International Conference
The Future of Education

Use of Learner Corpus in General English and Academic English Courses at the Higher School of Economics[1]

Olga Vinogradova[2]

Abstract

There have been many reports on advances in the development of learner corpora that have made it possible to effectively use these collections of texts for the benefit of the learning process. This paper lists all possible applications in English courses taught to Bachelor students of a middle-size learner corpus REALEC, which comprises student written works supplied with expert annotation of mistakes, browsing and search options, and some optional automated tagging system. Annotation in the corpus is given by either experts (mostly, EFL instructors), or by learners themselves under the supervision of their EFL instructors. As the first point, the paper argues that when EFL methodology requires that students apply the error classification in the process of annotating their peers' essays and gradually their own essays as well, their understanding of subtle areas of grammar, vocabulary and discourse improves, and correspondingly, the number of errors in their written works decreases. The second argument concerns the tool for the development of placement and progress tests, which makes use of sentences with mistakes made by other learners – contributors to the corpus. In the suggested design of the tests sentences are automatically extracted from the same corpus, manually divided into three echelons according to the complexity of the change required in the correction of the mistake, and then administered to learners as a way of automated measurement of their proficiency in English. The submitted test is scored automatically within minutes. The third possibility considered in the research is the possibility to supplement the corpus with the platform of trainers automatically or semi-automatically set up on the basis of frequently marked errors made by a particular group of students. In conclusion we point out the ease and usefulness of the proposed applications both for EFL instructors and English learners.

Introduction

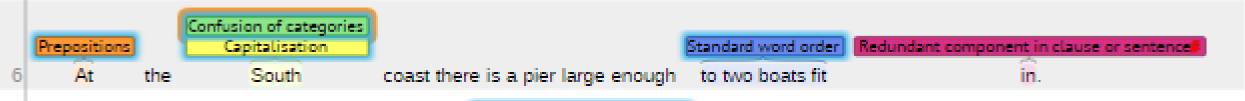
It has been proved over more than twenty years of research that having access to learner corpora is of great benefit for the process of L2 acquisition for both learners and instructors (see the overview of this area in [1]). Besides language acquisition, corpora studies have been in the focus of computer linguistics over the past two decades ([2], [3], [4]). Both these points account for the fact that EFL instructors teaching English to students specializing in computer linguistics at the School of Linguistics set up a learner corpus REALEC, Russian Error-Annotated Learner English Corpus (http://www.realec.org/). It is the first in the open access collection of texts in English written by Russian students learning English, and anyone interested in learner errors can carry out a search in it or even download the materials. The results of the first experiment evaluating inter-rater agreement in REALEC were presented at the 8th International Conference *Corpus Linguistics 2015* ([5]). The technological novelty of REALEC is the combination of a few original ideas.

First, text processing in REALEC includes two stages - automated tagging carried out in the open access tool - Freeling suite of linguistic analyzers ([6]), and manual expert annotation in another open-source tool - Brat annotation framework ([7]). Expert annotation is based on linguistically advanced error classification scheme. The present taxonomy comprises seven main areas of linguistic description: punctuation, capitalisation, spelling, morphology, syntax, lexis and discourse, and the last four are subdivided further into specific subcategories. Annotators have to suggest the correction of the error span, which can be seen below the name on the tag when a cursor hovers over the tag.

For both English instructors and their students it was important that mistake patterns typical of each student be demonstrated clearly, that was why REALEC's user-friendliness was paid much attention to. As a result, even a quick glance at their works allows students to see what mistakes are more frequent because the mistake tags of each type are marked in a certain colour (see Fig. 1).

[1] The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2015-2016, and the author is a member of the team that has won a Research Team Project Competition in 2016 (16-05-0057 at https://www.hse.ru/en/science/scifund/nug)

[2] School of Linguistics, National Research University Higher School of Economics, Russia



6 At the South coast there is a pier large enough to two boats fit in.

Fig. 1 A sentence from a student essay with errors annotated

The second factor is that practical activities in the corpus given in English classes are evaluated and followed up by the research team consisting of EFL specialists and computer linguists (the name of the team is “REALEC for Real Words” (https://realec-nug.wikispaces.com/ (in Russian)). Researchers in the project team are responsible for suggestions, changes and directions for development of REALEC.

