Corpus linguistic technology as a digital tool in teaching idioms’ interpretation to EFL students

Ekaterina Osipova
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation
E-mail: kitty_novgorod@mail.ru

Abstract. Corpus linguistics technology is considered from the linguodidactic perspective. Discussion of the results of extensive empirical research into efficient ways of teaching EFL (English as a Foreign Language) students to interpret idioms has been presented. The research comprises three steps: ascertaining, educational and control experiments. The aim of the paper is to investigate the efficiency of corpus linguistic technology as a tool in developing corpus-based idioms’ interpretation skill (CBIIS). The methodology has been tested on the EFL students of Bachelor study programs in Civil Engineering, Electrical and Electrical Power Engineering, Machine-building and Information Science and Computer Engineering implemented by Peter the Great St. Petersburg Polytechnic University in the years 2018-2019. The author shares personal experience of how to use corpus linguistics technology in foreign language teaching, presents a set of corpus-based exercises and describes the procedure of the learners’ work with the basic corpus tools while interpreting idioms. The result of the educational experiment shows that the average level of the CBIIS proficiency has increased that confirms the hypothesis of the research on the effectiveness of corpus linguistic technology as a tool in teaching idioms’ interpretation to EFL students.

1. Introduction
Corpus technology has already proved its validity for plenty of linguistic studies, which might be separated into several groups: (a) quantitative studies of English lexical and grammatical features [1,2], (b) studies on corpus methodology and design [3,4], (c) corpus-based studies on linguistic changes (lexical, syntactical and stylistic) [5,6], (d) corpus-based discourse analysis [7-12].

Some papers focus on some special tools of corpus linguistics, for instance, concordancer or KWIC (key words in context) and how to use them to solve linguistic issues. For example, Seretan and Wehrli [13] present “an enhanced type of concordancer that integrates syntactic information on sentence structure as well as statistical information on word co-occurrence”.

Not only might the linguistic issues be considered with the help of corpus linguistic technology, but also linguodidactic ones.

The advantages of corpus technology for language learning and second language learning are widely presented in the literature [14-18]. The researchers suggest to use corpus technology for the following purposes: a) as an empirical component of lectures, student assignments and projects; b) to determine the meaning of words and identify differences in usage between synonymous lexical items; c) to study lexical collocations; d) to focus on linguistic evidence that either supports or contradicts the prescriptive grammar rules [19-23].

The authors see the methodological value of linguistic corpora for some reasons. Firstly, they provide information about the real state of language, its historical, geographical and social variation.
Secondly, they illustrate speech registers and genre diversity. Thirdly, they expand understanding of the functions of some language units. Fourthly, they allow to identify the most frequent language phenomena and expand students’ vocabulary [24].

This research is targeted at supporting the idea of using linguistic corpora as an effective tool in teaching foreign language, because it develops students’ critical attitude to foreign language and facilitates deeper understanding of some linguistic phenomena, including idioms’ comprehension.

The object of the research is the process of teaching idioms’ interpretation to EFL students. The subject of the research is corpus linguistic technology as a digital tool in teaching EFL students. The aim of the research is to investigate the efficiency of corpus linguistic technology as a tool in developing corpus-based idioms’ interpretation skill.

The research tasks: (a) to assess the necessity of the corpus-based idioms’ interpretation skill (CBIIS) development (ascertaining experiment); (b) to assess the level of the CBIIS proficiency before the educational experiment; (c) to describe the methodology targeted at developing CBIIS (educational experiment); (d) to assess the level of the CBIIS proficiency after the educational experiment (control experiment).

2. Methods

2.1. Participants
A total number of 225 EFL students from Peter the Great St. Petersburg Polytechnic University took part in the educational experiment in the years 2018 and 2019. They were students of the first and the second year, getting Bachelor’s degrees in Civil Engineering, Electrical and Electrical Power Engineering, Machine-building or Information Science and Computer Engineering.

2.2. Procedure
The conducted research comprises three steps.

1) Ascertaining experiment is aimed at finding out (a) what ICT (Information Communications Technologies) are widely used by students in the process of idioms’ interpretation, (b) the level of English proficiency and knowledge of English idioms, (c) proficiency in translation techniques.

2) Educational experiment is targeted at building and developing the corpus-based idioms’ interpretation skill (CBIIS) and consists of a set of corpus-based activities.

