KnE Social Sciences

Convention-2019
Convention 2019 "Modernization and Multiple Modernities"
Volume 2020

Knowledge E
enriching | engaging | empowering

Conference Paper

Web Application "Linguistica": Concept and Possibilities of Utilization for Purpose of Increasing University's Virtual Interface Attractiveness

Yuliya Bogoyavlenskaya[1] and Maria Plotnikova[2]

[1]Doct. philol. sciences, professor of UrFU, Ekaterinburg, Russia
[2]Cand. philol. sciences, docent of UrFU, Ekaterinburg, Russia

Abstract

This article analyzes the limitations and problems of linguistic corpora utilization and attempts to solve them. The analysis revealed that among the main advantages of a significant part of the corpora, the possibility of analyzing large text fragments to determine their frequency of use in different contexts, as well as the high quality and speed of information processing can be noted, which contributes to time savings while working with corpora. In addition, a high degree of objectivity of studies using the corpus methodology is achieved with the help of wide possibilities for verifying quantitative data. Among the main limitations of most of the corpora we indicated the impossibility of discursive marking of the corpus, paid access to the corpus data or the limitations of the data available for work and the lack of meaningful context. The computer program Linguistica providing the basis for a web application of the same name, is designed to create different types of linguistic corpora: research and training, monolingual and multilingual and others. The Linguistica web application can also contribute to fulfilling the research projects of teachers and students specializing in linguistics as well as to increasing the attractiveness of the electronic environment of the university.

Keywords: corpus, web application, corpus linguistics, electronic environment of the university

Corresponding Author:
Yuliya Bogoyavlenskaya
jvbog@yandex.ru

Received: Month 2020
Accepted: Month 2020
Published: 28 September 2020

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the Convention-2019 Conference Committee.

1. Introduction

The rampant development of computer and, in particular, corpus linguistics, led to the creation of a significant number of corpora allowing linguists to do their research on the basis of a rich empirical material. The most famous and popular corpora include: Russian National Corpus, the General Russian Internet Corpus, British National Corpus, Frantext, the German Mannheim Corpus (COSMAS corpora), the University of Leipzig and many others. The advantages of corpus analysis are associated with the access to "large sections of textual discourse" and the determination of their frequency in different

contexts [1], with the quality and speed of material processing, and, consequently, significant time savings for researchers. The application of the corpus-based approach ensures the reliability of the obtained quantitative results, “representing the basis for further interpretations and identifying the qualitative characteristics of the objects under study” [2]. However, the use of corpora in research practice is contingent on certain restrictions. Some of the special computer programs attempt to overcome these restrictions but their functionality is limited to solving the research tasks in part. This article aims to analyze these restrictions and present the concept of the Linguistica computer program, on the basis of which a web application of the same name is currently under construction. The unique character of the web application consists in the possibility of creating a wide range of linguistic corpus types. The relevance of the work is also associated with the discussion of the practical utility of the application when it is available through the electronic learning environment of the university. The structure of the article is determined by the tasks set out above.

2. Analysis of Restrictions and Problems in the Use of Linguistic Corpora

The main types of automatic annotation are morphological and syntactic annotation. Much less common is semantic (NCRF), regional (GIKRYA, Global web-based English), genre and some other types of annotation. The current corpora, allowing the user to enter a word, a lemma or, much less often, a combination of words into the search line, *are* therefore *focused on* the *analysis of lexical or grammatical phenomena*. But communication units, having no standard modes of expression, and discursive phenomena, pragmatic features of texts, speech acts, etc., are not annotated in them, since the creation of such automatic annotation, at least at this point of the development of computer linguistics, is not possible. There are only a limited number of corpora that have discursive (for discursive-structural annotation, see, for example [3], accentological or other types of annotation, but such corpora are annotated manually, and the access to them is often restricted.

Restricted access or lack of access to some corpora is the second obstacle to their application. For example, the access to the largest corpus of works of the French literature Frantext requires a subscription on behalf of an academic or educational institution. The access to the database of the Russian press Integrum is a paid service. The full access to the SkethEngine platform is commercial-based, though it is possible to work with SkethEngine Skell – a corpus for the learners of English. The access to

the British Corpus, the American National Corpus or the International Corpus of English and some other corpora is limited, and sometimes it is only possible to search in the trial version of corpus.

The lack of access to the context of the phenomenon under study and to the text itself may be a significant inconvenience, as it reduces the possibility of the adequate interpretation of the phenomenon in the text. Most often, the researcher gets access to only one line, one sentence, or a certain number of characters (the limit depends on the corpus) where the requested word or expression is used.

