Corpus as the research object in linguistic domains

Yulbarsov Ochilbek
Teacher of the National University of Uzbekistan. Tashkent
Email: uzochilbek91@gmail.com

Annotation: This article explores the practical and theoretical value of Corpus-based approach in language research and its application, highlighting its connective use with computer technology in related fields. Furthermore, it explains its current use in modern linguistics in association with other fields. However, the hardships of Corpus-based approach are aslo mentioned in the article together with ways of how to fix them in language research.

Keywords: computer tools, concordance, contextual meaning, decontextualization.

INTRODUCTION

One of the global tasks of modern linguistics today is the purposeful use of computer technologies in language research and application. Classification of each direction or term being coined in linguistics as its own field, its practical use, information and communication of the Uzbek language. Tasks such as application through technologies require inclusion in the list of priority goals of language research. In particular, the field of Corpus Linguistics, which is an integral part of Computer Linguistics, is one of the important instruments in the practical-text aspect of language.

It is known that the first corpora existed without computer technology and researches were carried out. Later, the development of science required the search method (concordance) to be carried out in electronic format texts, and the need for interdisciplinary relationship with computer linguistics arose. At the same time, computer linguistics usually refers to the application of computer tools (programs, computer technologies for data organization and processing) in specific conditions, situations, problem areas and the scope of language models not only in linguistics, but also in other disciplines[1].

The creation of the first corpus dates back to 1812, when the German scientist Kadling analyzed the distribution of consonants in German words. However, at that time the term computer did not even exist. Currently, one of the global problems of the 21st century is to preserve the national character of natural languages. It has become an urgent task to consistently conduct research on NLP and language technologies in the creation and development of electronic corpora of world languages[2].

So, today the concept of Corpus appears as a practical presentation of its functions in the field of language and linguistics. Accordingly, the term "corpus" is used in many places in the language. Viewing the corpus as a linguistic instrument is an example of the sociolinguistic flexibility of contemporary linguistics. In fact, despite the fact that this field is one of the new scientific

---

[1] N.B.Ataboev. Main features of Corpus Linguistics. www.journal.fledu.uz/ "Foreign Langugaes in Uzbekistan". Scientific-methodical electronic journal.№2-2019. б. 38.

[2] Abduraxmonova N. Computer models of Uzbek electronic corpus (monograph). / Toshkent:Muharrir, 2021, 202 p.

research directions for linguistics, its theoretical essence has been clarified in most corpus-related researches. In particular, it has been noted about the meaning of the concept of corpus, its functional importance in linguistics, from the history of its origin (however, this does not mean that the theoretical essence of the concept of corpus in Uzbek linguistics has been fully studied). This will greatly help to understand the main content of the work faster and for the representatives of different fields to be able to imagine. It is correct to say that corpus has become an integral part of modern linguistics. Most of the researched language issues require its characterization as a subject (material). Because language phenomena - editing and analysis, translation, statistics, morphematic processes in the connection of words with each other - are described on the basis of the characteristic of the language. But considering that language is a social phenomenon, it should be taken into account that it occurs as a natural process. It can be argued that the natural features of the language, such as phonetic, dialectal, pragmatic, extralinguistic, oral, are phenomena of a stable form. The language corpus plays a key role in the study of these processes. At the same time, it is also used as a research object of corpus linguistics. The development of the term "corpus" into a branch of science in language has led to specific analyzes for the directions of language as a material. Therefore, Corpus Linguistics, as an integral branch of Computer Linguistics, can be said to be a field with specific theories. There are several functions of corpus linguistics, the most important of which is the lexicographic task - creating a computer database of textual resources. The text base effectively helps scientists to quickly find information for various fields and use it in their research. Scientists' research on corpus linguistics has become a new stage in modern linguistics.

In global linguistics, scientists in fields such as corpus linguistics and corpus lexicography conducted research with their theoretical views. Sara Laviosa D. Biber, S. Konrad and R. Reppen[3] researched the role of corpus linguistics in language learning. Isabel Duran-Munoz and Gloria Corpas Pastor, H. Lindquist, N. Leach, N. S. Dash, F. Meyer, T. Grays, S. Hunston, T. Mc. Enery, Z. Xiao, Y. Tono, A. Wilson and others studied this field based on different approaches. J. Sinclair discussed the importance of using corpus linguistics[4].

