The European Proceedings of
Social & Behavioural Sciences
EpSBS

Future Academy ISSN: 2357-1330

https://dx.doi.org/10.15405/epsbs.2018.12.02.30

18th PCSF 2018
Professional Culture of the Specialist of the Future

CORPUS LINGUISTICS TECHNOLOGY IN TEACHING ENGLISH AS A FOREIGN LANGUAGE

Ekaterina Sergeevna Osipova (a)*
*Corresponding author

(a) Peter the Great St. Petersburg Polytechnic University (SPbPU), Polytechnicheskaya 29, Saint Petersburg, 195251 Russia, kitty_novgorod@mail.ru, 89522705168

Abstract

The article considers corpus linguistics technology from the linguodidactic perspective. It focuses on a set of corpora, which might be used in English as a Foreign Language (EFL) teaching: the British National Corpus, the Corpus of Contemporary American English, the Times Magazine Corpus, the Corpus of Historical American English. The paper presents a definition of the instrumental competence as a coherent cluster of knowledge, skills and abilities necessary for the students to use corpus linguistics technology in foreign language learning. Students are supposed to learn how to work with some basic corpus tools: LIST, KWIC, CHART, COMPARE, COLLOCATE. 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. Using corpus linguistics technology in EFL teaching develops students` critical attitude to English language and facilitates deeper understanding of some linguistic phenomena.

Keywords: BYU corpora, concordance, corpus technology, EFL teaching, instrumental competence.

http://dx.doi.org/10.15405/epsbs.2018.12.02.30
Corresponding Author: Ekaterina Sergeevna Osipova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

Corpus technology has recently presented great potential for plenty of linguistic studies, which fall into some groups: (a) studies on corpus design and methodology (Iruskieta, Diaz de Ilarraza, & Lersundi, 2015; Weisser, 2016), (b) quantitative studies of English lexical and grammatical features (Schmidtke & Kuperman, 2016; López-Couso, 2015), (c) corpus-based studies on linguistic changes (lexical, syntactical and stylistic) (Koplenig, 2018; Perek, 2018), (d) corpus-based discourse analysis (Chernyavskaya, 2017, 2018) .

Some papers have a proper look at the tools of corpus linguistics, for example, concordancer or KWIC (key words in context) and how to use them to solve linguistic issues. For example, Seretan and Wehrli (2013, p. 158) present “an enhanced type of concordancer that integrates syntactic information on sentence structure as well as statistical information on word co-occurrence”.

As we can see above, most of papers are devoted to using corpus technology to solve linguistic issues.

However, we would like to look at corpus linguistics from another angle and focus on linguodidactic potential of corpus technology and its tools.

The advantages of corpus technology for language learning and second language learning are widely presented in the literature (Ackerley, 2017; Kogan, 2016; Leńko-Szymańska, 2017; Pavlovskaya & Gorina, 2017; Rodríguez-Fuentes, 2015; Tarnaeva & Osipova, 2016).

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 (Derybina, 2012; Kokoreva, 2013; Kuzminykh & Khoroshilova, 2017; Sosnina, 2017; Sysoyev & Evstigneev, 2014).

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 the students` vocabulary (Tarnaeva & Osipova, 2016).

2. Problem Statement

“Today’s challenging economic situation means that it is no longer sufficient for a new graduate to have knowledge of an academic subject; increasingly it is necessary for students to gain those skills which will enhance their prospects of employment” (Fallows & Steven, 2000, p.75). Taking this fact into consideration universities “have established an initiative to ensure that each of its students engages with these skills and has embedded this within the academic curriculum for all disciplines” (ibid).

The importance of humanitarian subjects for successful employment is in particular highlight. Krepkaia & Mamleeva (2015) conclude that the strong points of humanitarian subjects should be taken into consideration as emotional intelligence, communicative and social competencies developed by them help students to integrate into their professional areas effectively. Almazova & Rubtsova emphasize that liberal arts education is “a key foundation for a personality’s complete development and humanitarian culture, universal humanistic values, personal purport and moral guidelines.” (Almazova & Rubtsova, 2016, p.18).

In the effort to develop some skills that might be helpful for the specialists of the future we offer to use corpus technology in teaching one of the university disciplines – English as a Foreign Language (EFL). EFL is a discipline which is studied by students of all specializations. That is why we can speak about developing universal employability skills, which are necessary for professionals in all spheres. As far as we are concerned, universal employability skills might include the following abilities: independent learning, the retrieval and handling of information (texts, databases, diagrams, schemes, charts), analytical and critical thinking, problem-solving, ability to draw a conclusion etc.

