M. A. Klimova, V. K. Smilga, D. A. Overnikova

Using an Error-Annotated Learner Corpus (REALEC) in DDL Lessons[1]

Abstract. The paper describes the experience of introducing the Russian Error-Annotated English Learner Corpus in data-driven learning with regard to error-prone lexical items. The corpus data on confusables were selected and used for preparing teaching materials of two types: the traditional deductive approach and the DDL inductive instruction. Russian L1 first-year university students participated in the corresponding lessons. The results obtained from their pre/post-tests and questionnaires show improvement in the use of most of the target items as well as a positive attitude towards DDL but also point to issues related to DDL materials design and the development of the REALEC corpus.

Key words. Learner corpora, data-driven learning, confusables, REALEC.

1. Introduction

Data-driven learning, i.e. using the tools and techniques of corpus linguistics for pedagogical purposes, is known to present several advantages: it exposes learners to real examples, performs a corrective function and includes an element of discovery [Gilquin and Granger 2010]. Unlike corpus-based language teaching, DDL not only uses corpus data in the preparation of learning materials, but gives learners access to substantial amounts of corpus data, either indirectly by allowing them to learn about language use by studying concordances prepared in advance by the teacher, or directly by allowing them access to corpora and concordancing software to carry out their own searches [Chambers 2010]. Although the direct hands-on approach provides greater learners' autonomy, prepared materials are known to lead to immediate benefits for learners and teachers with little or no experience

[1] The research was carried out within the project of the Higher School of Economics Research Foundation «2020 - Automated Analysis of text written in English by learners with Russian L1 (ADWISER)»

in corpus linguistics [Boulton 2010]. DDL activities have mainly addressed the fields of lexis and lexico-grammar with concordances being the major way of presenting the material [Smirnova 2017].

Learner corpora can be considered an important source for DDL as they present students with typical interlanguage features, especially when the data were produced by learners from the same mother tongue background as the students' [Gilquin and Granger 2010]. The existing empirical DDL studies involving learner corpora have shown positive results for the study of error-prone items [Cotos 2014; Moon and Oh 2018].

English learner corpora containing output from Russian L1 learners, such as SBbEFL, are known to be a valuable source for interlanguage analysis that could potentially contribute to the development of teaching materials [Kamshilova 2012]. REALEC (Russian Error-Annotated Learner English Corpus) has served as a basis for developing an automated testing tool [Vinogradova 2019] but has not yet been introduced into teaching practice using DDL, which is the purpose of the current study.

This paper aims to describe the experience of introducing DDL induction based on the data from the REALEC. As teacher and learner participants have little or no experience in both DDL and using the REALEC corpus in particular, an indirect approach is chosen in order to make DDL immediately accessible. With common learner mistakes in confusables as the target language feature, the study provides learners with guided indirect DDL in the test groups and traditional instruction in the control groups. As a result, both the first experience in DDL and the REALEC as a didactic tool are assessed using pre- and post-tests as well as feedback from the teachers and learners.

2. The REALEC corpus

The Russian Error-Annotated English Learner Corpus (REALEC) is a collection of essays and learner texts that consists of 13569 pieces of writing, the absolute majority of which are annotated. The corpus contains essays by students of HSE University and is first introduced into language teaching at the same institution, which makes the materials suited for the needs of the learners. As a learner texts corpus, REALEC implements a complex system of error annotation: each error is assigned a correction and one of 98 error tags belonging to 6 major groups — Punctuation, Spelling, Capitalisation, Grammar, Vocabulary and Discourse errors. The REALEC data has several advantages that could make it usable for teaching English as a foreign language, especially to Russian native speakers. Firstly, it can provides an insight into the mistakes most common among Russian L1 speakers, including language-specific ones. Secondly, the system of tags allows for a selection of suitable material even for most specific topics. The sentences used in pre-tests, post-tests, example-based explanations and exercises were obtained from the REALEC corpus.

3. The target language feature

When choosing the target language feature, we relied on previous research, which was concerned with using patterns of error distribution to determine which genre of academic essay featured in the REALEC a certain text is more likely to belong to [Vinogradova et al. 2020]. This allowed us to narrow down the most common mistakes in the corpus that, nevertheless, provided enough clear patterns of misuse.