As a result of the studies of scientists, different theories and types of research are formed in corpus study. Accordingly, today corpus-based research is mainly carried out in 4 ways:

1) corpus-based lexicography;

2) corpus-based translation;

3) corpus-based analysis;

4) corpus-based language pedagogy.

Problems and theoretical views of corpus-based lexicology were studied in the works of V. Tuber,[5] A. Hurskeynin,[6] D. J. Prinsloo[7] and other scientists.

V. Tubert states that finding the meaning of words from text samples (context) is not the main principle of corpus linguistics. Touching on the main principles of corpus lexicography, he listed 4 differences between it and traditional linguistics: first, corpus linguistics does not use the corpus only for examples. He studies them systematically; secondly, corpus linguistics does not try to decontextualize the objects it describes. Contextual meaning does not generalize; thirdly, corpus linguistics tries to collect and apply examples of their various relationships to the database, taking into account the use of words in different forms. In contrast to traditional

---

[3] Laviosa S. Corpus-based tran Corpus-based translation studies: Where does it come from? Where is it going? Studies in the Languages of Africa. Vol., 35, 2004 - Issue 1. 2008. P. 6-27.

[4] Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. – Cambridge University Press, 1998. – P.300.

[5] Teubert W. Corpus Linguistics and Lexicography. Article in International Journal of Corpus Linguistics. December 2001. https://www.researchgate.net/publication/240510906_Corpus_Linguistics_and_Lexicography

[6] Hurskainen A. New Advances in Corpus-based Lexicography. https://hdl.handle.net/10520/EJC60458. 2003.

[7] Prinsloo D.J. Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus/ Article. Journ.: Lexikos / Vol. 25 (2015)

lexicography, it tries to place the meanings of words on the basis of an ontological concept independent of language; fourth, corpus linguistics is not a single text item or phrase. It also studies the semantic interaction between text elements and context.

The linguist A.Hurskainin justifies the fact that the text of a certain language can become the basis for a large and grammatically balanced dictionary through the process of corpus elematization. At the same time, he considers the lack of words used in oral communication as a disadvantage of the text-based lexicographic corpus. It is emphasized that such words should be considered separately from the point of view of their use, and the possibility of using the transcription of corpus texts as a solution to the problem, however, the currently large database of oral texts is insufficient.

In the opinion of scientists, it is specially explained that the lexicographic possibilities of the corpus are great. As one of the problems in the corpus, it is based on the electronic form of the existing traditional dictionaries in the practical process - the corpus model, which is the primary practical necessity of computer linguistics, in particular, corpus lexicography.

In the second method - corpus-based translation, scholars have conducted many researches. S. Laviosa,[8] Isabel Duran-Munoz and Gloria Corpas Pastor, Ger De Sutter and Marie-Aude Lefer have studied corpus-related aspects of translation in their research. In particular, S. Laviosa expressed his opinion about translation studies based on visual corpus and translation studies based on theoretical corpus. In Computational Linguistics, Isabel Duran-Munoz and Gloria Corpas Pastor[9] emphasized the benefits of language corpora for translators. In particular, he pointed to the importance of the correct management of terms at different stages of translation functions and the combination of sources in the corpus in optimal search processes. He also noted that various types of electronic resources that can be compared in different ways or provide access to parallel corpora are included in the corpus. In a co-authored article by Ger De Sutter va Marie-Aude Lefer,[10] it is mentioned that "On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach) critically analyzes the current state of corpus-based translation studies and pays attention to its description, methodology and theory". Touching on the empirical nature of translation studies, they emphasize that translation problems can be solved with methodological and theoretical concepts. Linguists have analyzed that the description of translation as a multifaceted linguistic activity and product limited by socio-cultural, technological and cognitive factors at the same time leads to an understanding of what exactly translation is and how it is formed.

A third method is corpus-based analysis of studies. According to the researches of scientists, this method has become an important stage of modern linguistics as corpus research is used as an instrument, that is, as a methodology. L. Flowerdev, K. Otis,E. Sagi,[11] M. Nelson, D. Bieber, Manning, D. Christopher and S. Hinrich and many other scientists have conducted research in this regard. The corpus is analyzed from a methodological point of view in the research works of linguistics, computer linguistics, NLP (Natural language processing), law, psychology. In the research of K. Otis and E. Sagi, one of the most useful features of corpus word analysis is the calculation and comparison of semantic vectors (techniques of mathematical expression of word meanings) on the example of words and phrases. Based on the corpus, the mechanism of random selection shows the semantic connection of the words in addition to their meaning in the language. Hence, Corpus-based analysis as a methodology in Linguistics and Natural Language

---

[8] Laviosa S. Corpus-based tran Corpus-based translation studies: Where does it come from? Where is it going? Studies in the Languages of Africa. Vol., 35, Issue 1. 2008. P. 6-27.