The main direction in the focus of this article is to show how effective the exposure to the learner corpus can be for both students and EFL/ESL specialists. I am going to demonstrate how the convenience of using the corpus as a learning management system can be increased and how the use of corpus materials allows English instructors to make their teaching better adjusted to the needs of their groups.

Methodology

It has been proved ([8], [9]) that attempts to apply the writing criteria to evaluating essays written by other students – peer evaluation – is a more efficient method than just presenting sample essays to students and highlighting those features in the essays that have led to the scores which these essays have been assigned by experts or examiners.. In our methodology the instruments that the learner corpus provides – namely, easy search tools and clear categorization of errors – seem to increase the educational efficiency manifold. That was why we set up work in class by involving students in work with REALEC at the following four stages:

(1) annotating mistakes that their instructor has outlined in the essays written by their peers;

(2) analysing mistakes annotated in their own essays;

(3) comparing the score they have assigned to the essay under consideration with their instructor’s grading;

(4) trying to spot errors in their own essays and annotate them under the supervision of their instructor.

One more – optional – activity, usually given just once as an experiment, asks the whole group of students to annotate and evaluate one and the same essay so that they can discuss differences in their annotations and scores they have ascribed. All these activities are supposed to give students the best idea of what writing strategies are expected of them in Academic IELTS examination essays.

Research context

The research was carried out over the essays collected in REALEC in the process of preparation for an IELTS-type examination and in the examination. Table 1 gives the breakdown across error domains.

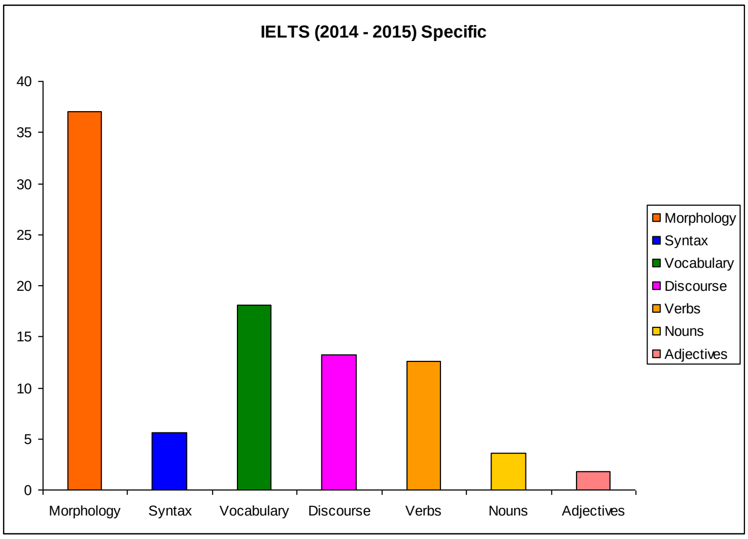


Figure 2 Variation of error types in IELTS

Research subjects were 1st- and 2nd-year undergraduate students, all in their late teens or early twenties, who took a course of general English as their first foreign language. The level of their English proficiency can be roughly described as upper-intermediate (C1 in CEFR). A number of English language instructors and examination experts are involved in the process of evaluating students' works. In spite of good level in English proficiency, students have to catch up on the areas of academic English necessary, on the one hand, for their studies in general, as some courses are taught in English, and, on the other, for their success in Academic IELTS: students are aware that Academic IELTS results constitute important criteria in students' rankings. Besides, for students majoring in linguistics, this approach is of double benefit as corpus research methods constitute a part of the curriculum in the course "Computational Methods in Linguistics."

Pedagogical tools based on corpus materials

The primary goal in the current research is to present a tool for EFL/ESL instructors to adjust their teaching to the needs of a particular group of students. With this in mind, REALEC research team developed two scripts, the first aiming at automated or semi-automated creation of lexical training exercises on the basis of the texts in any database or corpus, and the second devoted to semi-automated generation of placement and progress tests from the sentences with annotated errors. Both projects are still in progress, but the results of their application can already be discussed. The first design is to be reported in [10]. This paper presents the second script, a test-maker called RETM – REALEC English Test Maker, which extracts sentences with mistakes from student texts collected in a particular area of REALEC to form the pool of questions for a test. The technical side of the script is presented in [11]. Here I describe the strategies undertaken in the process of writing RETM script and present some results in the next section. The five main areas we had to address in the process of developing RETM were the following:

