European Proceedings of
Social and Behavioural Sciences
EpSBS

www.europeanproceedings.com
e-ISSN: 2357-1330

DOI: 10.15405/epsbs.2020.11.03.70

DCCD 2020

Dialogue of Cultures - Culture of Dialogue: from Conflicting to Understanding

VERIFICATION POTENTIAL OF EDUCATIONAL SEARCH TECHNOLOGIES (BIG DATA) IN SEMANTIC RESEARCH

Veronika V. Nikitina (a)*
*Corresponding author

(a) Moscow City University, 5b Maly Kazennyj pereulok, Moscow, Russia, veronikademch@gmail.com

Abstract

The global digitalization of education is changing the process of knowledge transfer, the methodology of education and self-education, new teaching media, as well as Digital Data inclusion in scientific research projects that allow us to bring educational technologies to a new level. Development of Big Data systems, including search engine results and text corpora, opens up new possibilities and allows us to set and solve research problems in an efficient way. With fast sharing of digital technology it perfectly opens up the completely new opportunities for today's researchers in modern teaching activities. The article focuses on the capabilities, verification potential of search engines and the procedure for implementing search queries in Big Data in a semantic research. The author suggests a description of the experimental methodology of semantic research which involves the use of Google / Yandex search engines, which are an effective tool for the semantic structure analysis. Empirical evidence of the effectiveness of this technique is provided as well. The data of search engines Yandex and Google coincide when determining dominant values. At the same time, the data of the Google search system still allows us to determine the semantic boundaries of the studied phenomena more accurate and fuller, compared with the data provided by the texts corpora.

Keywords: Big-Data, educational search technologies, search engines.

https://doi.org/10.15405/epsbs.2020.11.03.70
Corresponding Author: Veronika V. Nikitina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

Education, ultimately, assists learning in the form of transferring knowledge and skills. Having reliable information and proper evaluation of Big Data potential, as an advanced digital technology, is one of the most essential tasks and needs of modern education. With fast sharing of digital technology in modern teaching activities, it perfectly opens up the completely new opportunities for today's students and researchers. The global digitalization of education is changing the process of knowledge transfer, the methodology of education and self-education, new teaching media, as well as *Digital Data inclusion* in scientific research projects that allow us to bring educational technologies to a new level.

2. Problem Statement

Artificial Intellect can contribute to the education industry variously: *smart content*, *search engines, research engines* etc. *Smart content* includes a vast variety of products and technologies: *Cram 101 a*llows studying in the most effective way by dividing the material into logical manageable sections with briefing on the most essential facts and concepts casting away the secondary stuff; *chatbots* are software simulating human communication on certain topics. They are effective means in self-education when you are basically learning on your own but need some guidance on certain points etc. *Search engines* are the potential of Big-Data resources (Leung et al., 2019) They can be defined as technological capabilities for *Data store analysis* through the use of the entire world's data volume (Google, Yandex; national text corpora: BNC, COCA, National Corpus of the Russian Language, etc.), as human interaction has increasingly shifted online enormous streams of data have been generated in the wake of this shift, on the order of 2.5 quintillion bytes daily (Jones & Dye, 2018). Thus, which of the above platforms can be effective, reliable, and verifiable?

3. Research Questions

It appears that in order to qualitatively categorize the linguistic units to be analysed and to carry out a quaint quantitative analysis, the researcher can take advantage of the possibilities and verification potential of search engines (*Big Data*) in carrying out research, which have a wide range of possibilities (EMC Education Services, 2015). The validity of the data provided and interpreted is determinative. The question is, how do we iterate the validity of data?

4. Purpose of the Study

The question of the study allows determining the purpose of the study:

- to define big data query parameters;
- to determine through the experimental way their validity / non-validity.

For this purpose, the search is triangulated (Shcherba, 2004; Suleimanova & Lukoshus, 2015).

1) We specify search parameters of investigated lexical units or syntax constructs from text corpora BNC, COCA, etc., with restriction by type of discourse; Search engines Yandex/Google; web-

applications Google Trends and by time slices (implementation of search with these parameters is provided by text corpora if these parameters are significant for research).

