Using Google to Search Language Patterns in Web-Corpus: EFL Writing Pedagogy

https://doi.org/10.3991/ijet.v13i03.7712

Olga S. Kvashnina(✉), Olga V. Sumtsova
National Research Tomsk Polytechnic University, Tomsk, Russia
cuba@tpu.ru

Abstract—The use of the Web as a corpus of language data and Google as a concordancer has been considered as one of the promising directions in the development of the EFL pedagogy. It has been stated by several researchers that the use of the popular search engine Google to search and define authentic language patterns is mostly advantageous for teaching EFL writing as well as mastering learner autonomy of the EFL students. The present article focuses on the advantages and opportunities of Google as a search engine in the Web-corpus for EFL writing pedagogy in technical universities. The authors compare the results of the current investigation with the outcomes of their previous study aimed at examining the potential of traditional linguistic corpora for EFL classroom. Although there is a strong need for further in-depth study of the considered issue, the comparative results obtained at this stage reveal high potential and remarkable advantage of the global web as a corpus, and Google as a search and access tool to authentic language patterns.

Keywords—web corpus, Google, EFL writing, language patterns, linguistic corpus

1 Introduction

The concept considering the Internet not only as a source of materials for composing a corpus but also as a corpus by itself, has become a problem for discussion since 2001 when the paper “Web as corpus” was published in the journal of Lancaster University [1]. In that paper Kilgariff described more significant connections between the Internet and a linguistic corpus than it was believed at that point in time. Have any changes taken place from then onwards in considering the Internet as a potential linguistic corpus? In his research [2] devoted to the contemporary state of interrelation between the concept of web as a corpus and a traditional text corpus, Mordovin cited Gatto’s words confirming the thesis on in-depth potential of web for corpus-based linguistics proposed by Kilgariff in 2001: “[earlier] in the introductory courses into corpus-based linguistic, web as a corpus was generally described in the final chapters being referred to as “a field for future research”. Now this future has already arrived. In 2013 more than half of the world population has grown up in the environment where web establishes the way of communication, business operations and our lifestyle on the whole…In case we [as before] prefer a newspaper and book corpus to the corpus of blogs and chats regarding them as a peripheral novelty, you and I belong to people of the past…Today the language exists online, it is accessed through online tools and the ways of studying, processing and analyzing the language have become online as well… Google occupies a central position in our life, besides it does not have so many differences from corpus tools. Nowadays probably the first question being posed by a lay person to a corpus linguist concerned with his/her professional activity is - “Why don’t you use Google for all kinds of your work?” [3, p. 24]. However, in contradiction to it, the author of the paper quotes the statement of Sinclair, the father of corpus-based linguistic: “Web is not a corpus because its dimensions are unrevealed as they are constantly changing; furthermore, it was composed without a focus on the linguistic purpose”. Analyzing the existed discussions on the topic “web as corpus” and “corpus as web”, Mordovin comes to the conclusion that “web considerably differs from the corpus in the lack of linguistic conception, spontaneity, growth uncontrollability and specific representativity. Meanwhile an easy access and high data objectiveness in default of composite authors greatly compensate these drawbacks and functionally equate web to the corpus requiring a reinterpretation of the ontological definition of the last one”.

The statement that web as a corpus represents an irreplaceable source of authentic, natural, contextualized language patterns (concordances, established collocations, phrases, idioms etc.), has been proven by many scholars (Comelles et al [4]; Conroy [5]; Geluso [6]; Park [7]; Sha [8]; Yoon [9], Panah et al [10]). From the standpoint of cognitive linguistics, what is perceived as an authentic and natural language, is directly related to phraseology based on usage frequency rooted into the functional method of language acquisition [6]. Thanks to its inexhaustible content, worldwide web becomes a unique resource for analyzing frequency of use of particular language patterns meaning the natural, “living” language. It is evident that both written and oral speech is characterized by a vast number of such stable linguistic combinations and language patterns; therefore, a web corpus represents an irreplaceable tool for EFL students. [4, 5, 6, 8] At present, there is a fairly large number of search engines (Google, Alta Vista, Yahoo, MSN) providing access to the corpus of language data. Google is believed to be the most popular search system due to its ease of use, time rate, search functionality and representativity as well. For example, in linguistic pedagogy, Google as a giant search engine can be recommended as an access tool to data of the world network enabling to analyze usage frequency and meaning of one or another language pattern. [10]