[9] Isabel Durán-Muñoz & Gloria Corpas Pastor. Corpus-based lexicographic resourses for translators: an overview. 2019 - wlv.openrepository.com.

[10] Ger De Sutter & Marie-Aude Lefer. On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach / Studies in Translation Theory and Practice.Vol.: 28, 2020 - Issue 1. P. 1-23.

[11] K.Otis, E.Sagi. Phonaesthemes: A Corpus-Based Analysis/ Proceedings of the annual meeting of the cognitive science society, 2008. P. 66

Processing (NLP) refers to conducting analysis and research on a large collection of texts. Studying the structure and use of language based on empirical data forms the basis of the field of corpus linguistics, that is, corpus research. In addition, linguist M. Nelson in his article "Semanticassociations in Business English: A corpus-based analysis" (Semanticassociations in Business English: A corpus-based analysis) the connection is analyzed from the point of view of oral and written form. Using a million word combinations, the researcher substantiates the fact that the connection of words is often associated with compound words that have a negative or positive connotation. With this, it analyzes the mutual semantic influence of words related to entrepreneurship (business) in the English language. According to the results of the research, it was found that words in the business environment are regularly associated with word groups that have semantic similarity. Words also have business-related prosody, which is explained as not being exhaustive and often expressing strong tendencies for lexical relationships rather than strictly defined relationships. In general, considering the corpus only as the main object of study of computational linguistics may limit its functional possibilities. In our opinion, as a result of studying the recent researches of scientists: methods such as corpus linguistics, semantic association, field terminology, changes of language forms over time can be complementary research objects of all fields; the corpus increases the level of cooperation of different fields in scientific researches; along with interdisciplinarity, it leads to separate, specific theories in field studies and is evaluated as an object of study;

Therefore, it is correct to look at the corpus as a primary instrument in research in various fields.

The fourth method of analysis is corpus-based language pedagogy. In language learning and teaching and in achieving the main result, the use of the corpus serves as an important methodology. By analyzing large volumes of texts, he determines the linguistic features of the language, its possibilities, and the meanings of words in the context. Corpus-based language pedagogy allows language learners to learn a language effectively by providing them with practical examples and creating learning materials based on them. E.Cotos, R.Yuan, J.Yang, L.Flowerdev, A.Boulton, H.Tyne, G.Chalishkan, SİK Gönen, and other scientists have conducted research in this regard. In this regard, E.Cotos in his study entitled "Introduction to the handbook of technology and second language Teaching and learning" (Introduction to the handbook of technology and second language Teaching and learning) materials of LSP (Language for specific purposes) analyzed the importance of using corpora as sources of information for creating The researcher mentions that direct combination of corpus applications is realized through interactive CALL (Computer-Assisted Language Learning) technologies. Butechnology, on the other hand, emphasizes an inclusive corpus for genre-based writing pedagogy. The practical part of this method was analyzed by G.Çalışkan, SİK Gönenlar in their work[12]. In particular, the researchers put forward the opinion that the dictionary is considered as one of the important areas of language learning, that it is almost impossible to convey meaning without this basic element of language and communication. As a result of practice and questionnaires conducted with the participation of university students, they found that it is possible to show different aspects of vocabulary through the corpus and to achieve the result quickly and efficiently. They put forward the idea that language teaching should be integrally connected with the corpus in pedagogical activity. Because the oral and written form of words, their grammatical and collocational movement, frequency, stylistics, conceptual meaning aspects, association with other words form corpus-based language pedagogy. Some conclusions can be reached as a result of researchers' research. In our opinion, corpus-based language pedagogy allows language learners to learn a language effectively by presenting practical examples and creating educational materials based on them; in the corpus-based approach, natural texts are used as educational materials; helps students learn the language used in real life; corpus linguistics is used to study the meanings of words in different contexts. It helps students understand the subtle meanings of words and how they are used in different contexts. However,

---

[12] G.Çalışkan & S.İ.Kuru Gönen. Training teachers on corpus-based language pedagogy. /Journ.: Language and linguistic studies. 2018. 14(4), P. 190-210.

although corpus-based language pedagogy has many advantages in language learning and teaching, it also has some disadvantages. For example, corpus-based approaches require large volumes of electronic texts (corpora) and special programs for extracting information from them; analyzing corpus data and using them for pedagogical purposes is a complex process, and teachers and students must have special knowledge and skills to use such methods; corpora are often based on certain cultural situations. For students from other cultures or contexts, it may not be appropriate or difficult to understand. Despite these drawbacks, corpus-based language pedagogy can be very effective in language learning if well-designed and properly applied.