When analysing the errors of all types in the corpus, we noted a large number - approximately 14,000 - of errors related to lexical choice. This particular group of tags included all REALEC tags grouped under “Word choice”, which were further divided into two main tags - “Choice of a part of a lexical item” and “Choice of a lexical item”, which contains a single subtag “Words often confused”. This subtag, covering both incidences of paronyms and near synonyms, interested us, as it covered a specific word choice mistake, yet at the same time had enough examples to establish patterns and provide the necessary material. However, we soon found that many examples which interested us were actually annotated with the tag “Choice of a lexical item”. This can be explained by the fact that, while the tags are considered by the BRAT software, on which REALEC operates, to be separate, one of them is also a subset of the other; therefore, the annotators seem to opt for a more general tag when they are unsure if the tag "Words often confused" is suitable.

After extracting all the instances of this tag from REALEC, we manually analyzed the errors and established a list of clusters of words most commonly confused for each other by English learners whose texts were submitted to REALEC. As a result, we were able to select the three clusters of confusables to be used in our lessons - near-synonymous numerical nouns (*amount*, *number*, *quantity*), near-synonymous nouns related to possibility (*possibility, opportunity, ability*, *potential*), and a pair of paronyms in the form of the verbs *note* and *notice*.

4. Participants

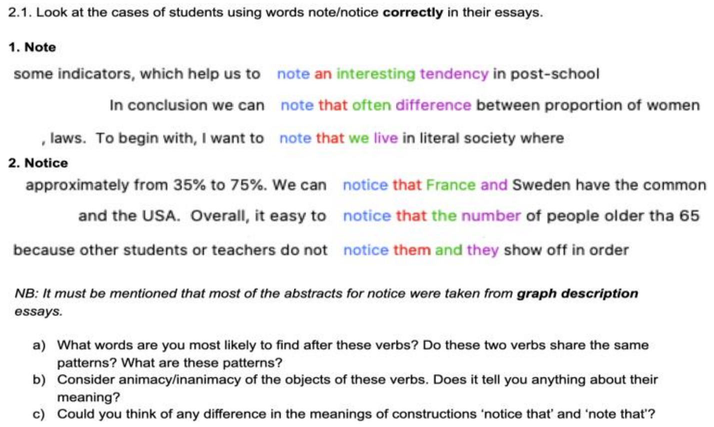
All the participants of the experiment were first-year Linguistics and Philology students of Higher School of Economics in Moscow and Nizhny Novgorod. 41 students attended the corpus-based lesson; 35 students attended the traditional one. In each case students and teachers from both of the HSE campuses were involved. The participants came from 8 learner groups; the lessons were given by 5 teachers. The specializations of the students made the experience especially useful for them as linguists and philologists will inevitably face corpora while studying or working in the future. Moreover, REALEC is primarily a collection of students' texts written after the completion of a 2-year English course at Higher School of Economics, which is why the material was immediately relevant to the participants.

5. Teaching materials and DDL intervention

Both lesson plans followed a similar structure, containing a pre-test, a three-part introduction of the target language feature, and the post-test. In the control group, this introduction was done using the deductive method: first, the rule outlining the proper ways of distinguishing between near-synonyms was presented explicitly to the students, followed by examples of sentences where the target confusables were utilized.

The test groups, on the other hand, were not presented with the rules explicitly. Instead, they received several concordance lines presenting instances of correct usage. The concordances were followed by a list of questions designed to encourage students to independently arrive at the rules controlling the correct usage of the target lexis.

After either familiarizing themselves with a rule or inducing it from the provided information, students in both groups completed 1 – 2 exercises testing their understanding of the rule. The exercises came in three types – multiple choice, where students had to choose the correct word out of several options, gap filling, which allowed students to write in their own suggestions, and error correction. The latter type of exercise contained 5 sentences, of which 4 contained mistakes and one did not, which students were explicitly told about. After each block of exercises was checked, the group moved on to the next rule.



