Russian Error-Annotated Learner English Corpus: a Tool for Computer-Assisted Language Learning

Elizaveta Kuzmenko, Andrey Kutuzov
National Research University Higher School of Economics
`lizaku77@gmail.com, akutuzov@hse.ru`

ABSTRACT

The paper describes the learner corpus composed of English essays written by native Russian speakers. REALEC (Russian Error-Annotated Learner English Corpus) is an error-annotated, available online corpus, now containing more than 200 thousand word tokens in almost 800 essays. It is one of the first Russian ESL corpora, dynamically developing and striving to improve both in size and in features offered to users. We describe our perspective on the corpus, data sources and tools used in compiling it. Elaborate self-made classification of learners' errors types is thoroughly described. The paper also presents a pilot experiment on creating test sets for particular learners' problems using corpus data.

KEYWORDS: learner corpora, English as a second language, computer-assisted language learning.

Elizaveta Kuzmenko and Andrey Kutuzov 2014. Russian error-annotated learner English corpus: a tool for Computer-Assisted Language Learning. *Proceedings of the third workshop on NLP for computer-assisted language learning.* NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107: 87–97.

1 Introduction

The present paper describes the learner corpus composed of English essays written by native Russian speakers, namely, students of National Research University Higher School of Economics. The corpus is error-annotated and available at `http://realec.org` (REALEC is for Russian Error-Annotated Learner English Corpus).

English learner corpora have been well documented in many countries [1], but in Russian linguistics not much attention has been paid to this research area. Therefore, there are not many corpora which comprise texts written by learners with Russian as the L1, Russian subcorpus of the ICLE being the best documented and representative of Russian-speaking learners' population (Granger et al., 2009). Another example of a corpus with Russian learners' data is RusLTC (Russian Learner Translator Corpus[2], which is composed of translations made by students. This corpus is of great help while studying translation process and mistakes typical for this kind of language task. At the same time, this corpus cannot be fully representative of Russian learners' language in general as translation is a very specific task. Also there is little or no sign of integrating corpus-based applications and techniques adapted for the needs of Russian-speaking learners into the study process.

Our research aims at developing a learner corpus with pure Russian-English interlanguage data available for use to anyone interested and at investigating how the corpus data can be used to facilitate the language learning. In particular, we describe the set of training exercises which are built on the basis of our corpus data and we compare their efficiency to the traditional exercises found in books for learners of English.

The structure of this paper is as follows: in section 2 we present our views of the tasks carried out by learner corpus research, describe the goals underlying the creation and usage of our corpus and give an overview of the corpus data. In section 3 we describe the annotation scheme adopted in our corpus. Section 4 presents external tools used for the creation of our corpus – the *brat* annotation tool used for marking up errors and the *Freeling* parser used for implementing part-of-speech tagging in our corpus. In section 5 we report our attempt to use corpus-based grammar exercises in the process of language teaching. Section 6 outlines the problems encountered during the process of setting up the corpus and discusses our future work.

2 Corpus overview (our perspective on LCR)

It is a common knowledge that having access to learner corpora is of great benefit for both learners and instructors – it has already been proved over more than twenty years of research in learner corpora (Granger et al., 2013). Learner corpora can be beneficial in several ways. First of all, learner corpus data are of tremendous help when studying a learner's interlanguage. Corpus linguistics methods have proved to be very successful when applied to the field of second language acquisition (SLA) as they help to observe patterns which are impossible to notice in the research of an individual subject (which is the traditional method in the field of SLA). Secondly, learner corpora offer major pedagogical help, much more than can be expected from purely descriptive SLA studies. This is why we pay so much attention to the aspect of error-annotation – students can improve their knowledge directly from corpus data.