The role of corpora in the development of computer lexicography.

A corpus embodies an important methodology known as lexicography or corpus lexicography. Understanding the essence of the lexicographic database in the language corpus means that we need to understand the term lexicography from the point of view of linguistics. Accordingly, if we consider "lexicography" as a branch of linguistics dealing with lexicography, composed of the Greek words "lexico" and "graphia", and corpus lexicography as an important tool for studying the possibilities of lexicography in computer linguistics and using it in practice. In the previous part of our work, we have touched on the analysis of the corpus based on lexicology. In this chapter, we will try to analyze the role of the corpus as a lexicographic instrument. In the field of corpus lexicography, practical and theoretical studies have been carried out by world scientists. In particular, J. Sinclair, Geoffrey Leach, Patrick Hanks, Sue Atkins, Janet Holmes, K. Green and J. Lambert and other scientists contributed to the development of the field. The "Collins COBUILD English Language Dictionary" project created in the 1980s is one of the first projects in this regard.[13] This project is a large corpus-based dictionary project that aims to build a dictionary based on texts written in real English. Geoffrey Lech[14] actively participated in the creation of one of the modern corpora of the English language, the Lancaster-Oslo/Bergen (LOB) Corpus, and greatly contributed to the development of corpus-based research. J. Sinclair[15] in his research work entitled "Prospects of automatic lexicography" (Prospects of automatic lexicography) explores the prospects of automatic lexicography and computer technologies and the possibilities of automating the processes of lexicography in corpus linguistics. By analyzing corpus data, it discusses the possibilities of obtaining information about word meanings and frequency of use, identifying new words, and continuously updating dictionaries. Considers methods of corpus lexicography that speed up automatic processes and lead to more accurate results. Patrick Hanks[16] argues that word association norms are derived by identifying the meanings of words in dictionaries and determining how people understand the relationships between words. In addition, the linguist suggests using mutual information to determine word collocations. That is, words that are often used together should be clearly indicated in dictionaries. These methods of mutual information exchange require working with a large amount of corpus data. This information helps you determine in which contexts words are used and with which words. Based on the views of the scientist, it is understood that the contextual meanings of words can be further defined and explained in dictionaries with the help of mutual information of the associative norms of words. With the help of these methods, dictionaries understand not only the main meanings of words, but also what they mean in different contexts. In our opinion, it is correct to say that this process is an important mechanism of corpus lexicography. However, despite the fact that solutions to the problems of the field are being found in theoretical research, the theoretical and practical possibilities of dictionary corpora have

---

[13] J. Sinclair. Collins Cobuild English language dictionary. / harvest@worldveg.org. 1989. P. - 1703.

[14] E.D.Beale. Grammatical analysis by computer of the Lancaster-Oslo/Bergen (LOB) corpus of British English texts/ Annual Meeting of the Association for …, 1985. - aclanthology.org. P. 293-298.

[15] S. Maior. Lexicographica. Prospects for automatic lexikography./ Symposium on Lexicography VII: Proceedings of the Seventh International. 1996. P. 1-11, https://books.google.co.uz/books?id=hX5dDwAAQBAJ&lpg=PA1&ots=DVRxlo9rXn&dq=J.Sinclair%20lexicography&lr&hl=ru&pg=PR5#v=onepage&q&f=false

