Л.Н. Гумилев атындағы Еуразия ұлттық университетінің ХАБАРШЫСЫ.
ISSN (Print) 2616-678X. ISSN (Online) 2663-1288

ЛИНГВИСТИКА / LINGUISTICS / ЛИНГВИСТИКА

DOI: https://doi org/10.32523/2616-678X-2026-155-2-119-136

IRSTI 16.31.21
Review article

A.N. Oraz[1*], K.T. Malikov[2], N.S.Khalikova[3]
[1, 3] O. Zhanibekov South Kazakhstan Pedagogical University, Shymkent, Kazakhstan
[2]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
*(E-mail: [1]*Aydana.oraz.1997@mail.ru, [1]k.malikov@mail.ru, [3]khalikova_nurila1978@mail.ru)*

BIBLIOMETRIC ANALYSIS OF SCIENTIFIC PUBLICATIONS ON THE KEYWORD "CORPUS LINGUISTICS": BASED ON SCOPUS DATA (2015-2025)

Abstract. This study aims to identify development trends, structural characteristics, and major research directions in scientific publications in the field of corpus linguistics. The purpose of the research is to determine the scientific productivity of the field, publication dynamics, leading authors and countries, patterns of international scientific collaboration, as well as dominant thematic trends by conducting a bibliometric analysis of publications on corpus linguistics published between 2015 and 2025 based on the Scopus database. Bibliometric analysis was used as the main research method. Data were retrieved from the Scopus database using the keyword "corpus linguistics," resulting in 23,073 documents; after applying a time filter (2015–2025), 16,475 documents were downloaded, and following data cleaning with the Bibliometrix software, 16,402 scientific documents were included in the analysis. Bibliographic data were processed in the RStudio programming environment using the Bibliometrix package and the Biblioshiny interface. The analysis examined the annual dynamics of publications, author and country-level collaboration, the most frequently used keywords, and thematic networks. The results indicate rapid growth in corpus linguistics publications, expanding international collaboration, and a shift toward computational approaches. The study systematizes current trends and provides a foundation for future research.

Keywords: corpus linguistics; bibliometric analysis; scientific publications; Scopus database; research trends; thematic evolution; international scientific collaboration.

Received: 08.02.2025; Revised: 23.05.2026; Accepted: 15.06.2026; Available online: 30.06.2026

Introduction

Corpus linguistics is one of the fields in linguistics whose foundations were laid in the mid-twentieth century and which has developed rapidly in recent decades as a result of advances in computer technologies. The term linguistic corpus emerged in the 1960s and was subsequently used to denote the compilation of large electronic language databases based on the machine processing of textual data. This approach opened up new possibilities for analyzing the structural and functional characteristics of language, as linguistic corpora enable precise, context-rich, and statistically grounded analyses of linguistic units. Corpus-based methods allow researchers to identify language structures in real contexts, patterns of lexical and grammatical usage, and discourse features through statistical and computational techniques. Therefore, this approach differs fundamentally from traditional intuitive analysis. The development of corpus methods, which began in the mid-twentieth century, evolved into a new linguistic paradigm often referred to as the "corpus revolution," introducing empirically based research as a core methodology in linguistic studies (Plungian, 2024).

Corpus linguistics contributes not only to the theoretical foundations of linguistic data analysis but also significantly advances applied aspects of linguistics. In particular, it plays an important role in language teaching and the development of instructional materials by emphasizing frequency, context, and structural patterns based on authentic empirical data. This enables learners to work with language materials that reflect real communicative situations (Crosthwaite, Ningrum, & Schweinberger, 2023). Studies focusing on the application of corpus methods in linguistics include, for example, research on corpus-based approaches to grammar instruction, which examines the relationship between corpora and language teaching, pedagogical applications, and improvements in learning outcomes (Alamri, 2022). A notable example is Alamri's (2022) work titled The Role of Corpus Linguistics in Grammar Instruction, which demonstrates the effectiveness of corpus-based methods in grammar teaching through an extensive literature review. Alamri shows that corpus data allow instructors to present materials grounded in authentic language use, facilitating the acquisition of language forms based on actual communicative practice rather than purely intuitive approaches. The study highlights three main advantages of corpus-based approaches. First, corpus data reveal the actual frequency of grammatical structures, enabling instructors to prioritize forms that are most commonly used in context, thereby making language learning more natural and effective. Second, corpus analyses demonstrate the contextual use of grammatical structures, helping learners not only memorize rules but also apply them in real communicative situations. Third, corpus methods reveal the variability and complexity of grammatical features, allowing learners to flexibly use language across different patterns of speech. In addition, corpus analysis extends beyond grammar to examine word meanings in context and collocational patterns, identifying frequency-based regularities and supporting evidence-based approaches to language teaching.

