenko, 2008. *Narrative analysis in the study of bi- and multilingualism*, pages 311–325. Blackwell, Oxford.

Maria Polinsky, Ekaterina Rakhilina, and Anastasia Vyrenkova. 2016. Linguistic creativity in heritage speakers. *Glossa*. In print.

Maria Polinsky. 2006. Incomplete acquisition: American Russian. *Journal of Slavic Linguistics*, pages 191–262.

Maria Polinsky. 2008. Heritage language narratives. *Heritage Language Education: A New Field Emerging*, pages 149–164.

Maria Polinsky. 2010. Russkij jazyk pervogo i vtorogo pokolenija emigrantov, zhivuschix v ssha. *Slavica Helsingiensia*, 40:336–352.

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

E.V. Rakhilina. 2015. Stepeni sravneniya v svete russkoj grammatiki oshibok. *Trudy instituta yazykoznaniya im. V.V. Vinogradova*, 6:310–333.

Olga Ramsajtseva, Aleksandr Ivankov, Robert Zakoyan, and Alina Ladygina. 2016. Morphchecker for non-standard data: a tool for morphological error correction in learner corpora. In print.

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01.

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann, 2013. *Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture*. Studies in Corpus Linguistics. John Benjamins Publishing Company.

M. Rusakova. 2013. *Elementy antropotsentrichnoj grammatiki russkogo yazyka*. Yazyki slavyanskikh kul'tur, Moscow.

Ilya Segalovich and Vitaly Titov. 1997. Mystem.

Barbora Štindlová, Svatava Škodová, Jirka Hana, and Alexandr Rosen. 2013. A learner corpus of Czech: current state and future directions. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Proceedings of , 15-17 September 2011*, Corpora and Language in Use, Louvain-la-Neuve. Presses Universitaires de Louvain. In print.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Harvard.

Y. Tono. 2003. Learner corpora: design, development and applications. In *Proceedings of the 2003 Corpus Linguistics Conference*, pages 800–809.

S.N. Tsejtlin. 1982. *Rechevye oshibki i ikh preduprezhdenie: posobie dlya uchitelej*. Prosveschenie, Moscow.

G. Valdés, 2000. *The teaching of heritage languages: an introduction for Slavic-teaching professionals*, pages 375–403. Slavica, Bloomington.

A.S. Vyrenkova, M.S. Polinsky, and E.V. Rakhilina. 2014. Grammatika oshibok i grammatika konstruktsij: heritage (unasledovannyj) russkij yazyk. *Voprosy yazykoznaniya*, 3:3–19.

E.A. Zemskaya, editor. 2001. *Yazyk russkogo zarubezha: Obschie protsessy i rechevye portrety*. Yazyki slavyanskoj kultury, Moscow.

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

7 Appendix A. Error tagset

Language level	Tag	Definition
Spelling errors	Graph	use of Latin alphabet
	Hyphen	error in use of hyphen
	Space	omission or insertion of space
	Translit	incorrect transliteration of a proper noun
	Ortho	incorrect letter
	Misspell	multiple severe misspellings (in one token)
Morphological errors	Infl	incorrect inflectional ending (which does not belong to a paradigm of a word)
	Deriv	made-up word
	Altern	error in stem alternation
	Num	non-existing number form (e.g. plural for singularia tantum)
	Gender	gender confusion
	Morph	other morphological errors
Syntactic errors	AgrCase	error in case agreement
	AgrGender	error in gender agreement
	AgrNum	error in number agreement
	AgrPers	error in person agreement (between subject and verb)
	AgrPers	incorrect subject for gerund
	Asp	error in verb aspect
	Passive	error in passive
	Tense	inappropriate tense form
	Mode	inappropriate use of verb mode
	Refl	incorrect use of a reflexive verb
	Gov	wrong case
	WO	word-order error
	Ref	pronominal reference error
	Conj	wrong conjunction
	Neg	error in negation
	Aux	incorrect use of auxilaries
	Brev	erroneous use of short-form adjective (or past passive participle)
	Syntax	other syntactic errors
Construction	Constr	Error in construction
Lexical errors	Lex	lexical error
	CS	code-switching
	Par	use of a paronym
	Idiom	error in idiom
Additional tags	Del	omission (of a character, a morpheme or a word)
	Insert	insertion (of a character, a morpheme or a word)
	Subst	substitution (of a character, a morpheme or a word)
	Transp	transposition (of a character, a morpheme or a word)
	Transfer	case of language transfer
	Not-clear	incomprehensible fragment