The most important principle of contrastive studies is the *principle of the accordance of the sources of linguistic material (texts of the same functional style and genre, and of the same type of discourse)*, according to which the multilingual material should be extracted from the works related to the same functional style or discourse. This requirement cannot practically be fulfilled, considering that the available base files of different corpora have *a different volume* and *structure*, contain different *types of text documents with different chronological relatedness*.

A part of the above-mentioned problems may be solved with the help of specialized computer programs allowing the researchers to create their own corpora. Among others, we can mention the program WordSmith Tools [4] – a paid set of tools for studying the behavior of words in a text, powered by MS Windows. The software package was developed by the British linguist Mike Scott at Oxford University. There are three modules in WordSmith Tools: “Concord” which is used to create concordances (a list of all the uses of a given language expression in a context); “WordList”, containing a list of all the words or word forms included in the selected corpus; “KeyWord”, creating a list of keywords and grammatical forms in accordance with certain statistical criteria. In fact, WordSmith has become a model for any similar programs.

The GATE linguistic processor (General Architecture for Text Engineering – GATE [5]) is an open source of natural language processing system. The system solves such tasks as information extraction, manual and automatic semantic annotation, coreference analysis, work with ontologies, machine learning, and analysis of the message flow in blogs [6]. The project developers are specialists from the University of Sheffield.

The AntConc (AntConc [7]) toolkit is a freeware set of tools for corpus analysis, concordancing and text analysis developed by Lawrence Anthony at Waseda University, Japan. In our opinion, the most interesting tools are: “Concordance Tool” (display of results in the “key word in context” format); “Collocates” (display of word combinations used within a search request, which allows exploration of inconsistent structures in a language); “Word List ”(calculation of all the words in the corpus and formation

of a list with them, which allows determination of the frequency of a word in the corpus);" Keyword List "(development of a list of words used in the corpus often or rarely, compared with the sample corpus). AntConc is a cross-platform-based corpus analysis toolkit, powered by MS Windows, Linux and Mac.

The UAM program (UAM [8]) developed by Mick O'Donnell is also freeware. The program is easy to use, has a friendly interface, provides the opportunity for automatic morphological annotation, and also allows creating your own schemes for meta-annotation and text profiling, which makes it stand out from the rest corpus analysis tools.

The Semograf Graf Semantic Information System (Semograph [9]), developed by D. Baranov, K. Belousov, et al. (Perm State Scientific Research University), is an innovation with a wide potential of use. The system is designed to extract knowledge about subject areas from data domain, including textual samples, metadata, semantic components and semantic fields, as well as frequency, language and thesaurus dictionaries.

The programs, information systems and toolkits under review, as well as many others (for example, Corsis (Corsis [10]), Mystem (Mystem [11]), have a number of both absolute advantages (see above) and disadvantages: the paywall and the function powered only by MS Windows (WordSmith Tools), the restricted set of tools (availability of only one type of annotation and/or working only with collocations and keywords, etc.), the inability to create your own multilingual corpora, the lack of comprehensive user documentation, and the interface only in English, etc.

3. ``Linguistica'' Concept

We tried to overcome the identified restrictions and problems and developed the Linguistica program (State Registration Certificate No. 2014660349 in Rospatent (The Russian Federal Service for Intellectual Property, commonly known as Rospatent) dated October 6, 2014). This desktop application helps create and process different types of corpora. The functions of the program include input and storage in the database of an unlimited number of texts in two or more corpora; meta marking (information about the source, the author, the date of creation of a text document); creation of a classification tree of text parameters and fragment parameters; selection of significant fragments in the text for their further analysis; output of general statistical information on selected parameters and when applying filters; saving the results in external applications (Excel, Word). The program makes it possible to display concordance and copy selected text fragments into a buffer, search for a text by keyword, etc. Linguistica can be used to

research a wide range of linguistic phenomena and may also be useful in solving applied problems, for example in the linguistic criminological analysis, as well as in establishing text authorship, etc.

The practical utility of the program, tried and tested in the context of our doctorate thesis, has motivated us for further research. Currently, we are elaborating a web application of the same name to update the functionality of the program, improve the package tools and ensure the reliability of data preservation. The new Linguistica is a free web application. The corpus formed with its help may have different typological characteristics depending on the objectives of the study: educational/research, single language/multilingual (comparative), specialized/non-specialized, open/closed, full-text and/or fragment-text, verbal/creolized (multimodal) corpora, and the corpora of written, oral or mixed texts, as well as synchronous/ diachronic corpus.