Google is a dynamic “corpus”, which provides access to the great amount of developing and constantly changing Internet resources. For instance, a word-group *perfect balance* is presented 37 times in the British national corpus, 8.020.000 times in the AltaVista search engine and 11.1000.000 times in Google. Thus, Shei declares that “Google is able to offer solutions for many research questions in the field of phraseology whereas corpuses, containing some billions of words, hardly ever solve these problems”. [11] A lot of users make a point that the search engine under study possesses a user-friendly interface and an info search rate, a possibility to check spelling, whilst the majority of the existed concordancer programs cannot display these important characteristics. Moreover, Google can perform a function of the huge dictionary comprising almost a total amount of English vocabulary including definitions, synonyms/antonyms, contexts and visualization.

Google can be realized as a concordancer program for studying frequency and context of using particular language patterns and for applying them in a more natural way. This calls for inputting a target phrase in quotation marks into the search line and study the obtained data. In that context, it is necessary to consider the results of searching such phrases as '*to discuss about the issue*' and '*to discuss the issue*'. In September 2016 a per-second search in Google Scholar gave 191 examples of using the first phrase and 59.700 of the second one which allows accuracy evaluation of each of these versions. We can also consider the examples of using the word-groups '*the study concentrates on*' (7.700 occurrences in Google Scholar) and '*the study focuses on*' (49.300 occurrences in Google Scholar) which are believed to be absolutely correct in terms of lexis and grammar. In reliance on frequency of using these phrases we can make a conclusion that a native speaker will definitely choose the last word-group as it seems more natural.

The benefits of Google are available to both ESL students and teachers.

For teachers, the tool under study is of particular assistance in picking over relevant language patterns from the perspective of frequency and contextuality of using them within the scope of determining a content of the education program of the discipline "Foreign Language", working out tasks on testing and drilling lexical and grammatical patterns based on authentic examples of using English, mastering learning autonomy of the EFL students by means of Google and web corpus.

In spite of the emerging opportunities of applying web corpus in acquiring English, only a few scientific papers focused on the study of practical experience of using Web as a language data corpus in the learning process are known. The authors of the present article attempted to study corresponding experience of using Google in the process of mastering written speech by the students of a nonlinguistic university in the course of the discipline "Foreign Language" as well as to compare this experience with the results of training outcome via the use of traditional linguistic corpora obtained before.

2 Experiment description

Two groups (experimental and control) of the second-year students of the Institute of High Technology Physics, National Research Tomsk Polytechnic University studying English as a Foreign Language, were chosen to investigate the potential of using the worldwide network as a linguistic corpus in teaching English. Students' level of English proficiency mainly referred to Intermediate (B1 according to Common European Framework of Reference for Languages). During the second year of studying the discipline "FL" within the framework of mastering written speech (formal letter, curriculum vitae and covering latter) the students of the experimental group (12 learners) were offered to use Google to search for, determine and select natural language patterns in the process of improving skills and developing a written speech competence. The experiment was carried out in the 2015-2016 academic year and consisted of four stages:

Stage 1. A preliminary questionnaire of the students aimed to examine their experience of using Google with the purpose of revealing general information or in the process of autonomous learning of English. The questionnaire was conducted in classroom environment and included several tasks on searching and analyzing information by means of the system under consideration along with the open-ended and closed questions. A questionnaire survey indicated a high level of students' awareness of Google opportunities and tools as well as sufficient experience of working with the given system when searching for information not only in Russian but also in English. Nevertheless, it emerged that the students had never used Google to analyze language data when studying the English language.