3) Control experiment is needed to check the hypothesis that corpus linguistic technology is an effective tool in teaching idioms’ interpretation to EFL students.

To assess the level of the corpus-based idioms’ interpretation skill (CBIIS) proficiency four criteria of correctness – linguo-cultural, rhetorical, technical, corpus – were taken into account.

Linguo-cultural correctness – the correctness of the transmission of cultural semantics in the process of idioms’ interpretation.

Rhetorical correctness – the correctness of the transmission of the communicative effect in the process of idioms’ interpretation.

Technical correctness – selection of the adequate translation equivalences in the process of idioms’ interpretation.

Corpus correctness – the correctness of the sequence of corpus queries, choice of correct corpus options (LIST, CHART, KWICK, COMPARE), relevance of the chosen linguistic corpus and ability to draw conclusions from the obtained corpus data.

If an idiom is interpreted correctly according to all the listed criteria, 1 point is given. The percentage of correct translation of idioms by one student in relation to the total number of idioms in the test was used to assess the level of CBIIS proficiency. Here is the formula (1):

\[
K = \frac{N_1}{N_2} \times 100\% 
\]
K – percentage rate of the level of CBIIS proficiency,
N1 – the number of correct idioms’ interpretation given by one student in the test,
N2 – the total number of idioms, offered to the students for interpretation in the test.

The obtained data was processed according to the scale of the level of the CBIIS proficiency (Table 1).

**Table 1.** The scale of the level of the CBIIS proficiency.

| The percentage of correctness | the level of CBIIS proficiency          |
|-------------------------------|-----------------------------------------|
| 0-20%                         | Low level                               |
| 21-40%                        | Weak level                              |
| 41-60%                        | Middle level                            |
| 61-75%                        | Communicative limited level             |
| 76-90%                        | Communicative sufficient level          |
| 91-100%                       | Proficient level                        |

The level of the CBIIS proficiency was assessed twice: before the educational experiment (pre-experimental stage) and after the experiment (post-experimental stage) to discover the development.

To evaluate the dynamics of changes of the level of the CBIIS proficiency the average percentage was calculated by the formula (2):

\[
x = \frac{(K_1+K_2+...+K_{25})}{n}
\]

X – the average percentage of the level of the CBIIS proficiency,
K – percentage rate of the acquired level of CBIIS proficiency of a student,
n – total number of students.

2.3. Methods

For the solution of the research tasks a series of methods was applied: a) methods of a theoretical analysis (analysis, synthesis, generalization, modeling, comparative); b) empirical methods (a questionnaire, a set of experiments: ascertaining, educational and control); c) a tabular, graphical and corpus-based presentation of information; d) mathematical statistics.

2.4. Materials

In the process of the corpus-based teaching idioms’ interpretation to EFL students open-access corpora have been used. That is a group of linguistic corpora, created by Mark Davies, Professor of Brigham Young University (BYU), which comprises the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), the Time Magazine Corpus (TMC), the Corpus of Historical American English (COHA).

3. Results and Discussion

3.1. The results of the ascertaining experiment

A questionnaire aimed at finding out what ICT are used by the students in the process of idioms’ interpretation showed that the vast majority of them (89.3 %) use bilingual online dictionaries, among which the most popular were Google and Yandex translators (63.1%), also students note some dictionaries ABBY Lingvo, Multitrans (26.2 %). Along with bilingual dictionaries some students (5.9 %) mention monolingual dictionaries Merriam-Webster Dictionary, Cambridge Dictionaries Online, Oxford English Dictionary. A small number (2.8 %) of students use specialized dictionaries (dictionaries of terms, phraseological units, synonyms, etc.) and encyclopedias British encyclopedia,
Wikipedia (2%). It should be noted that the resources of corpus linguistics have not been mentioned by the students among the ICT used in the process of idioms’ interpretation. Moreover, 94.8% of the student could not even give a single example of any linguistic corpus.

The results of the questionnaire are presented in Figure 1.

![Figure 1. The result of the questionnaire.](image)

The second task of the ascertain experiment is to evaluate the level of English proficiency and knowledge of English idioms. For this reason Entry English test (created by University faculty) was offered, the format of which is similar to PET (Preliminary English Test). Test results illustrated that the vast majority of the students (60%) have reached B1 (Pre-Intermediate), 25% - B2 (Intermediate), 12% - A2 (Elementary), 3% - B2 +/- C1 (Upper-Intermediate).