2.1. Look at the cases of students using words note/notice **correctly** in their essays.

1. Note

some indicators, which help us to note an interesting tendency in post-school
In conclusion we can note that often difference between proportion of women
, laws. To begin with, I want to note that we live in literal society where

2. Notice

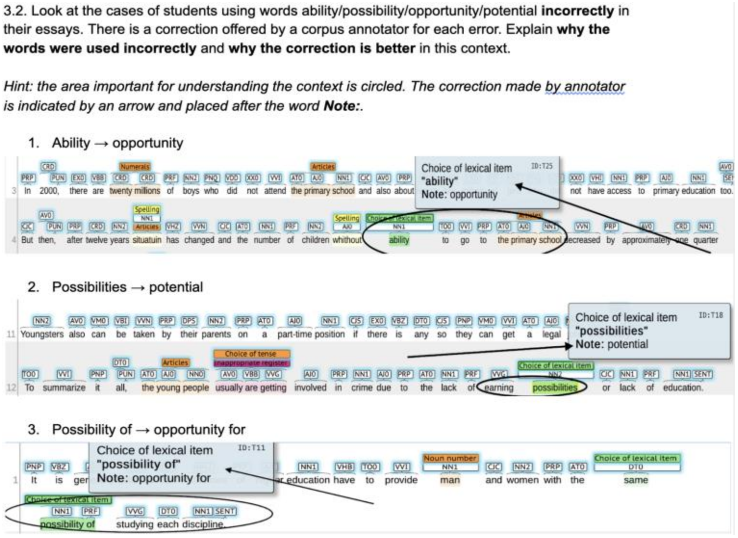
approximately from 35% to 75%. We can notice that France and Sweden have the common
and the USA. Overall, it easy to notice that the number of people older tha 65
because other students or teachers do not notice them and they show off in order

*NB: It must be mentioned that most of the abstracts for notice were taken from **graph description** essays.*

a) What words are you most likely to find after these verbs? Do these two verbs share the same patterns? What are these patterns?
b) Consider animacy/inanimacy of the objects of these verbs. Does it tell you anything about their meaning?
c) Could you think of any difference in the meanings of constructions ‘notice that’ and ‘note that’?

Fig. 1. **An example of the rule being induced in the DDL lesson.**

While both lessons utilized sentences obtained from REALEC as content for exercises, the test lesson introduced elements of DDL. The first element came in the usage of concordances, extracted from REALEC texts in advance using AntConc, which was emphasized by the use of screenshots including elements of its interface. This helped familiarize students with the concept of a concordance as well as corpus tools. Another way corpus data was integrated into the lesson was a task concerning the cluster of confusables related to possibility (*possibility*, *potential*, *ability*, *opportunity*). In this task, students were asked to analyse several examples of incorrect usage of these words, which were explicitly shown to come from REALEC and included error tags and corrections suggested by the annotators, as well as explanations regarding the interface of the corpus. This exercise was accompanied by an error-correction task, which closely followed the process of tagging in an error-annotated corpus.



3.2. Look at the cases of students using words ability/possibility/opportunity/potential **incorrectly** in their essays. There is a correction offered by a corpus annotator for each error. Explain **why the words were used incorrectly** and **why the correction is better** in this context.

Hint: the area important for understanding the context is circled. The correction made by annotator is indicated by an arrow and placed after the word ***Note:***.

1. Ability → opportunity

In 2000, there are twenty millions of boys who did not attend the primary school and also about … not have access to primary education too.

But then, after twelve years situatuin has changed and the number of children whithout ability to go to the primary school decreased by approximately one quarter

Choice of lexical item "ability" ID:T25
Note: opportunity

2. Possibilities → potential

Youngsters also can be taken by their parents on a part-time position if there is any so they can get a legal …

To summarize it all, the young people usually are getting involved in crime due to the lack of earning possibilities or lack of education.

Choice of lexical item "possibilities" ID:T18
Note: potential

3. Possibility of → opportunity for

It is ger… education have to provide man and women with the same possibility of studying each discipline.