Also, the number of teaching applications that can be built on the basis of a learner corpus is

[1] `http://www.uclouvain.be/en-cecl-lcworld.html`

[2] `http://rus-ltc.org`

beyond counting. Almost all learners' dictionaries are based on corpus data, *Longman Essential Activator* (LEA, 1997) being the first of this kind. Grammar books can be designed on the basis of learners' data too (Granger and Paquot, 2014). Such design is extremely helpful to learners, as they are shown concrete patterns in which mistakes often arise, not abstract rules that are to remember. Virtually, we define our main research purpose when building the Russian Error-Annotated Learner English Corpus as studying inconsistencies of students' interlanguage and getting insights about methods of language learning, which can later be embodied in the form of learning applications, such as corpus-based grammar and vocabulary exercises or recommendations on how to avoid typical learners' errors.

On the other hand, we consider the task of describing the Russian-English interlanguage in terms of CIA (Contrastive Interlanguage Analysis, as in (Granger et al., 1996)) secondary with respect to our aims. There are many studies carried out in the framework of CIA, whereas works on error analysis have been far less frequent for several reasons:

1. it is hard to strictly define what an error is,
2. error-annotation is a time-consuming task,
3. there is huge influence of a human factor, as errors are mostly identified by human annotators, who can have their own concepts of errors.

Thence, we hope to contribute to the field of error analysis by trying to overcome these difficulties and to formalize the basic concepts.

It should be noted that we acknowledge that the notion of error is rather vague and from the strict point of view it would be more correct to talk about 'variations of language'. The word '*error*' itself should, of course, be used with much caution (Kachru, 1992). However, from the practical point of view, learner corpora (including ours) are used in real teaching environment, where students are supposed to acquire a level of language skills enough to pass IELTS examination or other qualification tests. Good example is Russian Learner Translator Corpus, which among other tasks is employed to create exercises aimed at prevention of frequent translation mistakes (Kutuzov and Kunilovskaya, 2014).

So, from the point of view of IELTS etc., such 'variations' are outright mistakes. Thus, we believe that learner corpus annotation should include the notion of error, and it is practically useful to consider some language variations 'erroneous'.

Let us now describe the data presented in our corpus. As of September 2014, our corpus consists of 794 pieces of students' writing, which comprise more than 225 thousand word tokens, and the corpus is steadily growing.

The contributors to the corpus are 2, 3 and 4 year students (with Russian as their L1) from National Research University Higher School of Economics, Faculty of Philology, together with students of the first year of Master's program, Faculty of Psychology. The works presented in the corpus are either routine assignments or exam-type essays. Most of the works found in the corpus have the structure similar to that of IELTS writing tasks (Moore and Morton, 2005), as the main goal of English courses at the university is preparation for the compulsory IELTS examination.

The writing pieces in the corpus are initially processed with the help of a part-of-speech tagger (see section 4), and then error-annotated by experienced annotators (mostly teachers) using linguistically advanced error classification.

However, search options applied in our corpus that we have at the moment are not always sufficient for our needs. It is possible to search for a particular error tag or a string or substring containing part of a target word (using regular expressions). In future a morphologically enriched search tool, which allows searching a specified lemma, wordform or POS-tag alongside with error tags, will be developed.

3 Annotation scheme

We consider error-annotation the most important part of learner data analysis (Izumi and Isahara, 2004) . At the dawn of second language acquisition and learner corpus research it was already stated that errors made by a learner convey a lot of information about foreign language acquisition process (Corder, 1981).

However, the process of annotating errors is generally thought of as a very complex task, whose results cannot be truly reliable due to the human factor (Izumi et al., 2005). There were other points of criticism as well: ambiguous understanding of the term '*error*', neglecting structures that the learner generated correctly, etc. Despite this criticism, we believe that error annotation not only provides evidence concerning the process of acquiring language, but also can be of major pedagogical value, as stated in the previous section. Consequently, we have developed a detailed multi-layered annotation scheme to fully describe different aspects of errors.