In this article we have a look at corpus linguistics technology, the use of which might contribute to the formation of the universal employability skills.

3. Research Questions

- Which corpora should be used in EFL teaching?
- What competence should be formed to use corpus linguistics technology in EFL studying?
- What knowledge, skills and abilities must students acquire to use corpus technology?
- Which exercises might help students to acquire the appropriate competence?

4. Purpose of the Study

The article is written to prove linguodidactic potential of corpus technology in teaching EFL. In this survey we want to share our personal experience how to use corpus linguistics technology in EFL teaching for university students.

5. Research Methods

Analytic-synthetic methods are used in the search and analysis of literature on the topic; experimental research method is used to identify the linguodidactic potential of corpus technology; modelling method is applied in describing the instrumental competence.

6. Findings

6.1. Linguistic corpora in EFL teaching

In the process of the corpus-based teaching EFL we recommend to use open-access corpora, i.e. freely-available online, or requiring only registration. There is a group of linguistic corpora, created by Mark Davies, Professor of Brigham Young University. This group 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).

The British National Corpus is a monolingual corpus representing British English of the late 20th century. The corpus includes a collection of samples of both written and spoken British English from a

http://dx.doi.org/10.15405/epsbs.2018.12.02.30
Corresponding Author: Ekaterina Sergeevna Osipova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

variety of genres, sources and subject areas. The total number of the collection is about 100 million words (BNC, 2018).

The Corpus of Contemporary American English is the largest and the only freely-available corpus of American English. It contains 220 thousand texts, 520 million words of the period from 1990 to 2017. The corpus illustrates a wide variety genres: SPOKEN, represented by recordings of more than 150 television and radio programs (All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Today Show (NBC), 60 Minutes (CBS), Hannity and Colmes (Fox), Jerry Springer, etc); FICTION includes short stories, novels, and plays from literary magazines for adults and children, film scripts; POPULAR MAGAZINES (Time, Men's Health, Good Housekeeping, Cosmopolitan, Fortune, Christian Century, Sports Illustrated, etc.), NEWSPAPERS (USA Today, New York Times, the Atlanta Journal Constitution, San Francisco Chronicl etc.); ACADEMIC JOURNALS divided into classes according to the classification of the library of Congress (class B – philosophy, psychology, religion, D – world history, K – education, T – technology, etc.) (COCA, 2018).

The Time Magazine Corpus consists of more than 275,000 articles from the TIME magazine archive and contains more than 100 million words of texts written from 1923 to 2006 (TMC, 2018).

The Corpus of Historical American English is a historical corpus of 400 million words, including 100,000 texts from 1810 to 2009. It gives the possibility to trace the changes of the language over decades. The corpus contains texts from the digital library Making of America (MoA), including 267 monographs and over 100,000 articles of the magazines of the 19th century on education, psychology, American history, sociology, religion, science and technology, it also includes texts of different genres from the archives of universal digital library Project Gutenberg, Project Gutenberg (1810-1930) and from the website https://archive.org/ (1810-1900) (COHA, 2018).

All the linguistic corpora, have a similar structure. So, having acquired the skills to work with one corpus, students will be able to use the others.

Before using corpus linguistics technology in teaching EFL it is essential to form an instrumental competence.

6.2. Instrumental competence: a definition and a content

Instrumental competence is a coherent cluster of knowledge, skills and abilities necessary for students to use corpus linguistics technology in foreign language classes to solve linguistic problems.

Knowledge comprises:

- knowledge of the structure and content of the linguistic corpora of Brigham Young University (BYU): the British National Corpus, the Corpus of Contemporary American English (COCA), the Time Magazine Corpus (TMC), the Corpus of Historical American English (COHA);
- knowledge of the specific and distinctive features of the BYU linguistic corpora;
- knowledge of the possibilities of information search in linguistic corpora;
- knowledge of the possibilities of using corpus linguistics technology for the analysis of grammatical and lexical-grammatical phenomena;

Skills comprise:

- the skill to make a query in the search bar correctly;
- the skill to generate a concordance;
- the skill to use a "wide context" tool;
- the skill to generate frequency charts by genre and year;
- the skill to make a list of frequency collocates to a word, a phrase.