[16] P.Hanks. Word association norms, mutual information, and lexicography/ Computational Linguistics. 1990. Vol.: 16 Iss.: 1. pp 22–29

not been fully explored. That is, the improvement of the lexicographic process in the process of processing natural languages is still one of the actual tasks. Linguists such as I.I.Sazhenin, N.A.Shamova, and L.N.Zasorin from Russian scholars also tried to justify the role of the corpus in lexicographic development in their research. Jumalada, I. Sazhenin's[17] dissertation entitled "Corpus methods in lexicography: the experience of creating a dictionary corpus model" gave his views on the creation of corpuses. Explaining the corpus methods used in the field of lexicology, Sazhenin mentions that corpora include a wide range of texts and their use in linguistics, the possibility of analyzing the written word form and their changes. According to the scientist, the main problem in creating electronic resources is that the machine cannot fully work with natural language text to provide information that matches the possible range of user requests. Solving this problem can be said to be one of the urgent tasks facing language corpus specialists and linguists. In fact, each dictionary contains different information due to its uniqueness. However, despite the large number of electronic lexicographic resources on the network, there is still no resource that meets certain requirements: the volume of content material, a philologically appropriate search system that allows to obtain various types of linguistic information from the entire content volume, for example. Nevertheless, in our opinion, relatively accurate linguistic information can be obtained from dictionaries. Perfectly developed dictionaries reflect phonetic, morphological, syntactic, semantic and other information and embody all levels of the language. It is understood that the corpus is an important tool for language representatives to better understand a given language, to come to theoretical and practical conclusions, and the lexicographic database is an integral and complementary part of it. In her dissertation "Англоязычный кинодискурс в лексикографическом отражении" N.A.Shamova analyzes the lexicographic changes and concepts of the corpus of film discourse in English. The concept of linguist discourse mainly deals with the meaning of word meanings, communicative changes, lexicography analyzes the words used in the language and their meanings, external and internal features, stylistic dependence and pragmatic features. The Corpus believes that it can fulfill a functional role in the implementation of this practice. In addition, the study of lexicogrammar, that is, the influence of lexical factors in explaining grammar, is also reflected in the scientific article of Uzbek linguists professor H.Dadaboyev and N.Abdurakhmonova. According to him, the corpus is the most correct way to explain grammar…" it is emphasized. Linguist E. Khannazarov distinguishes between annotated and non-annotated types of corpora in his research work, and emphasizes that the term "annotation" is also used in the meaning of the term "annotation" in corpus linguistics, and the value of the corpus is related to its perfect annotation. In turn, it is mentioned that this can lead to the increase in the efficiency of practical use of the language and the integration of corpus technologies with linguistics. Therefore, natural language processing cannot be done without corpus technology. Corpus lexicography is distinguished by its importance in NLP. It has become an urgent task to consistently conduct research on NLP and language technologies in the creation and development of electronic corpora of world languages. Currently, the research of terminology with the help of a web corpus is a cross-cutting task in the field of linguistics. For such important areas as "Corpus based turn extraction, Corpus based information extraction, Corpus based machine translation", the lexicographic feature of the corpus comes to the fore. Because statistical analysis plays a key role in the work of the Corps. The corpus played the role of an instrument in the creation of educational dictionaries of the language and served as a basis for the creation of the Oxford, Cambridge, Kobild dictionaries. Its essence is that it played a very important role in the creation of concordance and statistical data in the texts. Because the British National Corpus and Brown and Corpus of Contemporary American English (COCA) corpora were used to create a language concordance dictionary and search for words according to the N-gram model to determine the modern features and properties of the English language. Nowadays, they are recognized as a modern electronic case. They are valuable resources for linguists and other professionals.