2) The validity of the search results is determined by the following factors: firstly, the lexemes entered for the search must strictly correspond to the lexico-semantic variants within the semantic structure under study. For example, the search engine specifies verb search parameters that accurately reflect their lexico-semantic variants: *shudder* (followed by the exclusion of the conversion-noun) and *shudder to*. Thus, in the first case we can claim about the direct meaning of the word: *The humiliation made me shudder* while in the second case - about the derivative metaphorical: *I shudder to think what injustice will accompany the solution* (Suleimanova & Demchenko, 2018, p. 470). Secondly, the lexemes entered for search have to be quoted *'shudder', 'shudder to'*; otherwise the global network provides a required combination as well as all available collocations at the time of a response.

Despite the fact that the quantitative data provided by the search engines are variational, system monitoring and access fixation have made it possible to establish their validity, provided that quantitative methods (mathematical statistics methods) are used to interpret the results and fix the access date to the Search engine.

5. Research Methods

Traditionally, conducting semantic research, a hypothetical-deductive method is used (O. N. Seliverstova, O. A. Suleymanova, L. V. Shcherba, A. Mustajoki, O. V. Belaychuk, etc.). The procedure for conducting semantic research is multi-stage (Fomina, 2007). First and the most, the study begins with the collection of material (examples of the use of the studied language units, collected from authentic sources) and the dictionary definitions analysis of semantically similar language units from lexicographic sources. Currently, the resources that make it possible to provide the researcher with sufficient language material are mainly text corpora.

First and foremost we use National Corpus of the Russian Language (NCRL) for the study of language units (Nesset, 2019, p. 157). For the study of language units of the English language, depending on the type of discourse and the purpose of the study we use British National Corpus (BNC), Corpus of Contemporary American English (COCA), Corpus of Historical American English (COHA), Time Magazine Corpus English (TIME), Wikipedia Corpus (WC), Corpus of American Soap Operas (CASO) and the like (Sha, 2010). Using the methods of corpus linguistics, it is possible to obtain linguostatistical data on the frequency of use of a language unit and, accordingly, to establish contexts, distinguish genre and stylistic features of units, and then put forward a hypothesis based on the analysis of their distribution. For example, researching into the semantic structure of the English verb *shudder* we refer to text corpora and *Big Data* resources. It was found that there is a redistribution of values between direct and figurative meanings (see Fig. CIDE) - *shudder 1;* and figurative, metaphorical *shudder 2*, 'tied' to a specific syntactic structure of type *X shudders 2 to V* (Demchenko, 2019; Suleimanova & Demchenko, 2018, p. 469). The very example shows that search engines are effective means as a research technology and highly-demanded in linguistic thesis research.

Secondly and specifically, the researcher classifies the language material, highlighting the integral and differential features of the studied units for the formation of mutually exclusive groups. Thus, a content analysis is 'a research technique for making replicable and valid inferences from data to their contexts' (Krippendorff, 2018, p. 403). In the manner now being indicated, it could be argued that content analysis is used as a method to demonstrate a real time communication patterns, where the inference which arise through objectively and systematically investigating the meanings embedded into communication are developing a shared set of interpretations which could be further replicated, due to their focus on objectivity, validity and explicit rules (Mihailescu, 2019). Thirdly, with respect to meaning verification, a representative sampling is generated from mutually exclusive groups; an experimental sample is formed, which is provided to the informant-native speaker to assess the use (adequacy) / disuse of statements for verification of the primary hypothesis. Finally, a secondary verification of the experimental sample by informants is carried out, varying semantically similar elements within an identical statement for the subsequent formation of an adequate interpretation of the studied language units.