The third task is to estimate the proficiency in translation techniques. The discussion of the students’ variants of idioms’ interpretation gives us a golden opportunity to discover a number of difficulties the students experience in the process of idiom’s interpretation:

- understanding the dominant, peripheral, and polisemous meanings of a word;
- using imagination and creativity;
- finding out the communicative intention of the text;
- activating knowledge relating to the different kinds of discourse;
- choosing an adequate equivalent for an idiom;
- understanding the figurative uses of a word and the relationship between the literal meaning and the figurative meaning etc.

Taken together, the observations of the results of the ascertain experiment lead us to some inferences (a) it is necessary to introduce corpus linguistic technology as a tool to interpret idioms, (b) 88% of students have sufficient level of English proficiency to work with figurative and idiomatic language, (c) the majority of students deal with the problems in the acquisition of the necessary skills of adequate idioms’ interpretation.

3.2. The results of the educational experiment

The educational experiment comprises a set of corpus-based activities which falls into two related modules. Module 1 (pre-translation stage) is aimed at developing key skills how to work with some basic corpus tools: LIST, KWIC, CHART, COMPARE, COLLOCATE in the process of idioms’ interpretation. Module 2 (translation stage) is aimed at developing CBIIS in the process of text translations.

Let us describe the part of the educational experiment with Module 1.

While working with the tool LIST of the Corpus of Contemporary American English (COCA), students were asked to find the most frequent collocates to the verb break and find out which ones are idiomatic.
75% of students completed this task successfully having made the conclusion that *break the law, break the cycle, break the news, break the rules, break the ice* are the most frequent idioms with the word *break* according to COCA (Figure 2).

15% of the students were unable to come to the correct conclusions from the received corpus data, having selected *break the ice* as the only idiom in the list. The remaining 10% made some mistakes in the query, therefore failed the exercise.

Furthermore, students were asked to choose an unfamiliar idiom with the word *break* and make a list of its contextual use. After studying the contexts, students had to guess the meaning of the chosen idiom. Most students (65%) chose the idiom *break the ice* for the contextual analysis and got the following result (Figure 3).
From the corpus data it has been concluded that the idiom *break the ice* has both literal (5th example) and figurative meaning. All the students (100%) managed to guess the idiomatic meaning (say something to relieve tension; get conversation going when strangers meet) and interpret it properly (make the first step / сделать первый шаг, melt the ice / растопить лед, break the silence / нарушить молчание).

While working with the tool CHART students were asked to choose one of the BYU corpora – COCA, COHA, BNC, TMC – to make the query on the idioms *golden handshake*, *golden hello*, *golden parachute* and study them in terms of genres, styles, years. Furthermore, the results were supposed to be discussed and compared in small groups. The obtained corpus data is presented in Figures 4-7.

![Figure 4. The query golden parachute. TMC.](image-url)
Figure 5. The query *golden handshake*. COCA.

Figure 6. The query *golden parachute*. COHA.
Figure 7. The query *golden hello*. BNC.

Taken together the results of the queries in different corpora students made the following inferences: (a) all the idioms under consideration *golden handshake, golden hello, golden parachute* function primarily in the public discourse; (b) the peak usage accounted for 80-90 years of the 19th century; (c) BNC corpus presents information only about the genres, TMC – only about the years, COCA, COHA – both genres and years; (d) COCA corpus turns out to be the most representative one because it contains information about the genres and the years alike and has the largest number of the tokens.

While working with the tool **KWIC**, the students were asked to build a concordance of the idiom *bang for the buck*. With the help of the vertical reading the students were supposed to analyze the right and left contexts and discuss lexical and grammatical compatibility of the idiom (Figure 8).
The students paid their attention that the verbs get/be are mostly used prior to the bang buck, prepositions in/for/from go after the idiom.

All in all, the corpus-based activities of Module 1 contribute to the development of linguistic and creative thinking, contextual guess, compensatory skill, abilities to deduce the meaning of unknown idiom and to identify its lexical and grammatical compatibility.

Module 2 (translation stage) represents a set of short texts containing idioms. Students were asked to interpret the texts, paying their attention on idioms. Having acquired some key skills of working with linguistic corpora in Module 1, students were supposed to use the obtained knowledge while interpreting the idioms and process the algorithm of the CBII.