Abilities comprise:

- the ability to work effectively with the LIST tool, which allows to find the most frequent collocates to the phrase and make lists of the context use of the desired unit;
- ability to work effectively with the CHART tool, which allows to make diagrams of the frequent use of the studied grammatical or lexical phenomena by genres and years;
- ability to work effectively with the KWIC tool, which allows to generate concordance lists;
- ability to work effectively with the COMPARE tool, which makes it possible to compare the frequency and stability of the use of two words in relation to the third;
- ability to work effectively with the COLLOCATE tool, which allows to find the most frequent phrases with the studied word.

6.3. A set of exercises aimed at the formation of the instrumental competence

6.3.1. The exercise on the formation of the skill to make a query in the search bar correctly

What phrase can NOT be found on the following query in the search bar?

[beat].[v*] * [nn*]

A. Beat the Yankees
B. Beat about the bush
C. Beaten a bill
D. Beaten dawn prices

The correct answer is B.

The part of the query [beat].[v*] means that we are looking for any forms of a verb *beat.* The part of the query * [nn*] means that before a noun only one word is possible. So, the phrase *Beat about the bush* can not be found on the query, because between a verb and a noun there are two words instead of one.

6.3.2. The exercise on the formation of the ability to work with the LIST tool and the skill to generate a concordance.

Refer to the Corpus of Contemporary American English (COCA), sub-corpus Finance and money. Use the LIST tool and generate the concordances with the prepositions *despite*, *in spite of*, and conjunctions *although, while, however, nevertheless.* Comment on the syntactical features of the prepositions and conjunctions.

2006	MAG	Forbes	Sexy as a matching washer and dryer. More important, nbc universal, **in spite of** its current malaise, has the ability to rebound sharply -- it is in
2006	MAG	Inc.	Have traditionally not been all that pro-business; it has managed to do well **in spite of** itself, " says Tom Devlin, founder of the Rent-A-Center retail chain and
2004	MAG	Fortune	The civil-rights generation's heirs to succeed independently in the mainstream-and to do so **in spite of**, not because of, their blackness. Their goal was to build companies
1998	MAG	Money	Three traditional discount brokerages, is also a good site for frenetic traders. **In spite of** its relatively low $7.95 commission, this new york city broker doesn't skimp

Figure 01. The result of the query. Tool LIST.

1990	MAG	Money	Preferred stocks, which pay fixed dividends, and are suitable for income investors. **Although** british banks have some troubled loans, their problems are far less severe than those
1990	MAG	Money	A maze. I never know how i wind up where i wind up, **although** i always wind up getting what i came in for. " the following steps
1990	MAG	Money	Stocks. Bonds posted healthy profits of as much as 6.4%. And gold, **although** a tiny part of the total portfolio, turned in the biggest percentage gain,
1990	MAG	Money	Two-income families climbed from 43.4% to 63%. In the sarneckis' case specifically, **although** tom's salary is more than twice that of the average single-paycheck family, there

Figure 02. The result of the query. Tool LIST.

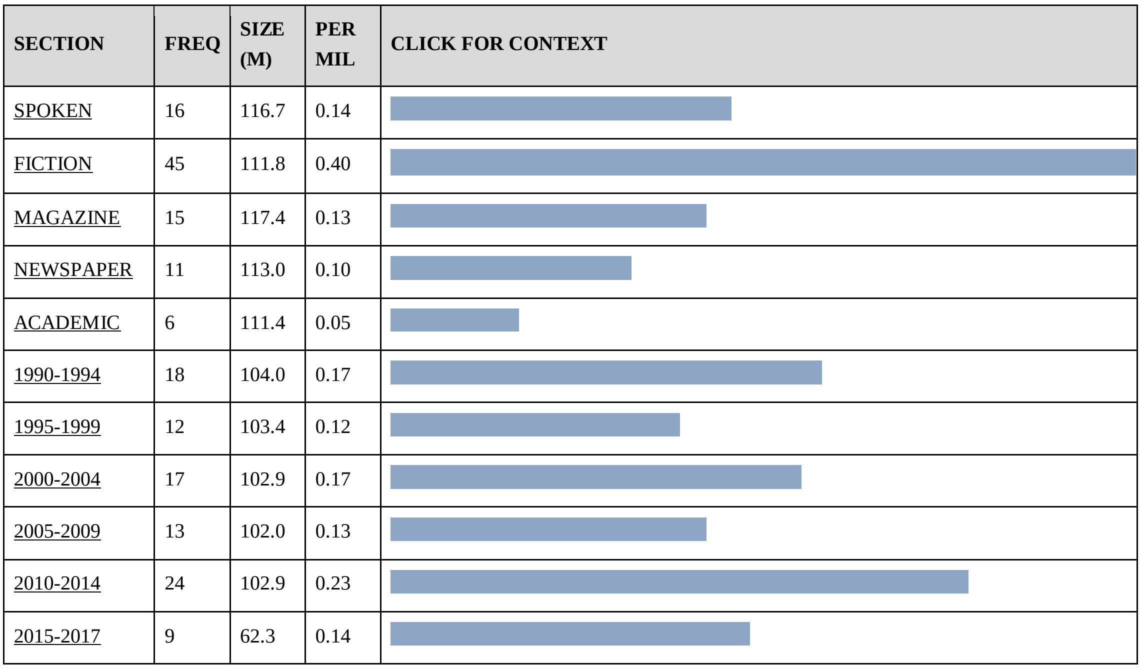
Based on the corpus results (figures 01, 02), the students conclude that after the conjunction *although* clause is used whereas after the preposition *in spite of* noun phrase is used.