Stage 2. The students of the experimental group were given the tasks focused on the search and identification of the language patterns with the aid of Google. The tasks were initially performed in classroom environment in the process of collaborative work under the guidance and with a direct involvement of the teacher; then individually, in pairs or micro groups; and further, in the scope of self-study, individually or in groups (including the online delivery platform Moodle).

Examples of the tasks:

- using Google for the search of the appropriate word to fill in the gap;
- using Google for the selection of the contextually determined word;
- selecting the most frequently used adjectives that go with the given noun;
- considering the use of two various prepositions in the given word collocation;
- selecting a grammatical structure on the base of the usage frequency (for example, *is used to determine* or *is used for determining*);
- using Google for identification and correction of mistakes in a small English text as a result of the language data analysis.

Stage 3. The students performed various written assignments (essay, abstract writing, and etc.) with the use of Google for language patterns analysis and selection.

Stage 4. The final stage included the discussion of the outcomes and impressions obtained by the students of the experimental group who used Google in the process of classroom and independent learning of the English language.

3 Discussion

The outcomes of the written assignments performed by the students of the experimental group need further thorough analysis. Nevertheless, the comparative results of the control written works carried out by the students of both groups with regard to language purity and relevance to the written speech of native speakers, allow the following conclusions: language purity in the experimental group was at average 55% higher than in the control group (correctness of using lexical items - 51% higher, grammatical structures - 59% higher). The comparison results of the works in the context of relevance to the speech patterns of native speakers need qualitative analysis rather than quantitative one as well as additional processing.

In the course of oral conversation with the students of the experimental group, the following advantages of using Google with the aim of developing written speech skills in English were distinguished:

1. the search engine Google is easy-to-use; a particular practice in applying it as a search engine for a language data corpus is not required;
2. the search engine is free of charge and generally accessible;
3. Google searching allows the analysis of correctness and appropriateness of using one or another language unit or pattern rather quickly raising confidence in outcome accuracy, thus it enables students to push their written speech closer to the written speech of native speakers;
4. Google searching raises students' awareness of using words/language patterns in the context;
5. the use of Google broadens comprehension of the functions of language units taking into consideration not only their form but also their meaning;
6. the work with a web-corpus by means of Google allows shifting the emphasis in favor of the autonomous cognitive language study;

Nonetheless, the following challenges were noticed in the process of conducting the present experiment: (a) not all the students equally coped with the analysis of obtained data generally because of insufficient level of English proficiency (below B1); (b) the process of completing the written works sometimes took a long time whereas some students had to analyze bulk information; (c) some students were not motivated enough to use a research approach to mastering English.

Drawing a parallel between the present experiment and the investigation related to using a language corpus in the process of teaching written speech in English pursued by the authors of the article in 2007 [12], the comparison table (Table 1) can be composed.

Table 1. Results comparison

Criterion	Google	Linguistic corpus
a need to train students	special training is not required	special training is required with the aim of raising awareness of working with search mechanisms
accessibility	generally accessible free resource	free online corpus Collins COBUILD Corpus was used, however, it was regularly inaccessible due to technical difficulties
interface	user-friendly interface for most students	quite difficult interface, special engine marking of parts of speech etc.
authenticity	a great number of authentic texts but not always essential in terms of the language and speech standards	a limited number of selected authentic texts
size	hundreds of billions of words	524 mil. of words (2007)
students' impressions of using	positive	55% - positive, 45% of the students mentioned work complexity of the resource

Having analyzed the obtained data, it is possible to come to the conclusion that web as a corpus can be reasonably used (through the search engine Google in particular) for teaching English written speech as a source of obtaining information about natural, authentic and widely-used language patterns as well as for the purposes of improving learning autonomy skills of the EFL students.

