USE OF LEARNER CORPUS IN GENERAL ENGLISH AND ACADEMIC ENGLISH COURSES AT THE HIGHER SCHOOL OF ECONOMICS[1]

Olga Vinogradova, School of Linguistics, National Research University Higher School of Economics

olgavinogr@gmail.com

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.

Key words: EFL/ESL methodology, learner corpus, error annotation, automated trainers; corpus-generated tests

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, for example, the overview of this area in Granger, S., Gilquin, G. and Meunier, F. (2013)). Besides language acquisition, corpora studies have been in the focus of computer linguistics over the past two decades (Leech 2005, McEnery , Granger 2003). Both these points account for the fact that EFL instructors teaching English to students specializing in computer linguistics set themselves the task of setting up a learner corpus.

REALEC, the corpus set up at the School of Linguistics (Higher School of Economics) (http://www.realec.org/), is the first in collection of English texts written by Russian students learning English the open access, and anyone interested in learner texts or errors can

[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 authors are 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**).

carry out a search in it or even download the materials. The task of developing a learner corpus was undertaken by computer linguists along with EFL/ESL professors. Some of the first observations were presented by Kutuzov, Kuzmenko and Vinogradova in 2014, and the results of the first experiment evaluating inter-rater agreement in REALEC were presented at the 8th International Conference *Corpus Linguistics 2015* (Kutuzov, A., Kuzmenko, E. and Vinogradova, O. (2015)). The technological novelty of REALEC is the combination of a few original ideas.

First, REALEC applies an open-source tool Brat annotation framework (Stenetorp, P. (2012)). The justification of this decision is given in Kutuzov, A. and Kuzmenko, E. (2013).

Second, text processing includes two stages - automated tagging, again in the open access - Freeling suite of linguistic analyzers[2] (Padro 2012). The results of this procedure are present in the database but are not conventionally visible; however, they can be displayed to the user on demand. This stage is then followed by annotation based on linguistically advanced error classification scheme, which provides for thorough detailisation in marking students' errors. The present taxonomy comprises seven main areas of linguistic description, namely punctuation, capitalisation, spelling, morphology, syntax, lexis and discourse, the last four of which are subdivided further into specific subcategories. Appendix 1 shows the full error classification scheme.Additionally, annotation process incorporates information about how serious the error is in the structure of English, how badly the error affects understanding, and the possible cause of the error, but these three layers of annotation are not reflected in the present paper. What is essential in this paper is the fact that annotators suggest their correction of the error span, which can be seen below the name on the tag by placing a cursor on the tag.

For both English instructors and their students it was important that mistake patterns typical of each student should 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 mistakes of each type are marked with tags of a certain colour (see example from REALEC in Fig. 1).

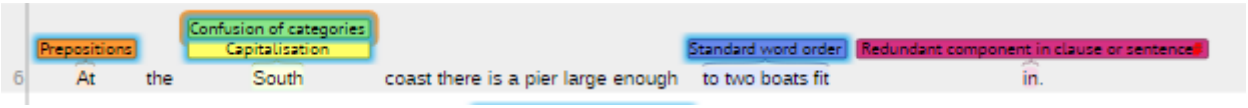

Fig. 1 A sentence from a student essay with errors annotated

The third factor that makes REALEC an interesting field of study is that students' and their instructors' practical activities in the corpus throughout their English course 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 a particular methodological approach can increase the convenience of using the corpus as a learning management system, how the students can improve their understanding of which score their own essay may get them in the exam, and finally, how the use of corpus materials allow English instructors to make their teaching adjusted to the needs of their groups.

Section 1. Methodology

[2] ***http://nlp.lsi.upc.edu/freeling/***

It has been proved 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. (Myles, F. (2005), Marinov, S. 2011). 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 and independent preparation for the exam involving students in work with REALEC at three stages: (1) annotating mistakes that their instructor has outlined in the essays written by their peers; (2) analyzing mistakes annotated in their own essays; (3) comparing the score they have assigned to the essays 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.

Section 2. 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. By the time these essays are looked at by the next generation of students preparing for IELTS, they have already been evaluated by IELTS experts.