6.3.3. The exercise on the formation of the ability to work with the CHART tool.

Make a chart of the frequency by genre and year to the phrase *Far be it from*?

While studying the “formulaic” subjunctive the students ask if the phrase *Far be it from* is still used in contemporary English. We don`t give the answer, but ask the students to do the exercise above.

This is the result they get.



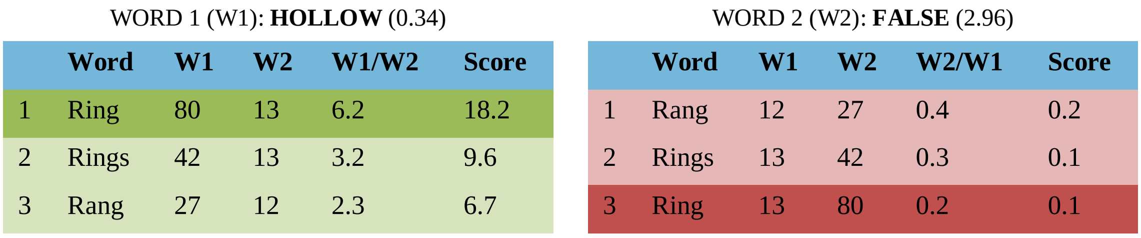
SECTION	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT
SPOKEN	16	116.7	0.14	
FICTION	45	111.8	0.40	
MAGAZINE	15	117.4	0.13	
NEWSPAPER	11	113.0	0.10	
ACADEMIC	6	111.4	0.05	
1990-1994	18	104.0	0.17	
1995-1999	12	103.4	0.12	
2000-2004	17	102.9	0.17	
2005-2009	13	102.0	0.13	
2010-2014	24	102.9	0.23	
2015-2017	9	62.3	0.14	

Figure 03. The result of the query. Tool CHART.

Based on the corpus results, the students come to conclusion that the phrase *Far be it from* is used in contemporary English and is very common in fiction.

6.3.4. The exercise on the formation of the ability to work with the COMPARE tool.

Use the COMPARE tool and choose the most frequent collocate from the following: it rings hollow/ it rings false?



WORD 1 (W1): **HOLLOW** (0.34)

	Word	W1	W2	W1/W2	Score
1	Ring	80	13	6.2	18.2
2	Rings	42	13	3.2	9.6
3	Rang	27	12	2.3	6.7

WORD 2 (W2): **FALSE** (2.96)