At the same time, the experimental method as educational search technology assumes the presence of both an informant and the use of *Big Data* resources (search engines Google / Yandex and the like) acting as a collective informant in the research. This stage involves the classification of the distributive boundary of the studied language unit their right and left co-occurrence. Co-occurrence is investigated depending on the object of research on the basis of NCRL or BNC and the like, as well as search engines Google / Yandex. The choice of a relevant search engine is determined directly by the language of the research material (English / German / Russian etc.). Russian language search engine Yandex handles search queries mainly of native Russian speakers, and is generally used by Russian-speaking users; in the study of Russian-language material it gives more reliable results because Russian texts are written mainly by native Russian speakers. The Google search engine is international and processes search queries depending on the language of the query, so in statistical analysis of data to exclude errors related to the responses of native or non-native speakers of the query language and to determine the reliability of the results provided by the search engine, it is important to use the public web application *Google Trends* of Google Corporation, based on Google search. The *Google Trends app* allows determining how often a particular unit is searched in relation to the total volume of search queries in different regions of the world and in different languages. The data is significant because, for example, in the query language 'English', which is the language of the entire world community, the probability of error provided by the collective informant (Google) is extremely high. Therefore, for linguistic research, territorial responses of the linguistic phenomena under study are extremely important, that is 'responses' of informants-speakers and not informants-non-speakers of the English language. Nonetheless, this approach of calculating the frequency presents technical, valid and time-consuming difficulties, due to the large amount of data collided (Roberts, 1997; Namenwirth & Weber, 2016).

Therefore, in order to reduce the objective 'noise' in the source data, the possibilities of implementing the *Google Trends* web application is used. Processing the data obtained by the search engine Goodle requesting the '*shudder to'* query, Google Trends application showed that the request for this token for the time slice from 2004 to 2019 'respond' exclusively in the United States (100%), in particular in the Central and Western parts of it (accessed February 2018). Therefore, since the collective informant is valid, this kind of data is also considered valid. However, the study of lexemes *shudders/shuddered/ shuddering* Google Trends has shown that these lexemes have a wider geographical use: New Zealand (100%), Australia (88%), Canada (69%), South Africa (65%), USA (52%), UK (52%), the Republic of Korea (28%), Philippines (15%), Pakistan, India (20%), Europe (from 1% to 6%), Brazil, Turkey, Japan (1%) (accessed February 2018). In this case, the objective 'noise' in the source data is eliminated using the methods of mathematical statistics. So, due to this fact, we claim that there is some difference, which has to be researched into.

6. Findings

The search string is sequentially entered quoted lexemes of verbs, accurately reflecting the syntactic construction '*sodrogat'sya, drozhat' i vzdragivat' ot otvrashcheniya, strakha, zlosti, nezhnosti, styda, radosti, udivleniya'(potential English correlates: shudder, tremble, shiver with disgust, fear, anger, tenderness, shame, joy, amazement)* In other words, to clarify the semantics of verbs *'sodrogat'sya, drozhat' i vzdragivat' (potential English correlates: shudder, tremble, shiver)*, we are to conduct quantitative monitoring of data on compatibility with the designations of basic emotions to clarify the distributive boundary, since preliminary corpus analysis showed the frequency compatibility of this group of verbs (see *Table 1*).

Table 01. Quantitative data monitoring on compatibility with the designations of basic emotions

	Yandex	**Otvrashcheniya**	**Strakha**	**Zlosti**	**Nezhnosti**	**Styda**	**Radosti, Udivleniya**
01.05.2018	sodrogat'sya ot (shudder with)	2000	308	10	4	2	NO MATCHES
08.08.2019		1000	238	8	4	1	
14.08.2019		1000	390	8	1	3	
17.08.2019		2000	449	9	2	2	
18.08.2019		1000	518	8	1	2	
02.11.2019		1000	505	9	7	65	
03.11.2019		710	343	9	1	4	
05.11.2019		2000	449	9	7	5	
06.11.2019		1000	518	8	1	2	

With identification of the Big Data variation we are researched into the validity / invalidity of the revealed data. So, in search databases Yandex and Google, with a time interval of 15 months, 1 week and for several days running were set quoted syntactic structures *sodrogat'sya, drozhat' i vzdragivat' ot otvrashcheniya, strakha, zlosti, nezhnosti, styda, radosti, udivleniya' (potential English correlates: shudder, tremble, shiver with disgust, fear, anger, tenderness, shame, joy, amazement)* followed by fixing the quantitative data.