An important contribution in this area is the study by Gapporov Baxriddin Baxtiyor Ugli titled The Role of Corpus Linguistics in the Study of Collocations (Gapporov, 2025), which systematically analyzes the structure, distribution, and contextual usage of lexical units. The author argues that corpus-based approaches enable researchers to identify collocations based on frequency, distribution, and context using empirical evidence rather than intuition. The use of corpus platforms and analytical tools makes it possible to determine word frequency and opens pathways for functional and discourse analyses. In particular, the study focuses on verb-based collocations and examines their significance in discourse texts.

In recent years, research directions in corpus linguistics have expanded considerably, and large-scale bibliometric analyses have become increasingly common across linguistics and the humanities. For instance, Crosthwaite, Ningrum, and Schweinberger (2022) analyzed publications indexed in the Scopus database over the past 20 years to identify key trends and developments in corpus linguistics (Crosthwaite et al., 2023). Other studies have conducted comparative analyses of recent trends in corpus research based on Scopus data, describing indicators such as publication volume, major research areas, popular topics, and the geographical distribution of research output. Based on data from 5,829 scientific articles, Crosthwaite, Ningrum, and Schweinberger (2022) demonstrate the multilingual scope, diverse research directions, and wide geographical coverage of corpus linguistics, noting in particular the growth of studies in Chinese, Russian, and Italian. Their work shows that corpus research extends beyond the analysis of linguistic structures within the Arts and Humanities to include applied aspects such as lexical bundles, academic writing, and multilingualism, highlighting the applied and transdisciplinary nature of corpus linguistics. This field is therefore not limited to theoretical linguistics but also addresses practical issues in language teaching, text analysis, and the comparison of linguistic materials. As is well known, corpus linguistics provides a strong statistical foundation for the application of research methods across various domains, distinguishing it from traditional descriptive approaches to language. Such analyses reveal an increasing methodological diversification and a deliberate expansion of research objects in contemporary corpus studies.

Today, corpus linguistics has become one of the leading directions of empirical and data-driven research in linguistics, enabling statistically robust analyses of morphological, lexical, syntactic, and discourse patterns through computational corpus data. The significant increase in scientific publications in recent years indicates the rapid development of corpus linguistics within the international academic community. Previous studies in this area have generally examined development trends, author distributions, and research directions through broad bibliometric reviews. The distinctive contribution of the present study lies in its comprehensive analysis of publications indexed in the Scopus database under the keyword “corpus linguistics” between 2015 and 2025. Specifically, the study examines annual publication dynamics, the most productive sources, the top ten most prolific authors during this period, changes in author productivity over time, the level of international collaboration, the identification of the most influential publications based on citation metrics, the visualization of dominant terms, the evolution of thematic trends, keyword co-occurrence networks and thematic clusters generated using Bibliometrix, thematic maps of corpus linguistics research, and international collaboration maps between countries. The results of this analysis provide an empirically grounded understanding of current research trends and future development trajectories in corpus linguistics. Accordingly, the aim of this study is to conduct a bibliometric analysis of scientific publications in the field of corpus linguistics published between 2015 and 2025 based on the Scopus database, in order to identify the field’s scientific productivity, publication dynamics, leading authors and sources, patterns of international scientific collaboration, and the evolution of key themes and research directions.

Materials and methods

This study is based on a bibliometric analysis aimed at identifying the development trends, structure, and main research directions of scientific publications in the field of corpus linguistics. Bibliometric analysis is a method that allows the quantitative study of scientific literature and

is widely used to determine the productivity, impact, collaboration patterns, and thematic evolution of a given scientific field. It is currently recognized as a standard tool for systematically evaluating scientific literature, enabling the quantitative description of publication flows, interconnections, and trends within a discipline (Roslim et al., 2023; Crosthwaite et al., 2022).