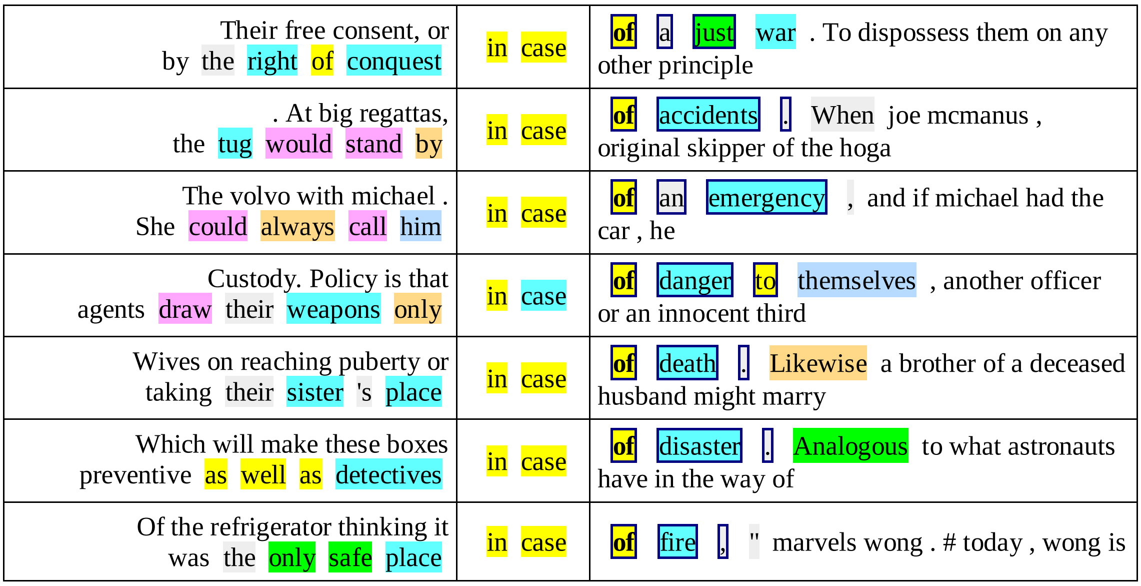
	Word	W1	W2	W2/W1	Score
1	Rang	12	27	0.4	0.2
2	Rings	13	42	0.3	0.1
3	Ring	13	80	0.2	0.1

Figure 04. The result of the query. Tool COMPARE.

From the figure 04 it is obvious, that the most frequent collocate is *it rings falls*.

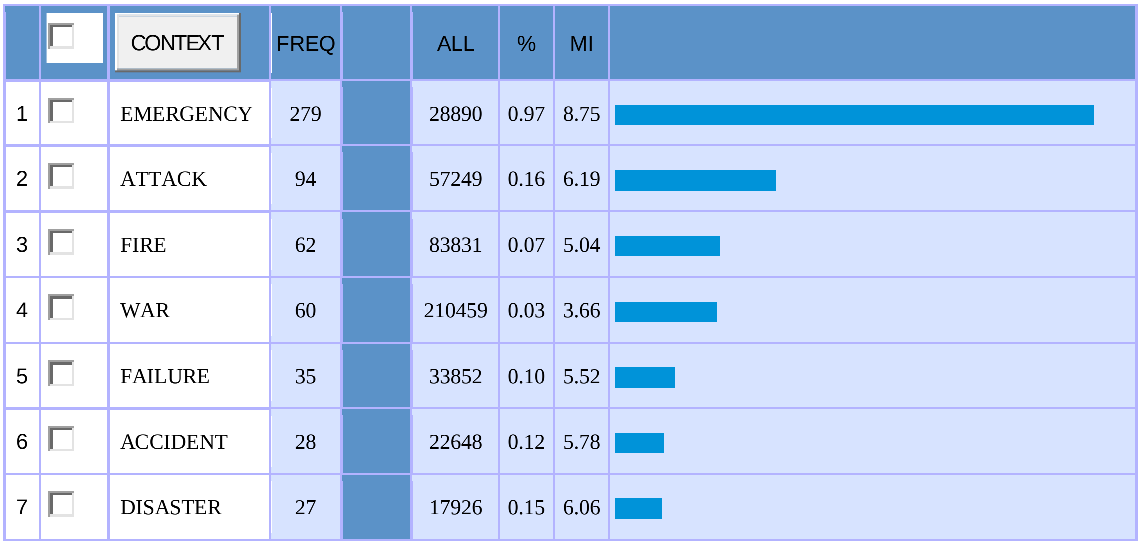
6.3.5. The exercise on the formation of the ability to work with the tools COLLOCATE and KWIC

Use the KWIC tool to make a concordance for the preposition “in case”. What preposition does it (“in case”) follow? Use the COLLOCATE tool and analyze the connotation of the expressions with the preposition “in case”.



