Review of Practices of Collecting and Annotating Texts in the Learner Corpus REALEC

Olga Vinogradova[1] and Olga Lyashevskaya[1,2(✉)]

[1] The National Research University Higher School of Economics, Myasnitskaya ulitsa 20, Moscow 101000, Russia
olgavinogr@gmail.com, olesar@yandex.ru
[2] Vinogradov Russian Language Institute RAS, Volkhonka Street 18/2, Moscow 119019, Russia

Abstract. REALEC, learner corpus released in the open access, had received 6,054 essays written in English by HSE undergraduate students in their English university-level examination by the year 2020. This paper reports on the data collection and manual annotation approaches for the texts of 2014–2019 and discusses the computer tools available for working with the corpus. This provides the basis for the ongoing development of automated annotation for the new portions of learner texts in the corpus. The observations in the first part were made on the reliability of the total of 134,608 error tags manually annotated across the texts in the corpus. Some examples are given in the paper to emphasize the role of the interference with learners' L1 (Russian), one more direction of the future corpus research. A number of studies carried out by the research team working on the basis of the REALEC data are listed as examples of the research potential that the corpus has been providing.

Keywords: learner academic writing in english · learner corpus · L1 Russian · corpus annotation · error taxonomy

1 Introduction

Researchers over the last four decades have claimed that learner corpora provide evidence necessary for second language acquisition theory and practices, as well as for many areas of linguistic studies (see [9–11,14]; and the important reviews by G. Gilquin [7] and by T. McEnery with co-authors [18], among many others). Learner texts themselves make up a valuable resource, and their value grows manyfold if the texts get annotation of features specific for a particular corpus. Russian Error-Annotated Learner English Corpus (REALEC), set up at HSE

The research was carried out within the project of the HSE University Research Foundation 2021 - Automated analysis of text written in English by learners with Russian L1 (ADWISER).

P. Sojka et al. (Eds.): TSD 2022, LNAI 13502, pp. 77–88, 2022.
https://doi.org/10.1007/978-3-031-16270-1_7

University, is a collection of essays written by 2nd- or 3rd-year university learners of English with Russian as their native language. REALEC is in the open access at the university portal. The errors in the texts have been manually annotated in the years 2014–2020, and Sect. 2 gives the details about the collection and annotation approaches adopted in REALEC.

2 Learner Corpora Available for Research Purposes

A number of large learner corpora have been presented to the research communities, and the results of using their data have been reported in numerous publications. The collection of smaller and larger learner corpora with different L1 of the contributors can be found on the site of the Learner Corpus Association [29]. The most frequently referenced corpora in corpus research community seem to be EFCAMDAT ([4,6] - 1st version and [13] - 2nd version) - accessible to the public big collection of short learner texts from learners with different levels of proficiency; ICLE [12] with 5.5 million words of essays written by learners with 25 different native languages; Cambridge Learner Corpus, CLC [21], a 45-million word corpus of student responses to ESOL exams, which can be accessed in Sketch Engine in two main parts - the error-coded learner corpus (CLC coded) and the uncoded learner corpus (CLC uncoded); and also two corpora of spoken learner production: the Louvain International Database of Spoken English Interlanguage, LINDSEI [8], and the Trinity Lancaster Corpus, TLC [5], with 4.2 million words of transcribed L2 spoken interaction. All these corpora differ in size, in the number of native languages of the learner authors, in platforms they were released on, and most importantly for this paper, in availability of different types of annotation assigned to the learner texts. These corpora have already been successfully used for studying a broad range of lexical, grammatical and pragmatic features. As all of them have been well documented, we tried to adhere to the same level of detail and pointed out the same important features in our presentation of the Russian Error-Annotated Learner English Corpus.

