ISSN: 2249-7137 Vol. 10, Issue 6, June 2020 Impact Factor: SJIF 2020 = 7.13

ACADEMICIA
An International Multidisciplinary Research Journal

ISSN (online) : 2249 –7137

Editor-in-Chief : Dr. B.S. Rai

Impact Factor : SJIF 2020 = 7.13

Frequency : Monthly

Country : India

Language : English

Start Year : 2011

Indexed/ Abstracted : Scientific Journal Impact Factor (SJIF2020 - 7.13), Google Scholar, CNKI Scholar, EBSCO Discovery, Summon (ProQuest), Primo and Primo Central, I2OR, ESJI, IIJIF, DRJI, Indian Science and ISRA-JIF and Global Impact Factor 2019 - 0.682

E-mail id: saarjjournal@gmail.com

VISION

The vision of the journals is to provide an academic platform to scholars all over the world to publish their novel, original, empirical and high quality research work. It propose to encourage research relating to latest trends and practices in international business, finance, banking, service marketing, human resource management, corporate governance, social responsibility and emerging paradigms in allied areas of management including social sciences , education and information & technology. It intends to reach the researcher's with plethora of knowledge to generate a pool of research content and propose problem solving models to address the current and emerging issues at the national and international level. Further, it aims to share and disseminate the empirical research findings with academia, industry, policy makers, and consultants with an approach to incorporate the research recommendations for the benefit of one and all.

ISSN: 2249-7137 Vol. 10, Issue 6, June 2020 Impact Factor: SJIF 2020 = 7.13

ACADEMICIA
An International Multidisciplinary Research Journal
(Double Blind Refereed & Reviewed International Journal)

DOI: 10.5958/2249-7137.2020.00749.1

PROGRAMS USED TO CREATE THE LANGUAGE CORPUS AND THEIR PRINCIPLES

Abdullayeva Oqila Xolmo'minovna*

*PhD Researcher of ????????
Tashkent, UZBEKISTAN
Email id: abdullayevaoqila@gmail.com

ABSTRACT

In the field of corpus linguistics, the creation of corpus programs that can directly analyze linguistic data and present different results to the researcher is classified and categorized according to the capabilities and characteristics of the programs. The advantages and disadvantages of each classified section are discussed. The language corpus analyzes the periodic sequence of the creation of programs that perform tasks such as searching and retrieving data for analysis.

KEYWORDS: *Corpus, Software, Linguistic Software, Abstract, Concord, Antwebconc.*

INTRODUCTION

Nowadays, language has become one of the areas that needs not only a theoretical but also a practical approach. Corpus programs are used in many languages to track unique and dissimilar linguistic features, speech patterns, and provide analytical conclusions through natural language texts (written or audio). computer technology creates programs and their design, special software. Due to the development of fast and convenient computer technologies, the creation of large-scale corpora of national languages, about a hundred programs for various purposes have been created.

Why do we need corpus programs? First, the creation of corpses is not possible without special software. Second, we need software in the analysis of broad-layer linguistic units in corpora. Third, the corps can be used not only by professional philologists, linguists, lexicographers, but also by language teachers and students. This requires corpus applications that can be used for a variety of purposes.

MATERIAL METHOD

T.McEnery and A.Laurence have provided information about the separation of corpus programs from corpses through textbooks and articles. The classification of corpus programs was initially described by T.McEnery and A.Hardie. Programs are grouped by period of creation and activity.The scientist divides the programs into four generations[1]. 1) First generation programs: Includes Concordance Generator, Discon programs. Although these programs serve as the foundation for subsequent programs, their functionality is limited, recognizing only a limited number of characters. More text is limited to tracking the amount and list of words. 2) Second-generation programs: Oxford Concordance Program, Longman Mini-Concordancer, MicroConcord, which could be installed on personal computers and used in language teaching from small analyzes.3) Third Generation Programs: WordSmith Tools are AntConc programs that can be recognized as high quality and wide range of applications with their use and various functions. 4) Fourth generation programs: corpus.byu.edu, SketchEngine, which are characterized by high speed and large amount of text. Although they are the latest and most advanced, fast search engine apps, they do have some limitations for the user. For example, SketchEngine, while fully copyrighted, makes it difficult to get complete information about the program, because it is a web-based program, the user can not download it to a personal computer, and pay a monthly fee for use. need Although downloading and using Corpus.byu.edu is free, the search engine charges a fee when it exceeds a certain amount. Many of the above four generations of programs have taken English or American-English to their research center or are tied to a specific language corpus that a researcher cannot use in any language search. However, among the programs of this generation, corpus programs created by A.Laurency can be highly appreciated in terms of openness and ease of use in the corpus. The AntConc program is also important in terms of language selection, meaning that different language researchers can download the program, monitor language and language features, and draw statistical and analytical conclusions. We also used this program in the Uzbek language. In an article published in 2013, A. Laurency provided information on the creation, structure and operation of a new generation program. So, the name of the fifth generation program is AntWebConc, which covers three integers. This is the Model-View-Controller[2], ie the model - the corpus database, View - the browser interface, Controller - the controller (for example, concord analysis, etc.). According to the information provided about the program, any researcher who does not have a deep knowledge of programming and language can understand and use the program. The program is expected to be announced.