A new, convenient and user-friendly interface in Russian is being developed for a web application. When working with the application, the user is able to create any number of research and educational projects in any language. Projects can be carried out both individually and in a team. A unique feature of the application is the ability to customize the parameters of the created project and consider different issues of the project through the chat application in the online mode. For some projects, the annotation of not only linguistic, but also graphic objects can also be valuable. The project data storage on a web server, as well as the ability to download it, ensures the reliability of data safety. For the purposes of researchers' convenience and time saving we optimize the procedure of data entry (text passports: static and dynamic parameterization with prefill), etc. An automatic calculation based on the method of verifying the reliability of the results, earlier developed by us (for more details, see [12], is integrated into the application to solve the problem of evaluation of the corpus representativeness.

4. Increasing the Attractiveness of the University Electronic Educational Environment

The interest in different ways to improve the availability and quality of higher education on the basis of the electronic educational environment leads to the development of new forms of its organization. In our opinion, a high functional potential of the application, comfortable usage and saving of time and labor resources provide a number of advantages. First, the use of an innovative software product can increase the motivation and effectiveness of the students working on their course projects, graduate papers and

master's dissertations in linguistics. Secondly, project advisors and directors who have access to the application will be able to better control the volume and quality of students' work, to guide, to improve and evaluate the results of students' research. Thirdly, the application can facilitate the working process of university lecturers and help increase their research activity. Fourth, the access to the application open for the contributors from other universities can develop inter-university cooperation. We do believe that the above-mentioned advantages present a strong case that the placement of such an application in the structure of the university's electronic educational environment can increase its interactivity and attractiveness for users – lecturers and students specializing in linguistics.

5. Conclusion

As the analysis shows, the linguistic corpora, programs, information systems and tools available on the Internet have a number of unconditional advantages and disadvantages such as the paywall, the limited set of tools (availability of only one type of annotation and/or working only with collocations and keywords, etc.), the inability to create your own multilingual corpora, the lack of comprehensive user documentation, the interface only in English, etc. We try to take into account all of the above disadvantages when developing the Linguistica web application designed to create linguistic corpora of different types (research and educational, mono-lingual and multilingual, etc.). We believe that this web-application could contribute to fulfilling the research projects of lecturers and students specializing in linguistics and to increasing the attractiveness of the electronic university environment.

Conflict of Interest

The authors have no conflict of interest to declare.

References

[1] Talmi, L. (2005). Metody kognitivnoj lingvistiki: predislovie / per. YU.YU. Ledeneva, G.N. Manaenko. *YAzyk. Tekst. Diskurs: nauchnyj al'manah*, issue 5, pp. 20-31.

[2] Bogoyavlenskaya, Y. V. (2016). *Parcellyaciya Kak Kognitivno-Semioticheskij Fenomen: Monografiya* Ekaterinburg: Izd-vo UrGPU, pp. 6, 180.

[3] Muhin, M. Y. (2016). Proekt sozdaniya kitajsko-russkogo parallel'nogo korpusa oficial'no-delovyh tekstov s diskursivno-strukturnoj razmetkoj. *Vestnik YUUrGU, Seriya Lingvistika*, vol. 13, issue 4, pp. 23-31.

[4] *WordSmith Tools*. Retrieved May 8, 2019 from https://www.lexically.net/wordsmith/.

[5] *GATE*. Retrieved May 8, 2019 from https://gate.ac.uk/overview.html.

[6] Rubajlo, A. V. and Kosenko, M. Yu. (2016). Programmnye sredstva izvlecheniya informacii iz tekstov na estestvennom yazyke. *Al'manah sovremennoj nauki i obrazovaniya*, vol. 12, issue 114, pp. 87-92.

[7] *AntConc*. Retrieved May 8, 2019 from https://www.laurenceanthony.net/ software.html.

[8] *UAM*. Retrieved May 8, 2019 from http://corpustool.com/index.html.

[9] *Semograph*. Retrieved May 8, 2019 from http://semograph.com.

[10] *Corsis*. Retrieved May 8, 2019 from http://corsis.sourceforge.net/index.php/Main_Page.

[11] *Mystem*. Retrieved May 8, 2019 from https://tech.yandex.ru/mystem/.

[12] Bogoyavlenskaya, Y. V. (2016). Reprezentativnost' lingvisticheskogo korpusa: metod verifikacii dostovernosti poluchennyh dannyh. *Politicheskaya lingvistika*, issue 4, pp. 163-166.