Let us describe the part of the educational experiment when we work with Module 2 activities. Students were asked to interpret the following extract, drawing attention on a new frontier.

Around 1994, a new frontier called the internet, was first being made available to the general public. In actuality, a primitive form of the internet had been around since 1969.

After studying the obtained results, we may derive two most common ways of idioms’ interpretation used by students who did not work with corpus linguistic technologies in advance.

1. Loan-translation
   ‘Приблизительно в 1994 году, новый рубеж, называемый интернетом, стал впервые доступен широкому кругу общественности’.

2. Elimination of idiomatic meaning – students made their mind to skip the idiom interpretation.
   ‘Около 1994 года интернет стал впервые доступен широкому кругу общественности. Приблизительно в 1994 году, общество вступило в эпоху Интернета’.

Most of the students who completed Module 1 activities and were familiar with corpus linguistic technology got different results.

1. ‘Приблизительно в 1994 году интернет стал доступен широкой общественности. Так был взят новый рубеж в развитии науки и техники’.

2. ‘Приблизительно в 1994 году интернет стал доступен широкому кругу людей. Это был новый рубеж в технологической модернизации общества’.

Here is the algorithm the students followed:

1) use KWIC tool; make the concordance of the word frontier to check its lexical and grammatical compatibility; outcome – new frontier is fixed expression;

2) use CHART; make the diagram of frequency; outcome – the idiom was significantly popular in 1960s; why?
3) use LIST; click on 242 and study the expanded context; outcome – new frontier is often used along with the name of John F. Kennedy in one sentence; why?
4) use on-line encyclopedia; outcome – the term new frontier refers to the economic and social programs of the presidency of John F. Kennedy;
5) it is getting clear that idiom new frontier is an allusion to the electoral program for the social reform of the United States, ‘New frontiers’, which was proposed by John F. Kennedy in the 1960 presidential election. This program included, among other things, the technological modernization of the country.

As a result, we can easily see that the quality of the idiom interpretation and comprehension is better of those students who know how to deal with linguistic corpora and their basic tools.

3.3. The results of the control experiment
The results of the level of the CBIIS proficiency before the educational experiment (pre-experimental stage) and after the experiment (post-experimental stage) are presented in the table 2 below.

Table 2. The results of the level of the CBIIS proficiency.

|                      | Low level | Weak level | Middle level | Communicative limited level | Communicative sufficient level | Proficient level |
|----------------------|-----------|------------|--------------|-----------------------------|--------------------------------|------------------|
| pre-experimental stage | 0.0%      | 28.9%      | 44.7%        | 21.1%                       | 5.3%                           | 0.0%             |
| post-experimental stage | 0.0%      | 7.9%       | 15.8%        | 28.9%                       | 39.5%                          | 7.9%             |

As can be seen from the above data, before the educational experiment most students had middle and weak levels of the CBIIS proficiency, which confirmed the need for targeted work on the developing idioms’ interpretation skill. At the post-experimental stage the level of the CBIIS proficiency has increased significantly. Most of the students achieved the communicative sufficient level, and some students initially having communicative sufficient level, were able to achieve the highest proficient level.

To assess the dynamics of changes of the level of the CBIIS proficiency the average percentage was calculated. The results are presented on the bar chart below (Figure 10).
4. Conclusions

In view of the aforesaid the following results have been recapped: (a) the ascertaining experiment confirmed the necessity of the use of corpus linguistic technology as a tool to interpret idioms, because the majority of students have sufficient level of English proficiency to work with idiomatic language but the lack of knowledge of corpus linguistic technology leads to the problems in the acquisition of the necessary skills of adequate idioms’ interpretation; (b) before the educational experiment most students had middle and weak levels of the CBIIS proficiency; (c) the designed methodology comprises module 1 (pre-translation stage) aimed at developing key corpus skills in the process of idioms’ interpretation and module 2 (translation stage) aimed at developing CBIIS in the process of text translations; (d) after the educational experiment most of the students achieved the communicative sufficient level.

From the bar chart (Figure 10) we may conclude that during the educational experiment the average level of the CBIIS proficiency has increased by 16.5%, which allows us to speak about the effectiveness of the corpus linguistic technology as a tool in teaching idioms’ interpretation to EFL students.