DISCUSSION

In the Oxford Handbook of Lexicography, Iztok Kosem emphasizes the need for well-designed corpus programs for corpus inquiries, especially for lexicographers. I. Kosem groups programs into 3 types according to their use: standalone programs, computer-based programs, online programs. Evaluates standalone programs as programs that are not connected to the case and are stored on the user's computer. WordSmit and MonoConc Pro are included in this group. Stand-alone applications process a limited amount of limited corpus data, so such programs may reduce their usefulness in modern large-scale corporations. Online applications are also classified as stand-alone and corpus-connected online corpus applications[3]. Among the online programs is the most popular and improved Sketch Engine program in recent years. The program is also

significant in that additional features complement it. However, any researcher cannot access this program quickly and easily; there are requirements and limitations of the program.

ANALYSIS

We think that it is appropriate to analyze the classification of programs used in the creation of language corpora and the implementation of linguistic operations in them directly using the methods of observation, experiment and comparison, because the similarities and differences of programs performing simultaneous and the same activity requires interpretation, direct observation, and experience. We try to group programs according to their function and application.

First of all, corpus programs that can be installed on the user interface or used online, depending on the application: 1) Corpus applications that can be installed on Windows, Mac, Linux operating systems: aConCorde, ANNIS, AntConc, ANVIL, CorpKit, Segment Ant, Corpus programs such as TagAnt, TextSTAT. For example, the AntConc concordant program has developed separate versions for Windows, Mac, and Linux operating systems, and the program can be downloaded either from the website or from other computers where the program is available.2) There are many programs available only online today, and due to the wide range of capabilities of these programs, a specialist can fully explore the capabilities of the corpus: AMALGAM, BNCWeb, CLiC, HeidelGram Web-Based tools, Sketch Engine, Programs like Cortext Manager work online. Some online corpus applications are housed in private language corporations and are designed specifically for that corpus. For example, the ANC2go program was created for ANC (American National Corpus), and BNCWeb was created for BNC (British National Corpus).For example, the ANC2go program was created for ANC (American National Corpus), and BNC Web was created for BNC (British National Corpus). It is not always easy and convenient to use programs that are linked to a specific language, it is impossible to observe the linguistic phenomenon of texts of different genres in any language. On the other hand, online corpus applications can cover large amounts of complex and multi-layered complex text and perform linguistic analysis in seconds, such as part-of-speech tagging or tokenization.

We have divided the corpus programs into several groups according to their function: text collector programs, annotation programs, concordance programs, statistical analyzer, tag, grammar analyzer (parsing), tokenizer, semantic analyzer, audio text analyzer. and programs that perform mixed analysis.

1. Text compilers: A collection of corpus texts can be collected from a database on a user's computer or from the Internet at a specific address, or in a variety of genres and fields. WebBootCat, a program that collects text from the Internet, and SketchEngine, which are available in computer memory or can scan a source to create a body of data.

2. Interpretive programs: @annotate, AMALGAM, Atomic, Dexter, PALinka, Synpathy are only annotating programs. These programs also comment on the issue at hand. For example, in oral corpuscles, semantic annotation is included if semantic analysis is intended in the corpus, just as syntactic analysis is more important in interpreting familiar, prosodic features of speech[4].Annotation programs provide complete information about the corpus: the author of the text used, the publication, the source (book, newspaper, web page), the size of the text, etc. Additional text may also include parts of speech, lemmas, and tokens. But scholars such as Sinclair and Laurency believe that annotation of each corpus causes a loss of textual integrity,

while on the other hand a researcher who is profoundly familiar with that language through corpus annotations explores the features of the text: can easily analyze.