The process of error annotation is carried out within *brat* text annotation framework, described in the next chapter. As of September 2014, more than 10 000 error spans are annotated in our corpus. Our annotation scheme comprises 4 tiers of information about a particular error. These tiers are:

1. type of grammar rule violated by the error,
2. supposed cause of the error,
3. the degree of grammar or 'linguistic' damage caused by the error,
4. the degree of pragmatic damage caused by the error (influence on understanding).

In addition, it is marked whether the correction of the error is insertion or deletion.

Let us observe every level of the annotation scheme in detail.

Primarily, the most important component of any annotation scheme is considered to be the classification of erroneously broken grammar, lexical and discourse rules. Approaches to this level of annotation differ by granularity. In our research we adopt a rather detailed annotation for grammar rule – it amounts to 151 types of grammar rules. This tier is divided into 5 categories with further subdivision (tree structure): punctuation, spelling, grammar, vocabulary and discourse. If an error affects more than one grammar rule, the mistake area can be tagged several times.

The figure 1 gives an idea of the ratio of different mistakes types across the corpus. Grammar and vocabulary mistakes dominate, accompanied by a serious amount of mistakes which annotators were not able to link to any type.

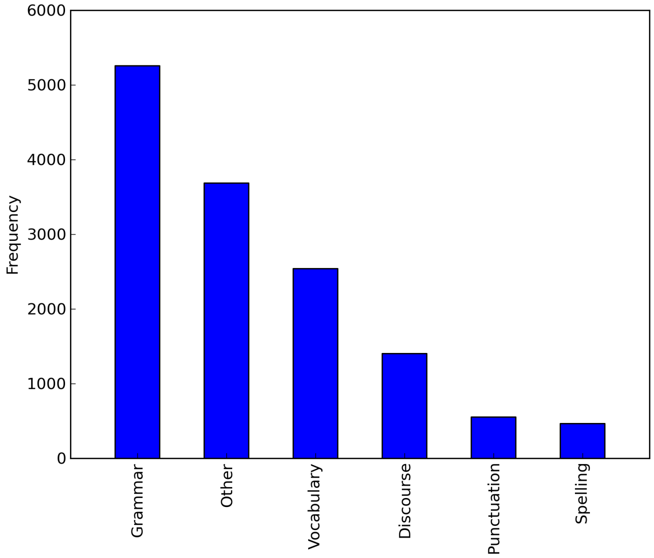

Figure 1: Mistakes types distribution over the whole corpus

The supposed causes of errors include typos and L1 interference (with a separate case of the absence of the category in L1). The evidence about typos can be of interest to psycholinguistic research, but in some cases it is not easy to distinguish typos from other errors. At the same time, the issue of transfer effects in the studied language deserve thorough consideration (Gas, 1979), so we are planning to refine this tier.

The next two levels of annotation – namely, grammar and pragmatic damage caused by the error – are by their nature scales with 3 values: 'minor damage', 'major damage' and 'critical damage'. Different types of errors are assigned a particular value of pragmatic and grammar damage according to our own classification of errors by this feature. These characteristics can be easily turned into integers and used for automated evaluation of students' work.

The unresolved problems of our annotation scheme are the following:

1. Lack of measurements for inter-rater and intra-rater agreement. Due to the lack of resources we can't have each essay annotated by several independent assessors and then compare their results, calculating inter-rater agreement. Neither did we check our assessors for intra-rater agreement by repeating annotation of the same text after some time, or by asking them to revise their own annotations. This knowledge can be of great use in assessing soundness of our error annotation.

2. Sophisticated classification by the violated rule, which reduces the possibility of direct pedagogical application as understanding tagging principles requires special training.

3. Inconsistent and subjective marking of degree of pragmatic damage (we have a more restricted principles of assigning degree of grammar damage).

Dealing with these problems will be our primary goal in the upcoming refinement of the annotation scheme.