While experiential learning all the students (100%) learned how to use linguistic corpora in order to interpret idioms. Linguistic corpora were recognized by the overwhelming majority (97.3 %) as an interesting and effective tool in learning FFL. 92.7 % of students are planning to use corpus tools in translation activities along with other electronic resources.

Thus, as a result of the educational experiment, the hypothesis of research on the effectiveness of the corpus linguistic technology as a tool in teaching idioms’ interpretation to EFL students was confirmed, provided methodically competent use of corpus linguistics resources in the learning process.

References

[1] Schmidtke D and Kuperman V 2016 Mass counts in World Englishes: A corpus linguistic study of noun countability in non-native varieties of English Corpus Linguistics and Linguistic Theory 13 135–64
[2] López-Couso M J 2016 Continuing the dialogue between corpus linguistics and grammaticalization theory: Three case studies Corpus Linguist. Linguist. Theory 12 7–29
[3] Iruskieta M, Diaz De Ilarraza A and Lersundi M 2015 Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque Corpus Linguist. Linguist. Theory 11 303–34

[4] Weisser M 2016 DART – The dialogue annotation and research tool Corpus Linguist. Linguist. Theory 12 355–88

[5] Koplenig A 2015 Using the parameters of the Zipf-Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis Corpus Linguist. Linguist. Theory 14 1–34

[6] Perek F 2016 Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis Corpus Linguist. Linguist. Theory 14 65–97

[7] Baker P 2012 Acceptable bias? Using corpus linguistics methods with critical discourse analysis Crit. Discourse Stud. 9 247–56

[8] Brindle A 2016 A corpus analysis of discursive constructions of the Sunflower Student Movement in the English-language Taiwanese press Discourse Soc. 27 3–19

[9] Beliaeva L and Chernyavskaya V 2016 Evidence-based linguistics: methods in cognitive paradigm Vopr. Kognitivnoy Lingvistiki 3 77–84

[10] Chernyavskaya V 2019 Corpus-Assisted Discourse Analysis of Russian University 3.0 Identity Tomsk State University Journal of Philology 58 97–114

[11] Kim K H 2014 Examining US news media discourses about North Korea: A corpus-based critical discourse analysis Discourse and Society 25 221–244

[12] Samaie M and Malmir B 2017 US news media portrayal of Islam and Muslims: a corpus-assisted Critical Discourse Analysis. Educational Philosophy and Theory 1 1–16

[13] Seretan V and Wehrli E 2013 Syntactic concordancing and multi-word expression detection Int. J. Data Mining, Model. Manag. 5 158–81

[14] Ackerley K 2017 Effects of corpus-based instruction on phraseology in learner English Lang. Learn. Technol. 21 195–216

[15] Kogan M, Gavriloa A and Nesterov S 2018 Training Engineering Students for Understanding Special Subjects in English: The Role of the Online Component in the Experimental ESP Course 4th Int. Conf. Inf. Technol. Eng. Educ. Inforino 2018 1–6

[16] Kogan M 2016 Ways of integrating approaches of corpus linguistics into the translators training programme 7th Int. Conf. Translation. Language. Culture. 215–19

[17] Pavlovskaia I and Gorina O 2017 Corpus-based cognitive oriented methods of lexical analysis in foreign language studies Cherеповец State Univ. Bull. 1 132–38

[18] Rodríguez-Fuentes R 2015 Review of corpus linguistics and linguistically annotated corpora Lang. Learn. Tech. 19 56–60

[19] Derybina I 2012 Control verbs studying based on corpus linguistics as prdagogical problem Bulletin of Tambov State University 10 154–8

[20] Kokoreva A 2013 Parallel corpus in foreign language teaching Bulletin of Tambov State University 2 57–62

[21] Kuzminykh I and Khoroshilova S 2017 Investigating the impact of corpus-based classroom activities in English phonetics classes on students’ academic progress Bulletin of Novosibirsk State Pedagogical University 7 40–51

[22] Sosnina E 2017 Research of pupils’ linguistic errors on the base of learner corpus Izvestiya of the Samara Russian Academy of Sciences Scientific Center. Social, Humanitarian, Medicobiological Sciences 19 39–44

[23] Sysoyev P and Evstigneev M 2014 Foreign language teachers' competency and competence in using information and communication technologies Procedia -Social and Behavioral Sciences 154 82–6

[24] Tarnaeva L and Osipova E 2016 Corpus linguistics resources use in training translators in the sphere of professional communication Philological Science. The theory and practice 9 205–9