3 Data Collection and Annotation Practices in REALEC

All Bachelor students at the HSE University take the Independent English Language Test (IELT) designed to evaluate English proficiency in academic register of English [28]. The exam format is the same as that of the leading international English certification tests, with IELTS being the closest. The examination is called Independent because EFL instructors from HSE do not participate in organizing this test or evaluating students' work in it. This task is done by independent certified examiners, who develop the materials every year and assess students' written and oral performance (essays and interviews, respectively) in accordance with international language standards. The test includes Reading, Listening, Writing and Speaking, and it is essays written in answer to the two tasks in Writing - a description of the graphical materials in 20 min and an opinion essay in 40 min, which have been submitted to REALEC since the year

2014. All students taking the IELT are at a similar academic level, as they are all undergraduate students (2nd or 3rd year at HSE), but because of differences in prior exposure to English language, examination essays show a wide range of levels, and our pilot experiments on automated predictions of CEFR levels attested CEFR levels from B1 to C1 for the majority of essays [1].

The collection in REALEC of these essays from the years 2014–2020 includes about 18,700 texts, with the total of approximately 4,336,000 words. When the administration of the examination involved only students of three departments typing essays on computer (in 2014–2019), we were able to annotate errors in those essays manually. This work was done by specially trained student annotators as their practical experience in corpus maintenance, and unfortunately we never had enough of those annotators to follow the conventional practice of double annotation of all texts. However, we did have an editing team responsible for editing student annotation to ensure some consistency in annotating approach. The new technological breakthrough came in 2020, when the test was administered online for students of all departments of the HSE university, and as a result REALEC received twice the number of texts as that in all the previous 5 years.

REALEC is made up of (1) the texts with the sentence borders established by using NLTK Punkt sentence tokenizer [2], (2) automated POS annotation tags received with the help of TreeTagger [22], and (3) of the files with manually annotated errors in the form of error spans, error tags assigned to them, and the correction of the error span suggested by the annotator. The learner corpus was released on BRAT platform [23] chosen for its convenience for annotating processes and for the highly satisfactory visualization opportunities. A team of specially trained Linguistics undergraduate students proficient in English annotated about 6,000 essays between 2014 and 2019. The annotators chose an appropriate label for each error they identified, and they could apply more than one error tag to the same error span if needed (see Figs. 1 and 2 for examples). The results of inter-annotator agreement experiment carried out across 2,128 errors annotated by 5 independent annotators were presented in (Vinogradova, 2016:743-748). Figures 1 and 2 illustrate REALEC annotations with two and four error tags assigned to one error span, and a pop-up window on the screenshots presents the corrected version suggested by the annotator.

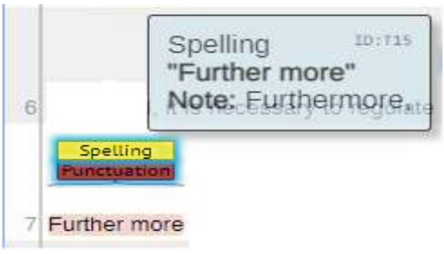

Fig. 1. Multiple categorisation of errors in REALEC - error span with 2 error tags.

There are 5,604 error spans with more than one error tag, which makes up 4% of the total number of error spans.

While choosing the appropriate tag for an error identified, annotators had to tick `delete` if an error span was to be deleted instead of being corrected. There was also an option to choose the tag `L1 Interference` as a possible cause of error, but so far it has not been marked consistently enough. An example of such error is the confusion in Russian learners' use of English verbs solve and decide Fig. 2: both English verbs have the same equivalent in Russian, so this wrong vocabulary choice (`Choice of lexical item` tag) was supposedly made under the influence of L1 interference.

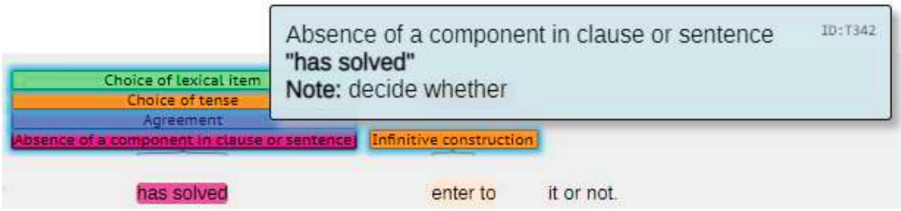

Fig. 2. Multiple categorisation of errors in REALEC - error span with 4 error tags.