Choice of lexical item "possibility of" ID:T11
Note: opportunity for

Fig. 2. **An exercise in the DDL lesson showing explicit usage of corpus data.**

In providing instructions for the DDL lesson, we had to consider our target audience, who came from programs centered around philology and linguistics, which usually do incorporate corpus usage in their curriculum. However, we assumed that while the concept of the corpora was not completely new to them, their knowledge as first-year students most likely would not be advanced. The survey conducted after the experiment confirmed that, while most of them did use corpora at some point, the majority used corpora only occasionally, and mainly used Russian corpora. Therefore, while we aimed to introduce students to the concept of using corpora in L2 studies, we also tried to familiarize them with the layout and inner workings of corpus tools, which we hope will be useful for their exploration of corpora in the future.

6. Pre- and post-tests

In order to evaluate the students' progress, we created a pre-test and a post-test. Both of them were similar in their content and consisted of two parts containing five questions each. The first part consisted of multiple-choice questions with four options, containing a combination of target units and other words of similar grammatical form and meaning. The second part was an error correction task, containing five sentences in which the target word was highlighted.

7. Results and discussion

As it was expected, in all groups students' performance increased after they got acquainted with the rules of word usage. However, the evaluation of students' performance before and after the lesson showed no major difference in the two approaches. In the traditional groups the average student performance improved by 13,03%, while in the corpus-based ones the improvement was 12,18%.

By the design of the experiment our observations for corpus-based and traditional groups were independent and the resulting data was homogeneous as Levene's test yielded the value of 0.5278592809.

Logarithmic transformation was applied to data to make its distribution closer to normal.

Table 1. **Student performance in groups with traditional and DDL-based approaches before and after the lessons.**

	Traditional lessons	**Corpus-based lessons**
Pre-test	70%	71,25%
Post-test	83,03%	83,43%
p-value	0.0001698674	0.0011476102

These three factors allowed for the use of two-way Analysis of Variance (ANOVA) to understand how two independent variables (*student's attending a corpus-based or traditional lesson* and *the test being pre- or post-test*) affected the students' performance (*the number of mistakes made in a test*). The resulting p-value was less than 0.05 for the *test* variable only. That lets us draw the conclusion that students' performance differed significantly on pre- and post-tests (p-value=0.0000000705), while the difference in performance between those who attended the corpus-based lesson and those who attended the traditional one was not statistically significant (p-value=0.9413137431). p-value was also calculated separately for the participants of traditional and corpus-based lessons using one-way ANOVA, as shown in Table 1. In both cases p-value was less than 0.05, meaning that the difference in performance on pre- and post-test was statistically significant in both groups.

Having proven that the difference in student performance on pre- and post-test, unlike the difference in student performance on corpus-based and traditional lessons, was statistically significant, we can now take a closer look at the percentages indicating student performance on different sets of words (tables 2 and 3).

Table 2. **Traditional groups' performance.**

	Amount, number, quantity	**Possibility, ability, opportunity, potential**	**Note, notice**
Pre-test	56,29%	80,07%	69,66%
Post-test	90,09%	72,29%	82,42%

Table 3. **Corpus-based groups' performance**.

	Amount, number, quantity	**Possibility, ability, opportunity, potential**	**Note, notice**
Pre-test	49,52%	79,99%	71,43%
Post-test	91,67%	74,25%	86,37%

If we take a closer look at the student performance, we will see that the most significant improvement was achieved for groups of words with unambiguous rules, such as *amount/number/quantity*, whose usage is entirely defined by the dependent noun's characteristics, and *note/notice*, which have unambiguously different Russian translations *отметить/заметить*. All groups' performance for the last set of words, *possibility/opportunity/ability/potential*, worsened after the lesson. This situation may be accounted for by the size of the word group and our decision to include the word 'potential', which caused most confusion among the students during the lesson, as one of the teachers observed. These words have no clear rules of usage motivated by grammar and are usually translated into Russian by the same word (*возможность* or *возможности*). While the approach was effective overall, more examples and explanations were needed in some cases, as one of the students noted in the survey.

The survey answers point out that the major feature of the DDL-based approach was it being unusual and interesting both for those who teach and those who are taught. The experience was generally evaluated as positive: the average willingness of students and teachers to participate in and conduct such classes in the future was 4 out of 5. They were primarily interested in topics similar to the one offered during the lesson, but also showed interest in classes on articles, prepositional phrases and set expressions with the use of corpus material.