4 Conclusion

Despite the need for further careful examination of the issue related to using Google and a Web-corpus in general as a source of infinite number of authentic patterns of the "living" language for the purpose of improving the English teaching outcomes, the presented intermediate results allow us to talk about a significant potential and advantages of using Google in teaching English writing to technical students. A Web-corpus as a generally accessible free resource meets the basic requirements applicable to the linguistic corpus: authenticity, representativeness and size. Google as one of the most popular search engines represents an access and search mechanism in the given Web-corpus which does not require any special training of students for using it as a tool of mastering English. In spite of the published results of some studies in this field, there is a necessity for further investigation of the question referred to search and analysis of natural language patterns in Web-corpus within EFL pedagogy.

5 References

[1] Kilgariff, A. (2001) Web as corpus. Proceedings of Corpus Linguistics 2001 conference, Lancaster University, 342-344

[2] Mordovin, A.Yu. (2015) "Web as corpus" or "corpus as web": new reality of corpus linguistics. Vestnik MSLU, 3 (714), 163-172 http://cyberleninka.ru/article/n/veb-kak-korpus-ili-korpus-kak-veb-novaya-realnost-korpusnoy-lingvistiki

[3] Gatto, M. (2014) The Web as Corpus: theory and practice

[4] Comelles, E., Laso, J. N., Forcadell, M., et al (2012) Using online databases in the linguistics classroom: dealing with clause patterns. Computer Assisted Language Learning, 1-13

[5] Conroy, M. A. (2010) Internet tools for language learning: University students taking control of their writing. Australasian Journal of Educational Technology, 26 (6), 861-882 https://doi.org/10.14742/ajet.1047

[6] Geluso, J. (2013) Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing. Computer Assisted Language Learning, 26:2, 144-157 https://doi.org/10.1080/09588221.2011.639786

[7] Park, K. & Kinginger, C. (2010) Writing/thinking in real time: Digital video and corpus query analysis. Language Learning & Technology*,* 14 (3), 31-50

[8] Sha, G. (2010) Using Google as a super corpus to drive written language learning: A comparison with the British National Corpus. Computer Assisted Language Learning, 23, 377-393 https://doi.org/10.1080/09588221.2010.514576

[9] Yoon, C. (2011) Concordancing in L2 writing class: An overview of research and issues. Journal of English for Academic Purposes, 10 (3), 130-139 https://doi.org/10.1016/j.jeap.2011.03.003

[10] Panah, E., Yunus, M. & Embi, M. A. (2013) Google-informed patter-hunting and pattern-defining: Implication for language pedagogy. Asian Social Science, 9(3), 229-238 https://doi.org/10.5539/ass.v9n3p229

[11] Shei, C. C. (2008) Discovering the hidden treasure on the internet: Using Google to uncover the veil of phraseology. Computer Assisted Language Learning, 21, 67-85 https://doi.org/10.1080/09588220701865516

[12] Kvashnina, O.S. (2007) Mastering EFL academic writing pedagogy in non-linguistic universities: corpus technologies. Yazyki v sovremennom mire: Proceedings of VI International conference, Moscow, 219-226.

6 Acknowledgment

The research is funded from Tomsk Polytechnic University Competitiveness Enhancement Program grant, Project Number TPU CEP_PTI_72\2017.

7 Authors

Olga S. Kvashnina is a head teacher at the Department of Foreign Languages of the Institute of Physics and Technology, National Research Tomsk Polytechnic University (Lenin Ave, 30, 634050, Tomsk, Russia), email: cuba@tpu.ru.

Olga V. Sumtsova is a head teacher at the Department of Foreign Languages of the Institute of Physics and Technology, National Research Tomsk Polytechnic University (Lenin Ave, 30, 634050, Tomsk, Russia), email: olgasumtsova0205@mail.ru

Article submitted 17 September 2017. Resubmitted 25 November 2017. Final acceptance 23 February 2018. Final version published as submitted by the authors.