One more function in annotation was to show that some changes had to be applied as a result of some other changes already made in the sentence, and for such cases there is a way to show with an arrow the relation between two tags called `Dependent change` coming from the initial suggestion of a change to the other tag depending on the former - see the example in Fig. 3.

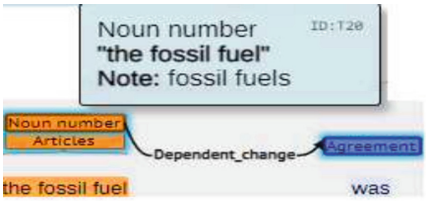

Fig. 3. Dependent change relation in REALEC.

After the annotators have completed their work, the supervisor of the annotation practice does some spot-checks, after which some decisions get reviewed. Annotation reviewing is an on-going process, and the specific numbers (in particular, numbers of texts without any annotations - see Table 1) and some of the choices made by annotators are still subject to changes. Currently, the total is 134,608 error tags for 4,918 out of 6,054 texts collected in the corpus from 2014 to 2019.

There have been observed cases when POS automated annotation from Treetagger produced some misleading indices - like in the following cases:

(1) *All of us have their leisure time and there is no secret that a lot of us* ***like*** *some kind of sport activity.*
(2) *For example, my friend Andrew really* ***like*** *basketball.* Both verbs *like*, the correct form in (1) and the incorrect form in sentence (2), are marked with the tag PRP (preposition) instead of the necessary verbal tag.
(3) *It can be one of the main reasons why the mobiles phone's part* ***increase*** *for 2 times.* Word *increase* is marked with the tag NN1 (singular noun) instead of the necessary verbal tag.

While the first confusion does not stem from any error in the learner text, the second and the third ones can be accounted for by agreement errors made by student authors.

4 Corpus Statistics

Table 1 gives the basic statistics of the REALEC corpus collected in 2014–2019. The table gives numbers of texts, the total numbers of sentences, words and tokens, the average numbers of sentences, words, and tokens per text, the maximum number of words in a text, the total number of error tags assigned, the average number of errors per text, the average numbers of tokens and error tags per sentence, and the total number of annotated and unannotated texts for Task 1 and Task 2 essays separately, as well as separately for the years 2014-2017 and 2019. For the much greater number of texts collected in 2020, manual annotation was out of the question, so we applied a BERT-transformer-type neural network for both identification and correction of errors, and the analysis of the results is still in progress and will not be included in the current report.

Table 1. Corpus statistics for the texts collected in REALEC before 2020.

Year – Task	Texts	Sent	Words	Tokens	Av. Snt /Txt	Av. Wrd /Txt	Av. Tok /Txt	Total Error Tags	Av. Err/ Txt	Av. Tok/ Snt	Av. Err/ Snt	Texts with annot
14 – 1	829	7,757	147,953	166,906	9	178	201	17,284	26	22	3	668
– 2	823	12,223	219,740	246,325	15	267	299	22,119	33	20	3	678
15 – 1	31	5,045	5,680	8	163	183	224	621	22	23	3	28
– 2	30	401	7,709	8,612	13	257	287	981	36	21	3	27
16 – 1	670	5,902	123,522	136,130	9	184	203	9,498	18	23	3	522
– 2	664	9,603	181,135	201,406	14	273	303	11,960	23	21	2	512
17 – 1	1,126	10,467	196,103	222,619	9	174	198	23,155	25	21	3	929
– 2	1,124	16,816	315,001	351,628	15	280	313	29,227	35	21	3	839
19 – 1	377	3,293	70,665	79,449	9	187	211	8,242	23	24	3	354
– 2	380	5,708	118,605	131,898	15	312	347	11,521	32	23	3	361
Task1	3,033	27,663	543,288	610,784	9	179	201	58,800	24	22	3	2,501
Task2	3,021	44,751	842,190	939,869	15	279	311	75,808	31	21	3	2,417
Total	6,054	72,414	1,385,478	1,550,653	12	229	256	134,608	27	21	3	4,918