Let us compare our scheme to other annotation schemes in use. Error categorization devised for use in the ICLE is in some ways similar to ours (Granger, 2003). It is based on linguistic error taxonomy developed in (Dulay et al., 1982) and focuses on grammatical categories being violated. The hierarchy of errors consists of 53 types divided into 8 major categories, which include standard lexical, grammatical, punctuational, syntactic errors and a number of intermediate subtypes. This scheme and ours are built according to the principle of tagging errors on the basis of the incorrect word/phrase, and not on the basis of the corrected word/phrase. Therefore, if an article stands in place for a personal pronoun as in the sentence *I saw the mother from a long distance*, this error belongs to the category *misuse of article* and not to *misuse of a possessive determiner*.

The differences between our annotation scheme and the one in the ICLE include:

1. greater detalization of error subtypes in our scheme,
2. variances in attributing particular error categories to major types (for example, errors on subordinating conjunctions are assigned to lexis in ICLE and to syntax in REALEC)
3. technical issues of juxtaposing markup and learners' texts (in ICLE markup is incorporated into the text and in REALEC texts and annotations are stored separately)

Now let us turn to another credible annotation scheme – the one deployed in Falko (Fehlerannotiertes Lernerkorpus, described in (Siemen et al., 2006)). This annotation does not seem to back on classical error taxonomies (Lüdeling et al., 2005). It is adjusted to the German language and reflects the grammatical categories which are frequently violated by the learners of German. The taxonomy itself comprises 8 types of errors: orthography, word formation, agreement, government, tense, mood, word order and expression. As well as in our corpus, annotation is carried out in the frame of the multi-layer standoff model, so that annotation is stored separately from texts. Also in Falko there is an annotation layer devoted to the description of error causes.

4 External tools

In compiling REALEC corpus we extensively used two external tools. The first is *brat*[3], open-source text annotation framework (Stenetorp et al., 2012). It allows simultaneous annotation of texts on server by multiple annotators at once using only their browsers, no matter their location. *Brat* also visualizes annotated texts in real time, making it easy to see and edit error spans and other markup.

It is very important that *brat* is highly customizable. One can adapt it to almost any type of markup (linguistic or not). Particularly, in our case it was easy to configure *brat* so that it offers our error classification as markup scheme.

One notable disadvantage of brat is its search interface which is not very rich. Although it is possible to look for particular entity (error type) in a document or in whole collection, one cannot combine several entity types in the query. Even more disappointing is the fact that *brat* does not support saving search results in any form. Due to open-source-nature of *brat*, this can be fixed, so we plan to do it.

[3] http://brat.nlplab.org/

Another external tool is *Freeling* suite of linguistic analyzers[4], also open-source. We employed only its lemmatizer and part-of-speech tagger module to assign POS and lemmas to word tokens in our corpus. With the help of our custom-made conversion tool, we managed to augment brat annotation files with data from *Freeling* output.

Freeling English tagger is trained on the widely known WSJ corpus, performs disambiguation and is reported to yield precision near 97-98% (Padró and Stanilovsky, 2012). It is hardly possible to somehow estimate Freeling performance on erroneous forms. Suppose we have a sentence *'The number are dropping.'* with the mistake in the verb form. *Freeling* marks *are* as a verb in plural. Context can't help disambiguation, because *are* simply does not have singular forms among its possible parsings. But we can hardly say that the tagger performs poorly here: it assigns the only possible POS tag to the token. It is suitable for us, as after that we theoretically can analyze annotated texts and look for 'strange' part-of-speech sequences.

Another case is wrong spelling. *Freeling* does not perform spell-checking before analyzing, thus it processes spelling errors as unknown words, and this is where the context comes into play. In most cases *Freeling* tagger is able to correctly assign a PoS tag to a wrongly spelled word based on word suffix and its neighbours. For example, in the sentence *'However, certain reseches disagree.'* the word **reseches* (*researchers*) is assigned a noun tag, obviously because it is preceded by an adjective and followed by a verb. To sum it up, we have not met many additional problems related to the fact that the authors of our texts are learners, not native English speakers.