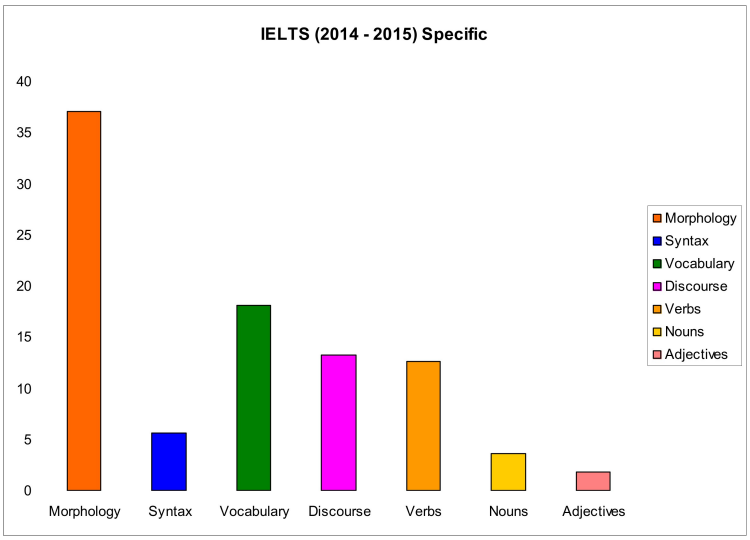

Figure 2 Variation of error types in IELTS

Research subjects were 1st- and 2nd-year undergraduate students majoring in Computer Linguistics, all in their late teens or early twenties, who took a course of general English as their first foreign language with preparation for IELTS as a part of it. 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 for their studies in general, as some courses are taught in English, and for their success in Academic IELTS in particular: students are aware that Academic IELTS

results constitute important criteria in students' rankings. Besides, this approach is of double benefit for students majoring in linguistics, as corpus research methods constitute a part of the curriculum in the course "Computational Methods in Linguistics."

Section 3. A pedagogical tool 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 is devoted to semi-automated generation of placement and progress tests from the sentences with annotated errors. The finalization of both projects is still in progress, but the results of their application can already be discussed. The first design is to be reported by Fenogenova and Kuzmenko in (Fenogenova & Kuzmenko 2016, paper in print). 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. The technical side of the script is presented in the paper "Design of Corpus-generated EFL Placement and Progress Tests for University Students" by Marina Kustova (Kustova 2016). 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. Methodological justification of the decisions on what to test. The statistics on the types of errors in the corpus (Fig. 1 above) shows that students make all kinds of mistakes in their writing, and, correspondingly, in theory, 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 extremely 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 first decide which tags of the 151 in the scheme are going to be used in questions, and after the automated process of extracting questions for the test with the tags outlined is finished, the process moves to the stage when the instructor will have to finally select appropriate questions among those automatically generated.
2. Selection of sentences for the pool of questions. At this stage, the instructor looks at each sentence and for a start decides whether it is possible for a learner to spot the mistake in this sentence (if not, the sentence is deleted). 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 are expected to be added.
3. Preparation of the selected sentences. There is one more decision to make about each question in the pool – it is about 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). In future distribution of sentences to the three levels can be further automated, but the need for manual approval and possible change will probably stay on. If for some reason it is necessary, the number of levels can be increased or decreased.
4. Structure and evaluation scheme of the results. 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 more correct answers the student gives, the higher the resulting score is going to be.

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 number of groups given by the instructor. If the test was administered as a progress test, a test-taker with the low score can be urged to take the test one more time to see if they can do better. The system then takes care of not including the same questions the second time. Then, there is an opportunity for the instructor to get the statistics across questions – howmany times each of them was answered correctly/incorrectly, and with this data, the system asks the instructor whether the question can stay on or should be deleted from the pool. As the final stage for the instructor, the system offers to draw the comparison between different groups.

Section 4. Conclusion 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 by students and as a result improve annotation practices in REALEC;

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.

One more issue to be considered is related to the predictive validity of IELTS results for academic performance. There have been many works devoted to this issue (e.g. Atkinson, D. and Valle, E. (2013), Davies, A. (2008), Dooey, P. and Oliver, R. (2002), Hall, G. (2010), Kerstjens, M. and Nery, C. (2000), O'Loughlin, K. and Arkoudis, S. (2009), Piggin, G. (2012)). There is no doubt this validity underlies selection of IELTS scores as the main criteria in language proficiency evaluation at the Higher School of Economics. However, the statistical processing of the correlation between IELTS scores and the grades earned for English proficiency demonstrated in the thesis submitted in English language in the fourth (final) year of the Bachelor's Programme has been carried out in the suggested research with all the specifics of Russian educational background at the Research University of Higher School of Economics taken into account.

References

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.

Kutuzov, A., Kuzmenko, E. and Vinogradova, O. (2014) Использование корпусных технологий для изучения ошибок: learner corpora на факультете филологии НИУ ВШЭ - в отчете о работе лаборатории корпусных технологий, май 2014. (Elizaveta

Kuzmenko, Andrey Kutuzov Using corpus technologies for research into student errors: learner corpora at the Philological Department in the Higher School of Economics)

Kutuzov, A., Kuzmenko, E. and Vinogradova, O. (2015)

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

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

Novikova, A. (2014) Verb pattern errors in Russian learner corpus. - 3rd-year course work, archive of Department of Philology, Higher School of Economics, Moscow

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.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, O. and Tsujii, J. (2012)

brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the*

Demonstrations Session at EACL 2012. Also http://brat.nlplab.org/