It should be noted that conducting statistical analysis of empirical data in the semantic experiment, grouping data frequency of listrequest syntax, that is, finding the relative frequency (f) should the observed frequency of a trait or units of absolute frequency (n) divided by the total number of observations of the studied phenomenon (Levitsky, 2007, p. 79). So, at the absolute frequency of the verb '*sodrogat'sya ot otvrashcheniya'(shudder with disgust)* f = 2000, identified by the Yandex search engine (accessed May 2018, see *Table 1*), with the total number of responses in the syntactic construction equal to 2324 responses (n=2324) mathematical calculation of empirical data showed that the relative frequency of collocation *sodrogat'sya ot otvrashcheniya (shudder with disgust)* is equal to 0.86 (= 0.86). To get the percentage frequency, multiply the relative value by 100% (0,86 *100% = 86%). Thus, the empirical probability is 86%. In statistical studies, 95% and 99% probabilities have the greatest evidence base, which in statistics are called confidence probabilities, which correspond to significance levels (the value obtained by subtracting from 100% probability). Therefore, the probability p=86% corresponds to the significance level P=14%. In linguostatistics, the significance level criterion shows the reliability of empirically obtained results and the percentage of cases in which error is possible. It should be noted that when studying the distribution of polysemantic, semantic close constructions and determining the dominant and periphery of the value, confidence probabilities cannot be reduced only to the numbers 95%, 99%. Therefore, confidence probabilities below 10% are considered as objective big Data noise and are not taken into account in the analysis.

Statistical analysis of constructions *sodrogat'sya ot* with nouns denoting emotions made it possible to identify the following patterns. Yandex search engine captures two dominant collocations that express negative emotions *sodrogat'sya ot otvrashcheniya (disgust)* and *strakha* (*fear*) with confidence probabilities of 71% and 26%, respectively.The Google search engine also identifies two dominant collocations that express negative emotions *sodrogat'sya ot otvrashcheniya (disgust)* and *strakha* (*fear*) with confidence probabilities of 17% and 48%, respectively.

Statistical analysis of the startle constructions showed that the Yandex search engine also records only two dominant collocations expressing negative emotions of *vzdragivat' ot otvrashcheniya (disgust)* and *strakha* (*fear*) with confidence probabilities of 70% and 25%, respectively. However, the Google search engine records only one dominant collocation expressing negative emotions *vzdragivat' ot strakha* (*fear*) (with a confidence level of 36%), and two more dominant collocations with nouns *radost' (joy)* and *nezhnost' (tenderness)* expressing positive emotions with confidence rates of 20% and 13% respectively.

Statistical analysis of trembling designs showed that the Yandex search engine captures three dominant collocations expressing negative emotions to *drozhat' ot otvrashcheniya (disgust)* and *strakha* (*fear*), with confidence probabilities of 32% and 19%, respectively, and a collocation expressing more positive emotions *drozhat' ot radosti (joy)*. The Google search engine also distinguishes two dominant collocations that express negative emotions *drozhat' ot strakha* (*fear*) and *zlosti (anger)*, with confidence probabilities of 47% and 13%, respectively, and also distinguishes dominant collocations with the noun *radosti (joy)* expressing positive emotions with a confidence probability of 20% and 17%.

As you can see, the data from the Big Data search engines revealed that the verbs under study are in the general semantic field, but their scope is different or limited. So, the verb ***drozhat'*** (*potential English correlate: tremble)* is the semantic core in the studied semantic field, since the dominant verb collocations are the most common and show equally negative and positive meaning; at the same time, information about the physical manifestation of emotion is provided: Ya *drozhu ot strakha* / Ya *drozhu ot radosti (I'm trembling with fear / I'm trembling with joy)*. The verb ***vzdragivat*** (*potential English correlate: shiver)* in this semantic field is similar in meaning to the verb **drozhat'** (*potential English correlate: tremble)*, the information about the physical action is provided: *ya vzdragivayu ot strakha (potential English correlate: I shiver with fear)*. At the same time, the verb ***sodrogat'sya*** is used purely in a negative sense * *ya sodrogayus' ot schast'ya (potential English correlate: I shudder with happiness (which is incorrect in Russian)* and information is introduced about the cognitive activity of the subject. The data of search engines Yandex and Google coincide when determining dominant values. Concurrently, the data of the Google search system still allows us to determine the semantic boundaries of the studied phenomena more accurate and full, compared with the data provided by the texts corpora.