Several students commended the use of corpora in the learning process for giving them an opportunity to 'identify the patterns of word usage by themselves' (5 students) and get acquainted with 'real contexts of using the words' (4 students). Another advantage of the DDL-based lesson was it being 'illustrative' (4 students) and implementing more practice than theory (3 students). As well as students, teachers noted that the material was illustrative and touched on important topics that tend to cause much confusion among students.

All in all, the lesson implementing DDL method was as effective in terms of rules acquisition as the one conducted in traditional style. What is more, students' and teachers' comments have shown that the use of corpus in learning may help to enliven English lessons and make the learning process more dynamic.

We can assume that the difference between the lesson implementing the DDL approach and the traditional one will be more visible in the long-term. That is why the future directions for our work may include conducting a similar experiment with delayed post-testing in order to reveal the retention of vocabulary knowledge, as suggested by T. Cobb [Cobb 1999]. What is more, in our further experiments we should introduce less information in one lesson, as it turned out to be rather difficult for the students to acquire three different groups of words within a relatively short time.

8. Conclusion

Being the first step in the introduction of both the DDL approach and REALEC as a didactic tool in first-year university EFL lessons, the study compared guided DDL induction with traditional deductive learning. The results suggest that the potential of exposing learners to DDL activities based on the REALEC corpus is as significant as using the same corpus data in traditional instruction with regard to confusables as the target language feature.

Although the findings for particular groups of lexis are controversial, the results of the survey confirm the pedagogical value of the DDL approach in terms of the motivation and attitude of both teachers and students. However, more examples in concordances and post-testing should be added to improve the design of further experiments. While the present study, like most DDL interventions so far, was focused on lexis, it has demonstrated that there is a call for data-driven learning related to a wide range of topics, including grammar and discourse. The range of application can be also broadened by combining REALEC data with the use of native speaker corpora.

Some implications can be drawn concerning the annotation of REALEC and its interface. Although the error annotation of the corpus proved usable for EFL teaching, tags for lexis, in particular “Choice of a lexical item” and its subtag “Words often confused”, need to be better differentiated in the annotation scheme. Following the indirect approach as a lead-in to DDL, the current study did not include independent use of REALEC by the students. Using more open-ended exploration in EFL lessons would require a user-friendly concordancing tool that would make it possible for students to carry their own searches.

References

1. Boulton A. (2010), Data-Driven Learning: Taking the Computer out of the Equation, *Language Learning*, 60(3), pp. 534–572.

2. Chambers A. (2010), What is Data-Driven learning, *The Routledge handbook of corpus linguistics*. Routledge, pp. 345–358.

3. Cobb T. (1999), Breadth and Depth of Lexical Acquisition with Hands-On Concordancing, Computer Assisted Language Learning, 12(4), pp. 345–360.

4. Cotos E. (2014), Enhancing writing pedagogy with learner corpus data, *ReCALL*, 26(2), pp. 202–224.

5. Gilquin G., Granger S. (2010), How can data-driven learning be used in language teaching, *The Routledge handbook of corpus linguistics*. Routledge, pp. 359–370.

6. Kamshilova O. (2012), Learner Language analysis in SPbEFL Learner Corpus, *LLLC 2012 Abstracts*, p.40.

7. Moon S., Oh S. (2018), Unlearning overgenerated be through data-driven learning in the secondary EFL classroom, *ReCALL*, 30(1), pp. 48–67.

8. Smirnova E. A. (2017), Using corpora in EFL classrooms: The case study of IELTS preparation, *RELC Journal*, 48(3), pp. 302–310.

9. Vinogradova O. (2019), To Automated Generation of Test Questions on the Basis of Error Annotations in EFL Essays: a Time-Saving Tool?, *Learner Corpora and Language Teaching*, 92, pp. 29–48.

10. Vinogradova O., Lyashevskaya O., Smilga V. (2020), Correlations between Accuracy, Complexity, and Task Type: *Learner Corpus Research* (in press).

Klimova Margarita A.
HSE University (Russia).
E-mail: mfokina@hse.ru

Smilga Veronika K.
HSE University (Russia).
E-mail: smilgaveronika@gmail.com

Overnikova Daria A.
HSE University (Russia).
E-mail: daovernikova@edu.hse.ru