Their free consent, or by the right of conquest	in case	of a just war . To dispossess them on any other principle
. At big regattas, the tug would stand by	in case	of accidents . When joe mcmanus , original skipper of the hoga
The volvo with michael . She could always call him	in case	of an emergency , and if michael had the car , he
Custody. Policy is that agents draw their weapons only	in case	of danger to themselves , another officer or an innocent third
Wives on reaching puberty or taking their sister 's place	in case	of death . Likewise a brother of a deceased husband might marry
Which will make these boxes preventive as well as detectives	in case	of disaster . Analogous to what astronauts have in the way of
Of the refrigerator thinking it was the only safe place	in case	of fire , " marvels wong . # today , wong is

Figure 05. The result of the query. Tool KWIC.



	□	CONTEXT	FREQ		ALL	%	MI	
1	□	EMERGENCY	279		28890	0.97	8.75	
2	□	ATTACK	94		57249	0.16	6.19	
3	□	FIRE	62		83831	0.07	5.04	
4	□	WAR	60		210459	0.03	3.66	
5	□	FAILURE	35		33852	0.10	5.52	
6	□	ACCIDENT	28		22648	0.12	5.78	
7	□	DISASTER	27		17926	0.15	6.06	

Figure 06. The result of the query. Tool COLLOCATE.

Based on the figure 05 the students can easily notice that the preposition *of* is used after *in case*. They also pay their attention on the nouns after the preposition *in case of*. They are *war, accident, emergency, danger, disaster, fire*. The students conclude that the connotation of the words after *in case of* is negative. To make sure that it is always true, the students are recommended to use the tool COLLOCATE (figure 06). This tool gives the most frequent collocates to a word or a phrase.

After acquiring the basic skills and abilities to work with corpus linguistics technology we offer our students more difficult linguistic tasks.

Examples of the linguistic tasks.

- Are there any differences in meaning between disinterested or uninterested? Are they absolute synonyms?
- Are there any differences in using disinterested / uninterested in context?

To solve the tasks the students are supposed to choose the appropriate corpus linguistics tools themselves and find the way to carry out the corpus search. As for the linguistic tasks given above, the students should use the tool KWIC.

I should have been the good daughter , studious and	uninterested	in boyfriends . But not at all . It was made clear
beat up the boys in infant school ! I was totally	uninterested	in boys , apart from playing football with them . up to
morale . Conscious that British audiences were almost wholly	uninterested	in British films , some members of the industry increased their
because it has little relevance but also to expose Labour as	uninterested	in cross-party co-operation . He has little sympathy with those

Figure 07. The result of the query. Tool KWIC.

a Solicitor # A solicitor 's ability to give impartial and	disinterested	**advice** is a fundamental element of his or her relationship with
because they want to believe it , that he is offering	disinterested	**advice** , as best befits their needs . The skilful dealer does
# Natural experts ? # The includers are not always merely	disinterested	**advocates** of a philosophy but people who have become so through
be the concern of one investigator , possibly with a senior	disinterested	**air** traffic controller working in consultation . The flight
discuss ways of deliberately cultivating and nurturing pure ,	disinterested	**altruism** -- something that has no place in nature , something

Figure 08. The result of the query. Tool KWIC.

Based on the corpus results, the students realize that disinterested and uninterested are not the synonyms. *Uninterested* means not interested in or concerned about something or someone whereas *disinterested* means not influenced by considerations of personal advantage. *Uninterested* is usually followed by the preposition *in* whereas *disinterested* is followed by a noun.

It is important to emphasize that all the presented above exercises are based on an inductive approach to teaching EFL. It means that some linguistic phenomena are presented to students in a real language context. Such exercises are designed both to engage the students` interest, and to raise students` sensitivity to linguistic features. It is valuable that students can learn to carry out their own investigations.

7. Conclusion

We absolutely agree with the opinion, that the methodology of corpus linguistics is congenial for students of all levels because it is a "bottom-up" study of the language requiring very little learned expertise to start with. Even the students that come to linguistic enquiry without a theoretical apparatus learn very quickly to advance their hypotheses on the basis of their observations rather than received knowledge and test them against the evidence provided by the corpus (Togini-Bonelli, 2001).

In conclusion, we would like to highlight that the use of corpus linguistics technology in teaching EFL contributes to (a) the development of students ' critical attitude to some linguistic phenomena, (b) the formation of deeper understanding and meaningful use of lexical and grammatical structures, (c) the development of linguistic guess, (d) the correct use of English language according to the genre.

References

Almazova, N., & Rubtsova, A. (2016). The role of liberal arts education in the system of vocational training of engineers. *The Scientific Opinion, 1(2)*, 18-27.