It can be seen that in roughly the same numbers of Task 1 and Task 2 essays, the numbers of sentences, words and tokens, both total and average/maximum, correspond to the proportion of the required length: Task 1 essay is supposed to be not less than 150 words, while Task 2 essays are required to be not less than 250 words. The only parameter with much smaller, almost no, difference between Task1 and Task 2 texts is the average number of tokens per sentence, which makes sense because each student wrote both Task 1 and Task 2 essays. The statistics related to error counts is discussed in the next section.

5 Error Taxonomy in REALEC

Hierarchical error categorization for the corpus was initially developed on the basis of the pedagogical tradition in Russian EFL error-marking practices and included over 150 error tags [25]. After about two years of manual expert annotation with this categorization scheme and as a result of annotator-agreement experiment, the number of tags was reduced to about 100 on the grounds of infrequent use of about a third of them. Further application of this reduced scheme revealed the inconsistency and/or the need for high-level linguistic knowledge for the appropriate use of some more of the error tags, so another portion of about 50 tags was eliminated.

The new version of the error tags has 54 errors tags (see Table 2), of which 7 are upper-level tags (Grammar, Vocabulary, Verbs, Nouns, etc.) used only for grouping tags of similar nature. At each stage of applying changes to the error tags, most error spans that had been annotated with the eliminated tags were automatically reassigned the remaining error tags, but there were six that required manual updating. One example of these six is the tag `Conditionals`: annotators used this tag either for wrong tense forms, which in all other types of clauses were labelled with the tag `Choice of tense`, or for the wrong uses of the negative form (for example, with the conjunction unless), which in turn can be marked with `Negation` tag. The reassigning of tags had to be implemented manually.

6 Distribution of Learner Errors in REALEC

From the general statistics, we can see that errors are quite frequent in student essays collected in REALEC - error density parameter is 9.72 errors per 100 words. Roughly, every tenth word in the corpus is grammatically incorrect. This shows that not many student authors of the essays in the corpus have achieved a very high level of English proficiency, which in terms of CEFR, which in terms of CEFR level implies somewhere between levels B1 and B2 for the majority of student authors.

When we look at the distribution of errors across documents, we can make some interesting observation. Figure 4 shows the histograms of the number of error annotations per document for Task 1 essays and for Task 2 essays in different colour (blue and orange, correspondingly). The distribution for both classes

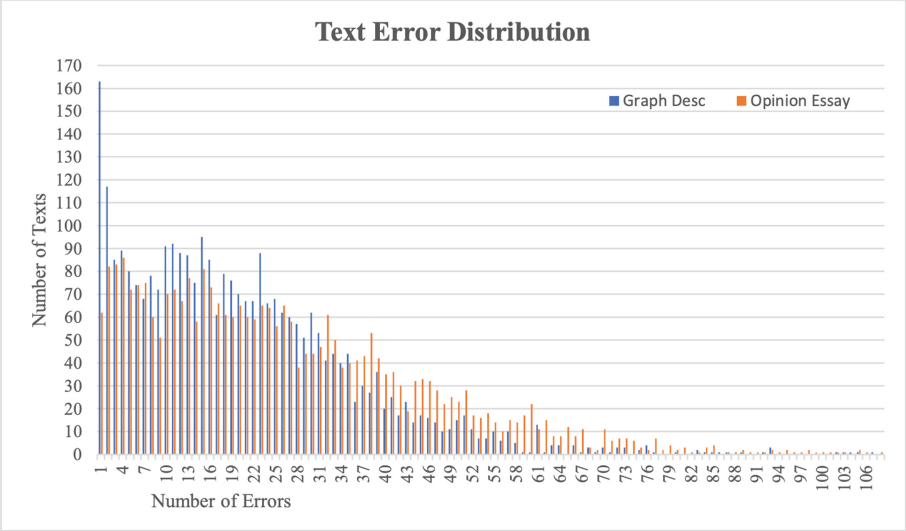

Fig. 4. Distribution of error annotations across documents in REALEC.

