Scopus-based bibliometric analysis on corpus linguistics for the period of 2017-2021

B.R. Mengliye[1*], *Sh.* Hamroyeva[1], and *O.* Abdullayeva[1]

[1]Tashkent State University of Uzbek Language and Literature named after Alisher Navoi, 100100 Tashkent, Uzbekistan

Abstract. This article aims to observe the latest scientific theories in the field of corpus linguistics, to analyze the latest research trends in corpus linguistics and the creation of language corpora. The results of our research are based on bibliometric analysis of scientific research results and review articles of universities, scientific research centers and well-known scientists of different countries where scientific and practical work is being carried out in the field of corpus linguistics. We analyzed the publications in the Scopus database in the field of corpus linguistics in 2017-2021 and found research results related to finding solutions to various problems in language corpora and problems in it and we observed bibliometric method through speech recognition, syntactic parsing problems, semantic tagging problems, automatic tokenization and lemmatization. This is the first research in Uzbek linguistics to report on the landscape of corpus linguistics in recent years. This research contributes to the general scientific understanding of corpus linguistics and provides insight into the past, present, and future of linguistics. 1353 publications were analyzed in the article. Although the field of corpus linguistics originated in the 1960s and 1970s, the fields of study have expanded and changed over time. Among the fields of linguistics, this direction is dynamic. In recent years, national corpora and target corpora have been created in various languages, and solutions to complex linguistic problems have been found.

1 Introduction

Corpora consist of large amounts of data and are used in natural language processing and in teaching languages [1]. In almost half a century of corpus linguistics, many research papers have been published. It has been shown that many research results have been related to the construction of language corpora, the use of language corpora in language learning, the importance of corpora for creating dictionaries, the use of corpora in the implementation of comprehensive linguistic analysis and synthesis of texts of various styles and genres, and the solution to other linguistic issues in the language through corpora. [1-7]. Especially in recent works, it has been researched whether a speech unit in language corpora gives the semantics of mutual synonym or homonymy in certain places according to contextual analysis, and it is possible to semantically annotate large texts in a short period of time [8].

* Corresponding author: bakhtiyormingliyev62@gmail.com

In recent years, a new direction of research such as the analysis of the corpus and related studies based on the data obtained by the bibliometric method has appeared. Among such studies, H. Park's article on citation analysis based on 5,600 scientific articles indexed in WOS in 1997-2016 and 172,352 citations in them [9], X. Lui's article on corpus linguistics in Chinese in 1998-2013 [10] includes bibliometric research based on the general analysis of the conducted researches, as well as, Sh.Liao and L.Lei conducted a bibliometric analysis of corpus-related research in 2000-2015 [11].

H. Park aims to solve 3 main questions in the analysis of research carried out in the field of corpus linguistics in 20 years: 1) What famous results have been realized in the last 20 years; 2) What issues have been discussed in the research carried out in the last 20 years? 3) How did problematic issues arise during the studied period and what solutions were found?

In the research of Sh.Liao and L.Lei, annual results in articles related to the corpus from 2000 to 2015, results of research carried out in countries, high publications and citations were analyzed using the bibliometric method. As a result, the development trend of the industry has been determined. Through this study, we have discussed bibliometrically 1353 research papers indexed in the Scopus database in the field of corpus linguistics in 2017-2021.

2 Experiment technique

All the data in the article is collected from the Scopus database. As we have already analyzed scientific articles of previous years based on data from various databases, we decided to take the last five years. We performed a bibliometric analysis based on the database of 2017-2021. A total of 1,353 corpus research publications were reviewed and analyzed using the terms corpus linguistics, language corpora as keywords in the search strategy. In our research, the countries with the most published research results in various fields, the famous scientists, universities and organizations that carried out the research, various sources indexed in Scopus, corpus linguistics and various other fields classified as including have their percentage of publications published in the last five years. Figure 1 shows the methodology flow chosen for the research. Analyzes were shown through various diagrams and visualization method in VOSviewer.