3. Concordant programs: AntConc, BNCWeb, CasualConc, Concordancer, ParaConc, Wordstatix programs. It is one of the simplest but most important programs in corpus programs. Because these programs allow you to see and analyze the number of times each word and word form occurs in context. Concordances define two important forms of analysis in corpus linguistics: quality and quantity[5]. Concordant programs allow you to search for words, word forms, or phrases of a certain length in the body.

RESULT

The corpus of Uzbek-language news items, "Saylovlar" was analyzed using the AntCons concord program.

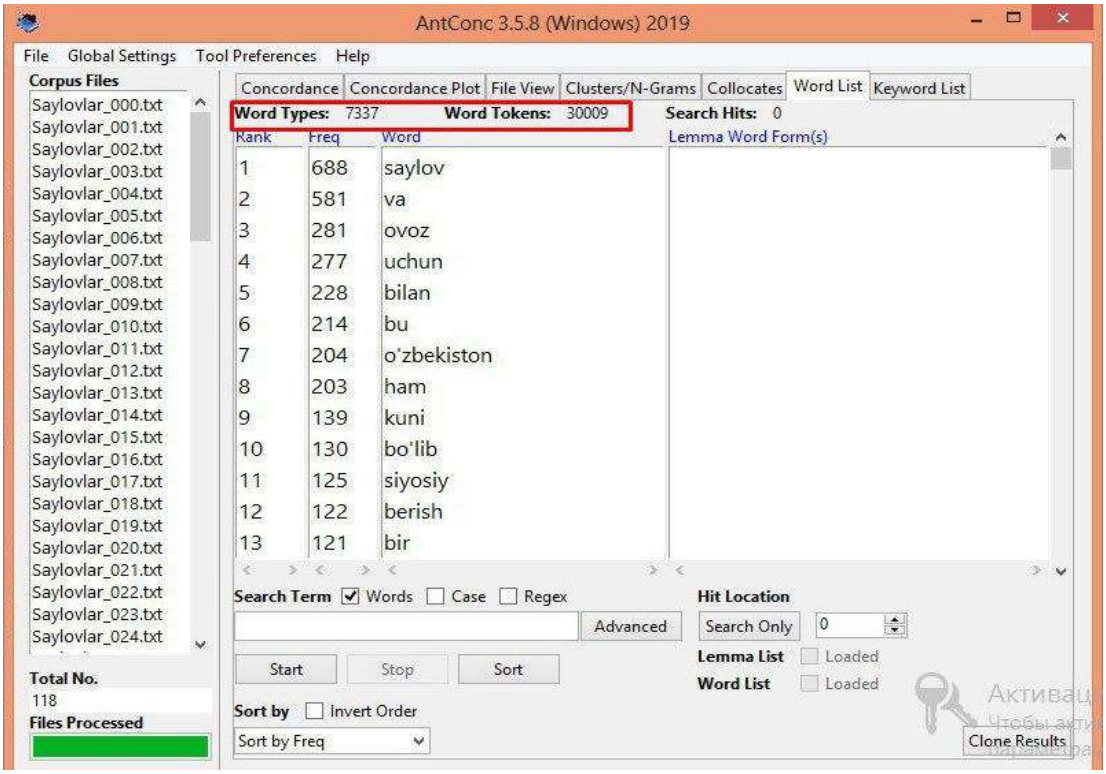

Figure 1. List of words

Of course, we prepared this very small case for experimental testing. There are 118 texts, as shown in Figure 1, showing a list of words.