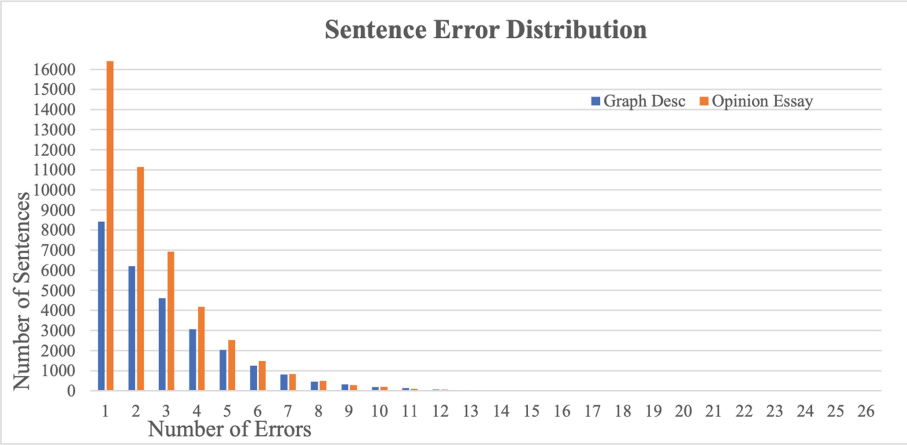

Fig. 5. Distribution of error annotations across sentences in REALEC.

Table 2. Error categorisation in REALEC.

Upper-level tag	Error tag	Error spans and ≫ their corrections for some tags
	Punctuation	
	Spelling	
	Capitalisation	
Grammar	Determiners	**The other** ≫ **Another** example
	Articles	**a** lowest figure ≫ **the** lowest figure of 35%
	Quantifiers	**much** ≫ **many** efforts
Verbs	Tense	
	Choice of tense	There **is** ≫ **was** a rise in 2012
	Tense form	It **has taken** many years
	Voice	**was fluctuated** ≫ **fluctuated**;
		interested ≫ **interesting fact**
	Modals	**must** ≫ **had to** do
	Verb pattern	**let them to create** ≫ **let them create**;
		Please **introduce** ≫ **introduce yourself**
	Gerund or participle construction	**Create** ≫ **Creating** modern house is
	Infinitive construction	**Doing** ≫ **To do** it means to develop
Nouns	Countable/uncountable nouns	**advices** ≫ **advice**
	Prepositional noun	... a **reason of** ≫ **reason for**
	Possessive form of noun	This **student** ≫ **student's** reaction
	Noun+infinitive	the way **of solving** ≫ **to solve**
	Noun number	I know **case** ≫ **cases** of injustice
	Prepositions	**in** ≫ **at** night
	Conjunctions	new opportunities appear, ≫ **and** the whole world becomes
Adjectives	Prepositional adjective	**independent on** ≫ **independent of**
	Adjective as collective noun	**poors** ≫ **the poor**
Adverbs	Prepositional adverb	**independently out of** ≫ **independently of**
	Degree of comparison	the **best** ≫ **better** of the two
	Numerals	**three millions of** ≫ **three million** people
	Pronouns	
	Agreement	
	Word order	
	Relative clauses	offices **which** ≫ **whose** role is
	Parallel construction	They want to study ... and **doing** ≫ **do** sports.
	Negation	They **have not** ≫ **do not have** time to do it
	Comparative construction	twice **more** ≫ **as many**
	Confusion of structures	**There is** ≫ **It is** very important to
Vocabulary	Word choice	**places in work industry** ≫ **work places**
	Choice of lexical item	**make** ≫ **fulfil** its function
	Change, deletion, or addition of part of lexical item	**the jury is still on** ≫ **the jury is still out on**
	Derivation	
	Formational affixes	**controversional** ≫ **controversial** issue
	Confusion of category	I **am agree** ≫ **agree**
	Compound word	**crowd sourcing** ≫ **crowdsourcing**
Discourse	Referential device	higher than **of** ≫ than **that of** male graduates
	Coherence	
	Linking device	**To sum**, ≫ **To sum up**,
	Inappropriate register	**tiny** ≫ **insignificant** increase
	Absence of a necessary component in clause or sentence	while **appeared** ≫ **there appeared** more people
	Redundant component in clause or sentence	from **both opposite** ≫ **both sides**
	Absence of necessary explanation or detail	The percentage of **people** ≫ **people** **in this group** is about 70%.