Thus, we have if fact two layers of annotation: errors and POS/lemma. This provides corpus user with the possibility to search for particular parts of speech and to find all grammatical forms of query word. However, as stated before, *brat* does not allow double query constraints (for example, one can't search for verbs with word choice errors). Also, our *freeling2brat* converter is not perfect and in some cases lemmas and morphological tags are assigned to wrong words. Mostly this is because *Freeling* deals with sentences and words, while *brat* operates with character offsets from the beginning of the document. Fixing the converter is one of directions for our future work.

5 Applications for language teaching

As it was pointed out earlier, various applications for language teaching can be built using the corpus data. There is no clear evidence on which learning stage corpus-based tools should be used. On the one hand, intermediate or advanced learners eager to eliminate the traces of their non-nativeness might want to know subtle problematic cases among learners in general. On the other hand, beginners might also be interested in error patterns in order to avoid them from the very beginning of their study. It can be noted, however, that most corpus-based tools are designed for self-tuition, making them more appropriate for advanced learners.

In particular, learner corpus data are used to present mistake patterns to students and to explain them the nature of their mistakes, as in (Altenberg and Granger, 2001). Also, error-tagged data can be used to identify which erroneous patterns are frequently encountered in learners' writing and to correct one's teaching methodology correspondingly, as described in (Seidlhofer, 2002).

Now we are going to describe the corpus-based exercises and give an experimental estimation of their efficiency. We believe that this can be regarded as a preliminary attempt to apply the results of our corpus work to practical uses in the learning process.

[4] `http://nlp.lsi.upc.edu/freeling/`

We hypothesized that students acquire grammar rules better if supported by corpus-based, focused on problematic cases exercises, whereas traditional grammar drills lose their efficiency compared to corpus-based exercises. To support this insight, we performed a pilot experiment described below.

The grammar rule chosen for testing in this experiment was using commas in defining and non-defining relative clauses.

The design of the experiment was as follows: students were assigned an essay, and all essays were uploaded in the corpus and error-annotated. After that the students were divided in two groups, 20 people each. The members of the first group completed exercises built on erroneous sentence instances from the corpus. Another group of students was studying the rule with the traditional material from grammar books ((McCarter and Roberts, 2010) and (Cotton et al., 2008)). After practicing the rule, students were once again assigned an essay, and the second portion of essays was error-annotated too.

The measure used for statistical analysis was errors per word ratio. The analysis has shown that the ratio has decreased in the data of experimental group and, on the contrary, increased in the control group (cf. Table 1).

Group	Error/word ratio before	Error/word ratio after	t-test confidence
Experimental	0.005122	0.001605	0.046853
Control	0.002785	0.004274	0.082329

Table 1: Error per word ratio

The obtained data show that so far corpus-based exercises proved to be influencing learners' performance in a positive way. If supported by future experiments, it can be confirmed that corpus-based exercises suit the study process better as they focus on problematic for learners' issues. Our future plans concerning this experimental schema are to design exercises for various rules and involve the measures of grammar and pragmatic damages into calculations.

6 Future work

As there are not many learner corpora for Russian-speaking learners of English and even fewer freely available corpora, our corpus seems to be a valuable contribution into the field. We plan to maintain the corpus, expand its volume and augment it with extensive error annotation cross-validated by numerous annotators. An important aspect of our refinement is developing a uniform way of tagging errors and verifying the details of the instructions for annotators to make sure that similar types of errors are tagged with similar tags. We also need to specify the tagging principles for the following tiers of annotation: causes of an error and the pragmatic damage caused by an error.

We are working within *brat* web annotation framework, so we are heavily dependent on its capacities. We are going to improve the framework by:

1. developing robust library to convert Freeling output to brat annotation files,
2. adding automatic tagging of typical spelling and grammar/syntactic errors,
3. improving search tool (possibility to save search results, multiple constraints, etc.),

4. adding the option to show only chosen tags (invisible annotation layers),
5. providing the possibility to annotate empty spans (missing articles, for example)
6. improving access control abilities (ability to show or hide annotations depending on user role).