7. Conclusion

Providing any research on the basis of *Big Data* the researcher is to consider the following: *Big Data* searching tools are extremely efficient and valid subject to the execution of the above algorithm of actions.

To sum up the Big Data searing tools are:

- Almost unlimited amounts of data (empirical material) (Suleimanova & Petrova, 2018)
- Data are generated at high speed
- Consistency and validity of ‘quoted’ data
- The results are processed using methods of mathematical statistics
- Search engines are efficient and accurate tool in semantic research

Consequently, the development of Big Data systems, search engines creates unique opportunities for linguistic research as educational technology.

References

Demchenko, V. V. (2019). *K voprosu o «logicheskih krugah» v leksikograficheskoj praktike: komponentnyj analiz gruppy glagolov tipa “shudder”* [To the question of ‘logical circles’ in lexicographic practice: component analysis of a group of verbs as *shudder* and the like]. *Filologiya. Teoriya yazyka. Yazykovoye obrazovaniye, 2*(34), 109-113.

EMC Education Services (Ed.). (2015). *Data Science & Big Data Analytics*. John Wiley & Sons.

Fomina, M. A. (2007). *Konceptualizaciya «pustogo» v anglijskom yazyke (empty, free, blank, spare, unoccupied, vacant i void)* [Conceptualization of ‘empty’ in English (empty, free, blank, spare, unoccupied, vacant and void)]. *Vestnik MGLU. Series Lingvistika*, 541, 272–281.

Jones, M. N., & Dye, M. W. (2018). *Research methods: Big data approaches to studying discourse processes.* In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), *Routledge handbooks in linguistics. The Routledge handbook of discourse processes* (p. 117–124). Routledge/Taylor & Francis Group.

Levitsky, V. V. (2007). Kvantitativnye metody v lingvistiki [Quantitative methods in linguistics]. Nova Kniga Publ.

Mihailescu, M. (2019). Content analysis: a digital method. https://www.researchgate.net/publication/333756046_Content_analysis

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publishing.

Leung, B. T. H., Xie, J., Geng, L., & Pun, P. N. I. (2019). Search Engines: Transferring Information Literacy Practices. Shanghai Jiao Tong University Press.

Namenwirth, J. Z., & Weber, R. P. (2016). *Dynamics of culture*. Routledge.

Nesset, T. (2019). Big data in Russian linguistics? *Zeitschrift für Slawistik, 64*(2), 157-174. http://doi.org 10.1515/slaw-2019-0012

Roberts, C. W. (Ed.). (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Lawrence Erlbaum Associates.

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*(5), 377-393.

Shcherba, L. V. (2004). *O troyakom aspekte yazykovyh yavlenij i ob eksperimente. YAzykovaya sistema i rechevaya deyatel'nost'* [About the triple aspect of linguistic phenomena and about an experiment in linguistics]. Editorial URSS Publ.

Suleimanova, O. A., & Demchenko, V. V. (2018). Ispol'zovanie big data v eksperimental'nyh lingvokognitivnyh issledovaniyah: analiz semanticheskoj struktury glagola shudder [Using big data in experimental linguo-cognitive studies: analysis of the semantic structure of the verb shudder*]. Kognitivnye issledovaniya yazyka, 33*, 466-472.

Suleimanova, O. A., & Lukoshus, O. G. (2015). Znachenie yazykovogo znaka kak lingvisticheskaya konstanta [The meaning of a linguistic sign as a linguistic constant]. Proceedings of the conference "The humanities: issues and development trends", Innovatsionnyj tsentr razvitiya obrazovaniya i nauki (pp. 78-83).

Suleimanova, O. A., & Petrova, I. M. (2018). *Eksplanatornyj potencial teorii klassov dlya lingvisticheskogo issledovaniya: poryadok sledovaniya opredelenij* [Explanatory potential of the theory of classes for linguistic research: Word order in attributive group] (pp. 52–64). Filologiya: Nauchnye issledovaniya.