of essays is heavily skewed to the left with most documents (4617 out of 6054) having less than 32 errors, while some documents have significantly more errors than the average document: 16 graph descriptions and 34 opinion essays have more than 80 error annotations, and the highest number of error annotations in a document overall is 133. The mode (the most frequent value in the histogram) is 1 error/text for graph descriptions and 5 errors/text for opinion essays, and the median is 16 errors for Task1 and 21 errors for Task2.

A similar pattern can be observed when we look at the distribution of errors per sentence. Figure 5 shows a histogram of the number of error annotations per sentence in the REALEC corpus. The histogram shows that the largest number of sentences have no errors or one error, both in Task 1 and Task 2 essays. The frequency decreases quickly for higher error counts, and the highest observed number of error annotations in a sentence is 34 in Task 1 and 25 in Task 2 essays.

The skewed distribution of errors in the corpus was observed in (Dahlmeier et al. 2013), in which the authors explicated the long tail of the distribution by stating that if a learner has made a lot of mistakes in the beginning of the essay, the chance of making more errors in the remainder of the essay increases at least because of systematic errors, which are likely to be repeated.

As for the types of errors that language learners make, Fig. 6 shows a histogram of most frequent error categories (having 1,000 occurrences and more). The top three categories are misspelled words (about 27 thousand, 23%), wrong uses of articles (about 15 thousand, 13%), and wrong vocabulary choices (about 17 thousand, 15%). These top three error categories account for 51% of all error annotations. The next 5 categories are errors in punctuation, wrong choice of verb tense, inappropriate prepositions, agreement errors, and uses of redundant components, ranging from 9 to 4%.

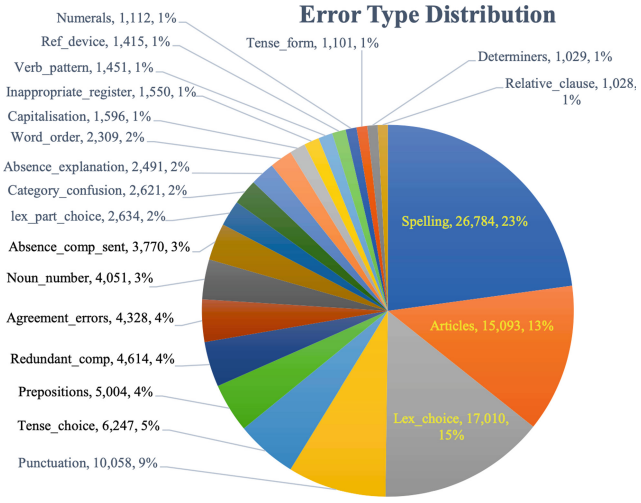

Fig. 6. Distribution of 22 most frequent REALEC error tags.

7 Corpus at Work

Drawing from the observations over the most frequent errors made by student authors with Russian L1, the research team working with REALEC set up the task to create, and reported the initial description of, a writing assistant for learners with Russian as L1 [26]. Annotated errors also formed the basis for a test-making program which worked at the HSE University as a placement program for a few years [26]. The third computer tool developed by the research team was a system for evaluating text complexity parameters [15].