Also we plan to use our corpus for the following applied tasks:

1. automated error detection,
2. automated grading of students' work (based on complex metrics and taking into account the predefined structure of exam essays),
3. massive creation of corpus-based error-focused grammar exercises,
4. adjusting morphological and syntactic parsers to learner corpus data.

Another important aspect of our future work is research into the process of second language acquisition based on the corpus data. In particular, we are planning to investigate the following aspects:

1. error patterns specific for Russian-speaking learners of English,
2. types error patterns typical of different stages in the learning process,
3. common causes of errors and peculiarities of L1 transfer effects for native Russian speakers,
4. introduction of the procedures of self-editing and mutual annotation among students,
5. features of interlanguages of learners with different past exposure to English (we have correspondent metadata available to large part of texts in the corpus).

Finally, the corpus can be of great value to the research into didactic issues, such as formulating the concept of an error in general, elaborating criteria for assigning errors, or studying the rate of agreement between EFL instructors.

7 Acknowledgements

We would like to express our deep gratitude to Olga Vinogradova for inspiring this project and for guiding us through the jungle of learners' errors.

The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2014.

References

Altenberg, B. and Granger, S. (2001). The grammatical and lexical patterning of make in native and non-native student writing. *Applied linguistics*, 22(2):173–195.

Corder, S. P. (1981). *Error analysis and interlanguage*, volume 112. Oxford Univ Press.

Cotton, D., Falvey, D., Kent, S., Albery, D., Kempton, G., and Hughes, J. (2008). *Language Leader: Upper Intermediate*. Pearson Education.

Dulay, H., Burt, M., and Krashen, S. D. (1982). *Language two*, volume 2. Oxford University Press New York.

Gas, S. (1979). Language transfer and universal grammatical relations. *Language learning*, 29(2):327–344.

Granger, S. (2003). Error-tagged learner corpora and call: A promising synergy. *CALICO journal*, 20(3):465–480.

Granger, S., Dagneaux, E., Meunier, F., Paquot, M., et al. (2009). The international corpus of learner english. version 2. handbook and cd-rom.

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

Granger, S., Gilquin, G., and Meunier, F. (2013). *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, volume 1. Presses universitaires de Louvain.

Granger, S. and Paquot, M. (2014). The louvain eap dictionary (lead): A tailor-made web-based tool for non-native academic writers of english.

Izumi, E. and Isahara, H. (2004). Investigation into language learners' acquisition order based on the error analysis of the learner corpus. In *Proceedings of Pacific-Asia Conference on Language, Information and Computation (PACLIC) 18 Satellite Workshop on E-Learning, Japan.(in printing)*.

Izumi, E., Uchimoto, K., and Isahara, H. (2005). Error annotation for corpus of japanese learner english. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora*, pages 71–80.

Kachru, B. B. (1992). *The other tongue: English across cultures*. University of Illinois Press.

Kutuzov, A. and Kunilovskaya, M. (2014). Russian learner translator corpus. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 8655 of *Lecture Notes in Computer Science*, pages 315–323. Springer International Publishing.

Lüdeling, A., Walter, M., Kroymann, E., and Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005*, pages 15–17.

McCarter, S. and Roberts, R. (2010). *Ready for IELTS Coursebook*. Macmillan Education.

Moore, T. and Morton, J. (2005). Dimensions of difference: a comparison of university writing and ielts writing. *Journal of English for Academic Purposes*, 4(1):43–66.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. *Computer learner corpora, second language acquisition and foreign language teaching*, pages 213–34.

Siemen, P., Lüdeling, A., and Müller, F. H. (2006). Falko – ein fehlerannotiertes lernerkorpus des deutschen. *Proceedings of Konvens 2006*.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for nlp-assisted text annotation. In *EACL*, pages 102–107.