---

[17] И.Саженин. Корпусные методы в лексикографии: опыт создания модели Словарного корпуса. Автореферат дисс. 2013 С.
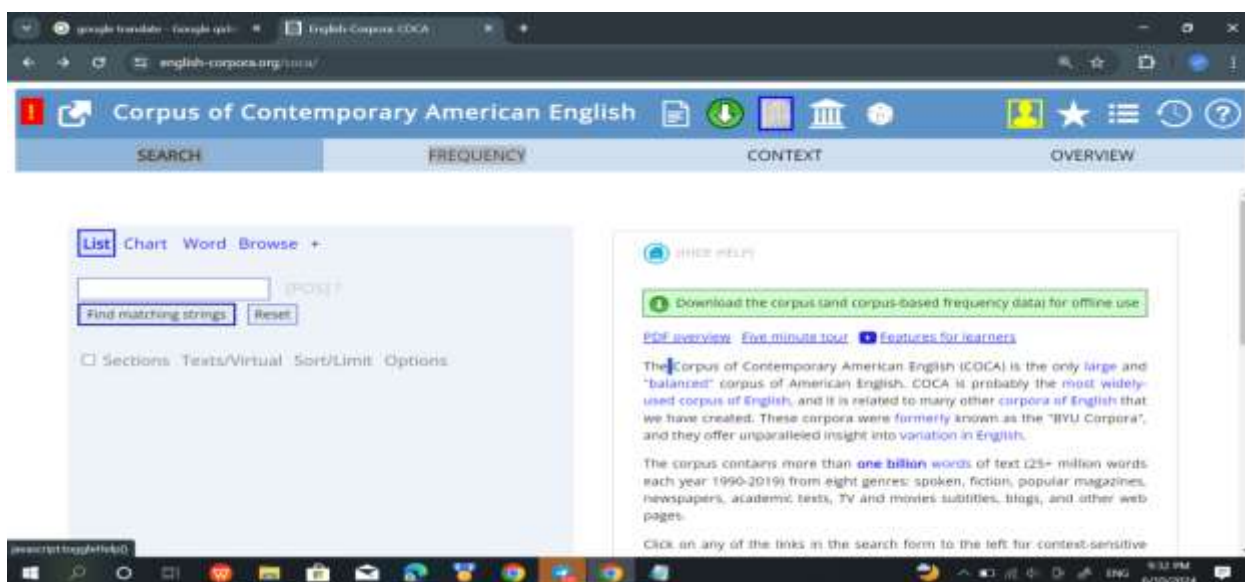
Image 2. A view of Coca's home page. We can see that there is a search menu.

Image 3. Statistical table of the collected texts in the Text section of COCA

Conclusion

In short, corpus lexicography is a combination of corpus linguistics and lexicography, which deals with the construction and analysis of vocabulary using corpora (large text databases) to determine the actual use of language. Corpusgai dictionary databases provide the following possibilities:

1. Determines how often words are used. This is important when building a vocabulary, and more attention is paid to the most common words.

2. Analyzes how words are used in different contexts, identifying and classifying their different meanings.

3. Learns how different expressions and phraseological units are represented in corpora.

4. Observes how the language changes over time and analyzes the process of emergence of new words and phrases.

5. Analysis of various linguistic trends and their causes.

REFERENCES

1. Abduraxmonova N. Computer models of Uzbek electronic corpus (monograph). / Toshkent:Muharrir, 2021, 202 p.

2. Abdurakhmonova N. Uzbek ontology of Uzbek language as example of adjective // Шестая Международная конференция по компьютерной обработке тюркских языков 363 "TurkLang-2018". (Труды конференции) – Ташкент: Издательско-полиграфический дом, 2018. – 320 с.

3. Ataboyev N. Problematic issues of corpus analysis and its shortcomings // ISJ Theoretical & Applied Science, 10 (78), 2019.; Ataboyev N. Compiling dictionaries by using corpus analysis and its advantages // International Journal of Progressive Sciences and Technologies (IJPSAT) http://ijpsat.es/index.php/ijpsat/article/view/508

4. N.B.Ataboev. Main features of Corpus Linguistics. www.journal.fledu.uz/ "Foreign Langugaes in Uzbekistan". Scientific-methodical electronic journal.№2-2019. б. 38.

5. Laviosa S. Corpus-based tran Corpus-based translation studies: Where does it come from? Where is it going? Studies in the Languages of Africa. Vol., 35, 2004 - Issue 1. 2008. P. 6-27.

6. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. – Cambridge University Press, 1998. – P.300. https://books.google.co.uz/books?id=2h5F7TXa6psC&printsec=frontcover#v=onepage&q&f=false

7. Sinclair J. 1991. Corpus and text-basic principles. – Oxford: Oxford University Press.

8. Teubert W. Corpus Linguistics and Lexicography. Article in International Journal of Corpus Linguistics. December 2001. https://www.researchgate.net/publication/240510906_Corpus_Linguistics_and_Lexicography

9. Isabel Durán-Muñoz & Gloria Corpas Pastor. Corpus-based lexicographic resourses for translators: an overview. 2019 - wlv.openrepository.com.

10. Hurskainen A. New Advances in Corpus-based Lexicography. https://hdl.handle.net/10520/EJC60458. 2003.

11. Laviosa S. Corpus-based tran Corpus-based translation studies: Where does it come from? Where is it going? Studies in the Languages of Africa. Vol., 35, Issue 1. 2008. P. 6-27.

12. Prinsloo D.J. Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus/ Article. Journ.: Lexikos / Vol. 25 (2015)

13. Teubert W. Corpus Linguistics and Lexicography. /Journ.: Corpus Linguistics. P. -137.