Currently, two more directions for the researchers working with REALEC is to explore the relations between text complexity values and error counts in learner texts, on the one hand, and comparisons of English learner production by learners with different L1 [17,27]. The work on the writing assistant is still in progress, and an interesting question is how adding syntactic parsing with SpaCy [31] allowed us to increase the efficiency of the writing assistant in identifying four of the eight most frequent errors attested in REALEC, namely:

- errors in the subject-predicate agreement
- errors in the determiner-noun agreement
- errors in the use of commas
- errors in the use of verb tenses
- errors in the use of some prepositional constructions.

Some limitations in the ability of the parser to cope with the erroneous learner production, especially when errors were made by Russian learners of English under the influence of L1, have been observed by the REALEC research team. The possible ways to tackle the problems with wrong parsing was discussed in [17], but this specific line of research is beyond the scope of the present paper.

The difficulties of carrying out research across different learner corpora were noted by many authors (see, for example, [24, :44–45], and one of those were specific errors made under the influence of the interference with learners' L1. That is why annotation in REALEC has an additional focus on marking all errorenous occurrences that in some way resemble the features that exist in Russian. The examples can be brought in from such different areas as spelling (*democraty* instead of democracy - cf. Russian *demokratia*; *standarts* instead of *standards* - cf. Russian *standart*), word formation (*tendention* - cf. Russian *tendentsia*; *expluatated* instead of *exploited* - cf. Russian *ekspluatiroval*, lexical choice (see example in Fig. 3; *close their eyes to* instead of *turn a blind eye to* - cf. Russian *zakryvat' glaza na*). Errors in word order often have to be given as the complete sentences, and the corresponding Russian sentences have exactly the same word order as in the erroneous English sentence, as in the following example: *What it leads to?* instead of *What does it lead to?*

8 Conclusions and Future Research

The paper has reviewed our recent work towards development the REALEC corpus. Texts from our corpus can be downloaded, and the fact that the time-consuming and costly error annotation has been done and is being improved will

hopefully make REALEC a valuable resource for EFL professionals, for SLA researchers, for linguists working in different walks in Linguistics, for NLP specialists and, finally, for students learning to become EFL instructors to practice error detection and correction in English classes. At HSE University, REALEC data are being used by both undergraduate and graduate students in Computer Linguistics program for their research activities.

Now that the first results of automated identification and correction in a large portion of learner texts have been received, it becomes even more important to increase the consistency of manual error annotation in the smaller part of the corpus in order to be able to create a procedure of the automated error classification as a follow-up to deep learning model.

References

1. Bailler, N., Buzanov, A., Gaillat, T., Vinogradova, O.: A Cross-platform Investigation of Complexity for Russian Learners of English. EUROCALL 2021, presentation at the conference (2021)
2. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. (2009)
3. Díaz-Negrillo, A., Valera, S., Meurers, D., Wunsch, H.: Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. Language Forum **36**, 139–154 (2010)
4. Education First. http://www.englishtown.com. Englishtown (2012)
5. Gablasova, D., Brezina, V., McEnery, T.: The trinity lancaster corpus: development, description and application. Int. J. Learner Corpus Res. **5**(2), 126–158 (2019)
6. Geertzen, J., Alexopoulou, T., Korhonen, A.: Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In: Selected Proceedings of the 2012 Second Language Research Forum, Somerville, MA, USA (2013)
7. Gilquin, G.: Learner corpora. In: Paquot, M., Gries, S.T. (eds.) A Practical Handbook of Corpus Linguistics, pp. 283–303. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46216-1_13
8. Gilquin, G., de Cock, S., Granger, S.: The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM. Press. univ. de Louvain, Louvain-la-Neuve (2010)
9. Granger, S.: Learner corpora. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics. An International Handbook, vol. 1, pp. 259–275. Walter de Gruyter, Berlin, New York (2008)
10. Granger S.: How to use foreign and second language learner corpora. In: Research Methods in Second Language Acquisition: A Practical Guide, ch. 2, pp. 5–29. Blackwell, Oxford (2012)
11. Granger S.: The contribution of learner corpora to reference and instructional materials design. In: The Cambridge Handbook of Learner Corpus Research, pp. 485–510. Cambridge University Press, Cambridge (2015)
12. Granger, S., Dupont, M., Meunier, F., Naets, H., Paquot, M.: The International Corpus of Learner English. Version 3. Press. univ. de Louvain, Louvain-la-Neuve (2020)