1. Choice of what to test. Any sentence with the mistake tagged in it can be regarded as relevant material for the test. Automatic generation implies that a test-taker will have to correct what (s)he sees as an error, and his/her correction will be compared with the one given by an expert in the annotation: if they coincide, then the test-taker has won a score assigned to the question. However, some mistakes are more difficult to spot than others, and, moreover, a few are very difficult, if not impossible, to categorise. There are also mistakes that learners make very rarely, as well as accidental slips, and these should not be included on the test. As a result, instructors have to decide which tags of the 151 in the scheme are going to be used in test questions.
2. Selection of sentences for the pool of questions. The instructor looks at each sentence of the automatically generated pool and first deletes those where it is impossible for a learner to spot the mistake. If the sentence is approved, the instructor chooses between three options allowed by RETM – highlighting the error span, giving the sentence without any highlighting, or giving the sentence as a multiple-choice question. In future, other types of tests will be added.
3. Preparation of the selected sentences according to the level of difficulty it poses for a learner. At present the system allows to assign any question one of the three levels – the lowest (1 point), middle-level (2 points), and the highest (3 points). If for some reason it is necessary, the number of levels can be increased or decreased.
4. Test strategy and evaluation. The test is organised in the following way: all test-takers get the same number of questions randomly chosen from the pool. The first question is always at the lowest level, and if a student gives the correct answer to it, the next question is taken from the pool of middle-level difficulty, but if the answer was wrong, the next question is also of the lowest level. As a result, the score is going to be higher if a test-taker reaches the higher levels more often.
5. Analysis of the testing statistics. At the end of the test, a test-taker gets the number of correct answers, the number of correctly spotted error spans with the wrong correction suggested, and all the wrong answers are presented along with the expected answers in a way of feedback. The instructor, in turn, gets the statistics for the whole group in the form of the list from the best to the worst. If the test was administered as a placement test, the system offers to add other criteria to sort out the division of students into the necessary number of groups. In case of a progress test, a test-taker with the low score can be urged to take the test one more time. The same questions are excluded then.

Conclusions and implications

According to the research, work with the learner corpus has helped

- English instructors to set up new methodology of efficient preparation for the Writing section of IELTS-type examination;
- students to develop efficient methods of searching learner corpus for DOs and DONTs in writing essays;
- computer linguists to point out systematic discrepancies between annotations given by experts and those given by students (as a result annotation practices in REALEC will be improved);
- ESL/EFL instructors to get insights into the teaching methods required for their particular groups of students;
- ESL/EFL instructors to use automated tests customized to the needs of their groups instead of making those tests by hand.

References

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

[2] Leech, G. *Adding linguistic annotation*. In Developing linguistic corpora: a guide to good practice. Oxbow Books, Oxford, pp. 17-29, 2005

[3] McEnery, T. and Xiao R. *What corpora can offer in language teaching and learning* In Hinkel, E. (ed.), Handbook of Research in Second Language Teaching and Learning. London: Routledge. 2011

[4] Granger, S. *The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research*, TESOL Quarterly, 37, 3, 2003, p. 538-546

[5] Kutuzov, A., Kuzmenko, E. and Vinogradova, O. *Evaluating inter-rater reliability for hierarchical error annotation in learner corpora* in The proceedings of 8th International Corpus Linguistics Conference, Lancaster 2015, pp. 211-214.

[6] Lluis, P. and Stanilovsky, E. *Freeling 3.0: Towards wider multilinguality*. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012. Also http://nlp.lsi.upc.edu/freeling/

[7] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, O. and Tsujii, J *brat: a Web-based Tool for NLP-Assisted Text Annotation.* In Proceedings of the Demonstrations Session at EACL 2012. Also http://brat.nlplab.org/

[8] Myles, F. (2005) *Interlanguage corpora and second language acquisition research* - in Second Language Research 21, 4 373-391

[9] Marinov, S (2011) *Training ESP students in corpus use - challenges of using corpus-based exercises with students of non-philologic studies* -Teaching English with Technology, 13(4), 49-76

[10] Fenogenova, A. and Kuzmenko, E. *Automatic Generation Of Lexical Exercises* - in Proceedings of the International Conference “Dialog 2016”

[11] Kustova, M. *Design of Corpus-generated EFL Placement and Progress Tests for University Students* - in The Future of Education, 2016