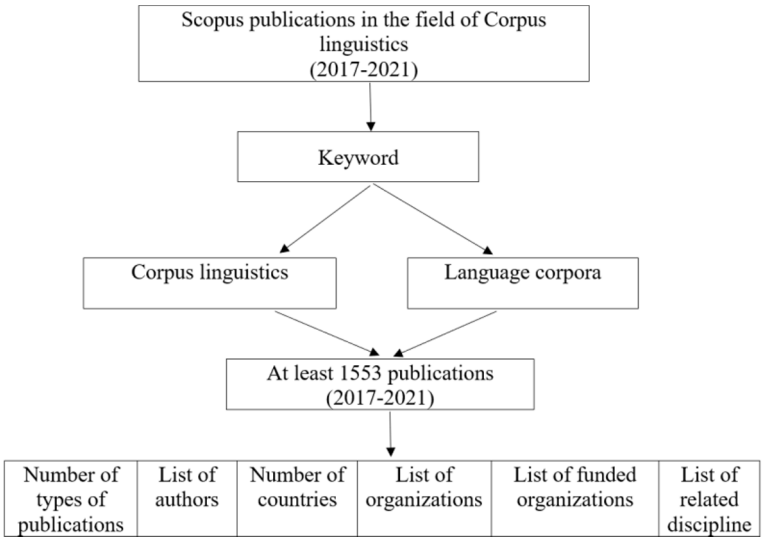

Fig. 1. Scheme of methodology for research

3 Results and discussion

In the bibliometric analysis of research results, we analyzed the number of publications in five years in different directions. The results of the annual survey of corpus linguistics publications are presented in Figure 2. The number of studies published and indexed in the Scopus database was 190 in 2017, but by 2018, the number of studies has increased sharply, reaching 316. The top indicator can be observed in 2020 as well.

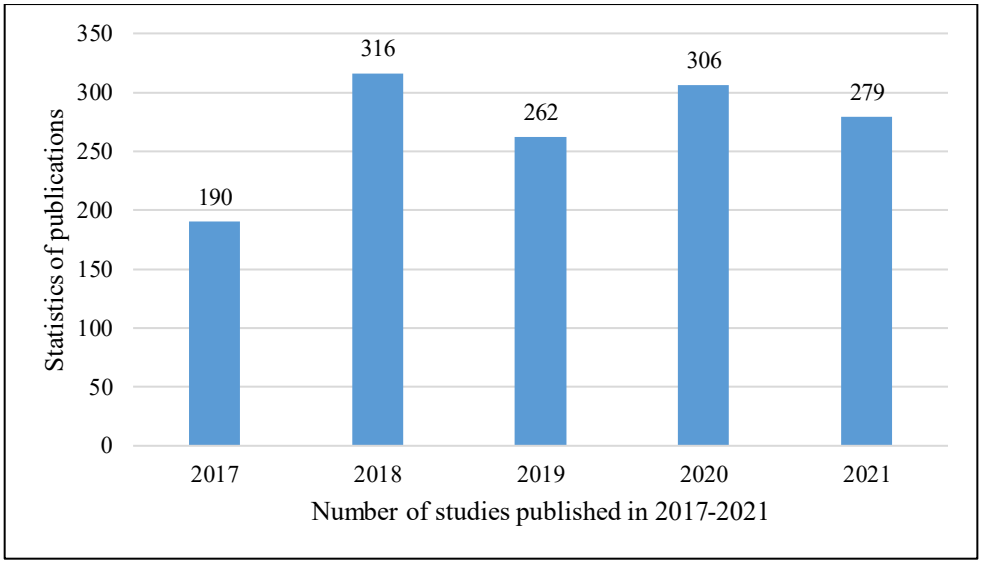

Fig. 2. Number of publications (by years)

The growth trend of research in the field of corpus linguistics can be seen in the creation of language corpora and the study of problems related to it, and its application to various related fields. Because research work related to corpus-based analysis, natural language processing, artificial intelligence, discourse analysis has grown rapidly. The most important keywords in the 1353 analyzed works were studied and their interrelationships were shown in Figure 3 through the visualization method. The term computational linguistics, which covers a wide range of research works, occupies the highest percentage, is the basis of various scientific researches, and its interconnection has become complex.

When corpus linguistics and the most important directions related to it were bibliometric analyzed as research topics in the visualization method (Figure 4), computer linguistics was considered one of the effective research topics due to the high percentage. If the topics of linguistics, natural language processing, and semantics are considered next, then the topics of speech recognition, artificial intelligence, syntax, ontology, and discourse analysis can be considered as a new research direction.

In the last five years, the total number of countries whose research results have been published in the Scopus database on the problems of corpus linguistics is 67. We have selected the top 11 countries with no fewer than 10 publications in Figure 5. Among the countries, the USA is in the highest place and owns 82.6% of publications. The statistics on the top 10 of the index showed that the USA (1118), Russia (196), China (66), Turkey (57), Great Britain (54), Germany (46), Canada (24), India (24), Israel (18), France (14) are leading in the number of publications. According to the bibliometric results of the last five years, among the Central Asian countries participating in Scopus with indexed publications are Uzbekistan (3), Kazakhstan (2), Tajikistan (2).

Scientists have a special role in researching scientific problems in the field and publishing the results. In the years determined by our research, 160 authors published articles related to the field of computational linguistics and corpus linguistics. In the last five years, Kai-Wei Chang, a scientist whose research on natural language processing and machine learning problems is indexed in Scopus, has published 13 articles. This is the highest indicator. Ryan Cotterell published 12 articles on language corpora, lemmatization, morphological tagging, and Graham Neubig published 12 articles on natural language processing and machine translation. We can see Tatyana Litvinova from Russian scientists in the field, she has published 10 articles. Figure 6 shows 15 researchers who published 7-10 articles.