13. Huang Y., Geertzen, J., Baker, R., Korhonen, A., Alexopoulou, Th.: The EF Cambridge Open Language Database (EFCAMDAT): Information for Users, pp. 1–18. https://corpus.mml.cam.ac.uk (2017)
14. Lindquist, H.: Corpus Linguistics and the Description of English. Edinburgh University Press, Edinburgh (2009)
15. Lyashevskaya, O., Vinogradova, O., Panteleeva, I.: Automated assessment of learner text complexity. Assessing writing **49**, 100529 (2021)
16. Lyashevskaya, O., Panteleeva, I.: REALEC learner treebank: annotation principles and evaluation of automatic parsing. In: TLT 16, pp. 80–87 (2017)
17. Lyashevskaya, O., Vinogradova, O., Scherbakova, A. Accuracy, syntactic complexity, and task type at play in examination writing: A corpus-based study (forthc.)
18. McEnery, T., Brezina, V., Gablasova, D., Banerjee, J.: Corpus linguistics, learner corpora, and SLA: employing technology to analyze language use. Ann. Rev. Appl. Linguistics **39**, 74–92 (2019)
19. Meurers, D., Dickinson, M.: Evidence and interpretation in language learning research: opportunities for collaboration with computational linguistics. Lang. Learn. **67**(1), 66–95 (2017)
20. Nesi, H.: ESP and corpus studies. In: Paltridge, B., Starfield, S. (eds.) The Handbook of English for Specific Purposes. Handbooks in Linguistics Series, pp. 407–426. Wiley-Blackwell, Oxford (2013)
21. Nicholls, D.: The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In: Proceedings of the Corpus Linguistics Conference, pp. 572–581. Lancaster University: University Centre for Computer Corpus Research on Language (2003)
22. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK (1994)
23. Stenetorp, P., Pontus, P., Sampo, T., Goran, O., Tomoko, A. Tsujii, J.-I.: BRAT: A Web-based Tool for NLP-Assisted Text Annotation. In: EACL 13, Demonstrations, pp. 102–107. Stroudshourg, PA (2012)
24. Tetreault, J.R., Filatova, E., Chodorow, M.: Rethinking grammatical error annotation and evaluation with the amazon mechanical Turk. In: Proceedings of the Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 45–49. Los Angeles, USA (2010)
25. Vinogradova, O.: The Role and Applications of Expert Error Annotation in a Corpus of English Learner Texts. In: Computational Linguistics and Intellectual Technologies: Proceedings of Dialog 2016, pp. 740–751. Moscow, Russia (2016)
26. Vinogradova, O., Ershova, E., Sergienko, A., Overnikova, D., Buzanov, A.: Chaos is merely order waiting to be deciphered: corpus-based study of word order errors of Russian learners of English. In: Learner Corpus Research Conference, p. 115. Warsaw (2019)
27. Vinogradova, O., Smirnova, E. The L1 influence on the use of the English present perfect: a corpus analysis of Russian and Spanish learner essays (forthc.)
28. HSE Independent English Language Test regulations page. https://www.hse.ru/en/studyspravka/indexam. Accessed 20 Apr 2022
29. Learner Corpus Association page. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html. Accessed 5 May 2022
30. REALEC homepage. https://realec.org/index.xhtml#/exam. Accessed 5 May 2022
31. SpaCy homepage. https://spacy.io. Accessed 5 May 2022