Ackerley, K. (2017). Effects of corpus-based instruction on phraseology in learner English. *Language Learning & Technology, 21*(3), 195-216. doi:10125/44627

BNC (2018). British National Corpus. Retrieved from http://www.natcorp.ox.ac.uk/

Chernyavskaya, V. (2017). Towards methodological application of discourse analysis in corpus-driven linguistics. *Tomsk State University Journal of Philology, 50*, 135-148. doi: 10.17223/19986645/50/9

Chernyavskaya, V. (2018). Discourse analysis and corpus approaches: a missing evidence-based link? Towards qualitative and quantitative approaches in language studies. *Issues of Cognitive Linguistics, 2(55)*, 31-37. doi:10.20916/1812-3228-2018-2-31-37

COCA (2018). Corpus of Contemporary American English. Retrieved from http://corpus.byu.edu/COCA/

COHA (2018). Corpus of Historical American English. Retrieved from http://corpus.byu.edu/COHA/

Derybina, I. (2012). Control verbs studying based on corpus linguistics as prdagogical problem. *Bulletin of Tambov State University named after G.R. Derzhavin, 10* (114), 154-158.

Iruskieta, M., Diaz de Ilarraza, A., & Lersundi, M. (2015). Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory, 11*(2), 303-334. doi:10.1515/cllt-2013-0008

Fallows, S., & Steven, C. (2000). Building employability skills into the higher education curriculum: a university-wide initiative. *Education + Training, 42* (2), 75-83. doi:10.1108/00400910010331620

Kogan, M. (2016). Ways of integrating approaches of corpus linguistics into the translators training programme. *Translation. Language. Culture.* 215-219.

Kokoreva, A. (2013). Parallel corpus in foreign language teaching. *Bulletin of Tambov State University named after G.R. Derzhavin, 2* (118), 57-62.

Koplenig, A. (2018). Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis. *Corpus Linguistics and Linguistic Theory, 14* (1), 1-34. doi: 10.1515/cllt-2014-0049.

Krepkaia, T., & Mamleeva, A. (2015). The importance of humanitarian subjects for successful engineering graduates` employment. *Teaching Methodology in Higher Education, 4 (18)*, 83-91.

Kuzminykh, I., & 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* (4), 40-51. doi: 10.15293/2226-3365.1704.03

Leńko-Szymańska, A. (2017). Training teachers in data driven learning: Tackling the challenge. *Language Learning & Technology, 21* (3), 217-241. doi: 10125/44628

López-Couso, M. (2015). Continuing the dialogue between corpus linguistics and grammaticalization theory: Three case studies. *Corpus Linguistics and Linguistic Theory, 12*(1), 7-29. doi:10.1515/cllt-2015-0069

Pavlovskaya, I., & Gorina, O. (2017). Corpus-based cognitive oriented methos of lexical analysis in foreign language studies. *Bulletin of Cherepovets State University,1* (76), 132-138. doi: 10.23859/1994-0637-2017-1-76-19

Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 65-97. doi:10.1515/cllt-2016-0014

Rodríguez-Fuentes, R. (2015). Review of Corpus Linguistics and Linguistically Annotated Corpora. *Language Learning & Technology, 19*(3), 56-60. doi:10125/44431

Schmidtke, D., & 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*(1), 135-164. doi:10.1515/cllt-2015-0047

Seretan, V., & Wehrli, E. (2013). Syntactic concordancing and multi-word expression detection. *International Journal of Data Mining, Modelling and Managemen, 5*(2), 158-181. doi: 10.1504/IJDMMM.2013.053694

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* (5), 39-44.

Sysoyev, P., & Evstigneev, M. (2014). Foreign language teachers' competency and competence in using information and communication technologies. *Procedia -Social and Behavioral Sciences, 154,* 82-86. doi: 10.1016/j.sbspro.2014.10.116

Tarnaeva, L., & Osipova, E. (2016). Corpus linguistics resources use in training translators in the sphere of professional communication. *Philological Science. The theory and practice, 9-1* (63), 205-209. Retrieved from http: // www.gramota.net/materials/2/2016/9-1/

TMC. (2018). Time Magazine Corpus. Retrieved from http://corpus.byu.edu/time/

Togini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam, The Netherlands: John Benjamins Publishing Company. https://doi.org/10.1075/scl.6

Weisser, M. (2016). DART – The dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory, 12*(2), 355-388. doi:10.1515/cllt-2014-0051