We can see 160 universities and scientific research institutes that published research results in 2017-2021 on the issues of computer science, natural language processing, and artificial intelligence. Figure 7 shows lists the top 15 countries. In terms of the number of articles published by Carnegie Mellon University in the last five years, 6.87%, Saint Petersburg State University – 4.13%, and Google LLC – 3.25% have the highest index. American and Russian universities and research institutes are leading the way in conducting and publishing the most important research in the field.

When studying the organizations that financed the industry in 2017-2021, their statistics were shown in Figure 8. The National Science Foundation funded 226 research results, accounting for 16.7%, the Agency for the Support of Advanced Research Projects funded 73 studies (5.39%), the Russian Foundation for Fundamental Research funded 61 studies related to the field (4.50%). In the bibliometric analysis of the top 10, it can be observed that organizations supporting scientific research in America and Russia have taken place.

We have analyzed that the researches carried out in the field of corpus linguistics were carried out within the framework of journals of different directions. In 2017-2021, we witnessed the publication of a total of 20 scientific journals. We selected the top 10 journals with the highest publication rate for analysis. 1087 research results were published in Computer Science journal and indexed in Scopus. As a result, it is 80% of the last five years. 584 works (43.1%) in Art and Humanities journal, 566 (41.8%) in Social Sciences journal, 181 (13.35%) in Mathematics journal, Engineering 105 (7.7%) in the journal, 50 (3.6%) in Medicine. The percentage of research published in other journals is not high.

Finally, we analyzed the types of research published in the last five years in the field of corpus linguistics. It was observed that there are 7 types of publications in the field. The most common type of publication is conference papers, with 1084 published articles. Research results were published in 233 scientific journal articles, 17 reviews, 14 book chapters, 2 books, 2 notes and 1 short survey

4 Conclusion

The field of corpus linguistics is one of the most crucial topics in linguistics and related disciplines. In our research, we tried to analyze the results of research in corpus linguistics in the last 5 years based on the Scopus database. We identified important aspects through bibliometric analysis. First, the results of research in the field began to increase extreme, this indicator was clearly visible in 2017 and 2018. We observed that clustering in the published results was done within the topics of Computational Linguistics, Linguistics, Corpus Linguistics, Natural Language Processing, and Speech recognition. America and Russia are leading among 67 countries. It is possible to see that the results of the scientific research carried out by the scientists of Carnegie Mellon University and Saint Petersburg State University in these countries are increasing gradually. Next in line are China and European countries with a high index. Among 160 scientists in 2017-2021, K.W.Chang, R.Cotterell, G.Neubigs are the leading authors. We got the statistics of the best journals

according to the number of published articles. In the analysis of 20 journals, it was confirmed that the Computer Science journal has the highest index, accounting for 80% of the published articles. According to the obtained bibliometric results, it can be observed that corpus linguistics is becoming increasingly popular not only in computer linguistics and natural language processing, but also in speech understanding, artificial intelligence, discourse analysis, ontology, and thesaurus. Here, the results of scientific research in the field of corpus linguistics had a significant impact on the grammar of languages, which we observed in the published research results on the syntactic and semantic levels. The data used in our research is limited to the Scopus database. The results of such analysis allow better understanding, and comparison of the development trend of the tendency.

References

1. D. K. Graeme, An introduction to corpus linguistics, London, Longman (1998)
2. J. Sinclair, Corpus concordancer collocation, UK, Oxford University Press (1991)
3. G. Leech, Corpora and theories of linguistic performance. In Jan Svartvik (ed.), Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, Berlin: Mouton de Gruyter, 105-122 (1992)
4. T. McEnery, C. Gabrielatos, English corpus linguistics. In B. Aarts, A. McMahon (eds.), The handbook of English linguistics, Oxford: Blackwell Publishing, 33-71 (2006)
5. T. McEnery, A. Hardie, Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press (2012)
6. P. Hanks, The corpus revolution in lexicography. International Journal of Lexicography, **25(4)**, 398-436 (2012)
7. A. O'Keeffe, M. McCarthy, R. Carter, From corpus to classroom: Language use and language teaching, UK, Cambridge University Press (2007)
8. S. Nirenburg, V. Raskin, Ontological semantics, Cambridge: MIT Press (2004)
9. H. Park, D. Nam, Corpus linguistics research trends from 1997 to 2016: A co-citation analysis. Linguistic Research, **34(3)**, 427-457
10. X. Liu, J. Xu, L. Liu, An overview of corpus linguistics research in China (1998-2013): A CiteSpace based analysis. Corpus Linguistics, **1(1)**, 69-77 (2014)
11. S. Liao, L. Lei, What we talk about when we talk about corpus: A bibliometric analysis of corpus-related research in linguistics (2000-2015)
12. A. Boulton, T. Cobb, Corpus use in language learning: A meta-analysis. Language Learning, **67(2)**, 348-393 (2017)
13. P. Crosthwaite, S. Ningrum, M. Schweinberger, A bibliometric analysis of two decades of Scopus-indexed corpus linguistics research in arts and humanities, International journal of Corpus linguistics (2022)