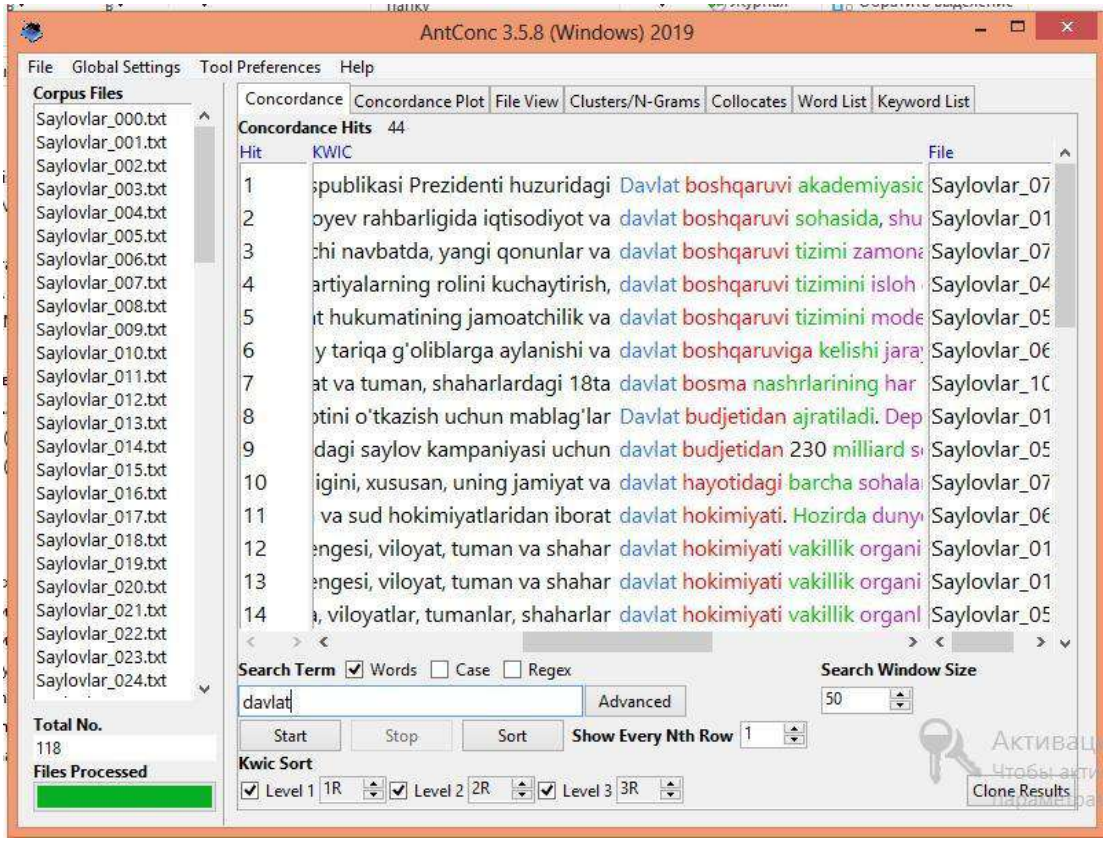

Figure 2. The concord is described.

Figure 2 above shows the words and phrases in the context. The word is described in the context in which it is used. This process is called KWIC, or "key word in context" in corpus linguistics[6].

1. Statistical analysis programs are similar to concordance programs. For example, the Wordsmith program, but the UCS Toolkit program, which detects data statistics in large libraries, or the Log-Likelihood and Effect-Size Calculator program, which checks the size and frequency of two corpses, belong to this group. We thought it was a complex corpus for the average researcher.
2. Programs that separate parts of speech or speech differ from other programs in their complexity and uniqueness in each language. Even when each part of a sentence is tapped, that is, encoded, this tag may change depending on the content and structure of the sentence in another text.

In addition, corpus applications can be classified as grammatical analysis, semantic analysis, audio text analysis, and mixed multi-part applications according to the query.

Differences in the classification of software can be observed in corpus linguistics research and sources. For example, Humboldt University's website on Corpus Linguistics and Morphology categorizes software as online applications, API programming interfaces and frameworks, corpus creators, annotators, tags, corpus analysis tools, and more[7].

Applications can also be categorized into open or closed, paid (for a certain period of time or without restriction) or free for any user.

CONCLUSION

As a conclusion, software that displays the capabilities of language corpora can be further classified according to its internal features, but this does not mean that any type of software can be used in corpora. Typically, separate sites are created for each national language corpus, with software that provides analysis and conclusions, or the simplest user prefers to use free and easy-to-install software on the Internet. In our opinion, corpus programs are considered to have the most users if they do not focus on a specific language, if the access and use instructions for the researcher are complete, and if the user does not require in-depth knowledge of the programming language. Because such programs can be used not only in research, but also in various stages of language teaching and education in general.

REFERENCES

1. McEnery T., Hardie A. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press, 2012
2. Laurency A. A critical look at software tools in corpus linguistics. Linguistic Research, 2013 https://www.researchgate.net/publication/267631312_A_critical_look_at_software_tools_in_corpus_linguistics
3. Kosem I. The Oxford handbook of Lexicography. – United Kingdom, 2016
4. McEnery T., Xiao R., Tono Y. Corpus-based language studies: an advanced resource book. – London and New York, 2006
5. Software.https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/korpuslinguistik/links-en/software