The data for this study were retrieved from the Scopus database. Scopus is a reliable source frequently used in bibliometric studies due to its broad multidisciplinary coverage and indexing of high-quality peer-reviewed journals and conference proceedings. Scopus serves as a tool for monitoring high-level publications and assessing research quality and scientific development trends, as its indexed materials undergo strict selection and peer review according to international standards. The significance of the Scopus database lies in its comprehensive bibliometric information: it provides detailed metadata on articles, authors, institutions, and journals, as well as indicators for analyzing citations and publication dynamics. Such a broad and high-quality dataset allows researchers to quantitatively evaluate scientific trends, changes in trends, and international activity of authors and publications, ensuring the objectivity and reliability of bibliometric analysis (Baas et al., 2020).

The data collection and retrieval process was carried out in January 2026. Thus, the dataset includes publications indexed in Scopus up to the end of 2025. Searches were conducted using the keyword “corpus linguistics.” The initial search identified 23,073 documents. To align with the study’s objectives and to accurately reflect current research trends, publications were filtered to the period 2015–2025. After applying this temporal filter, 16,475 documents were selected, and following data cleaning with the Bibliometrix software, 16,402 scientific documents were included in the analysis.

No prior restrictions were imposed on the types of publications (e.g., journal articles, conference papers), as conference publications also contain significant scientific contributions in the field of corpus linguistics. All data were downloaded in BibTeX format for analysis. The bibliographic data were processed using the R programming environment within the RStudio interface. Bibliometric analysis was conducted using the Bibliometrix package and its graphical interface, Biblioshiny. Bibliometrix is a powerful R-based tool for comprehensive analysis of scientific publications, enabling the identification of scientific productivity, citation structures, author and institutional collaboration, thematic networks, and research trends.

The Biblioshiny platform provides an interactive environment for analysis, visualization, and systematic description of bibliometric indicators, ensuring the accuracy and reproducibility of the study results. During the analysis, Google Translate, an AI-based platform, was used solely as an auxiliary tool to maintain grammatical accuracy in English texts and assist with partial content translation. This tool did not directly influence the interpretation or analysis of the scientific content. The study is entirely based on openly accessible secondary data. No human participants, experimental interventions, or collection of personal data were involved; therefore, ethical committee approval was not required..

Results and Discussion

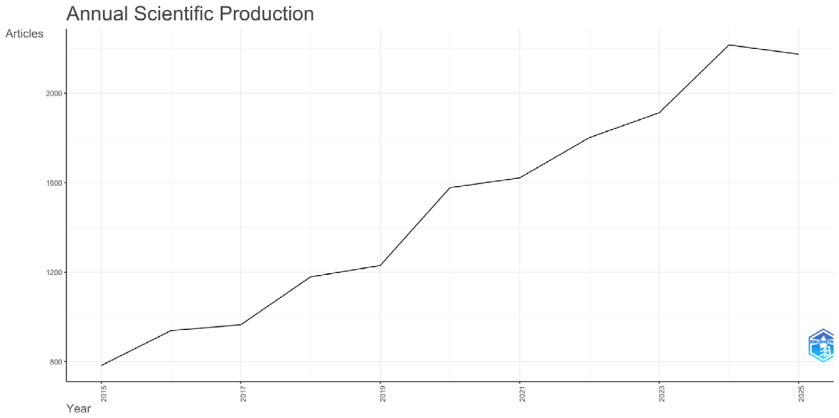
A total of 16,402 documents indexed in the Scopus database under the keyword “corpus linguistics” between 2015 and 2025 were included in the analysis. These publications appeared across 2,667 sources, with a total of 25,788 authors registered in the dataset. The covered period spans the last decade, with an average document age of 4.99 years. The average annual growth rate of publications was 10.76%.

By document type, the dataset included 10,063 articles, 3,784 conference papers, 1,326 book chapters, and 615 review articles. The average number of authors per document was 2.47,

with 6,096 single-author publications. Publications resulting from international collaboration accounted for 16.36% of the total. The average number of citations per document was 7.709.

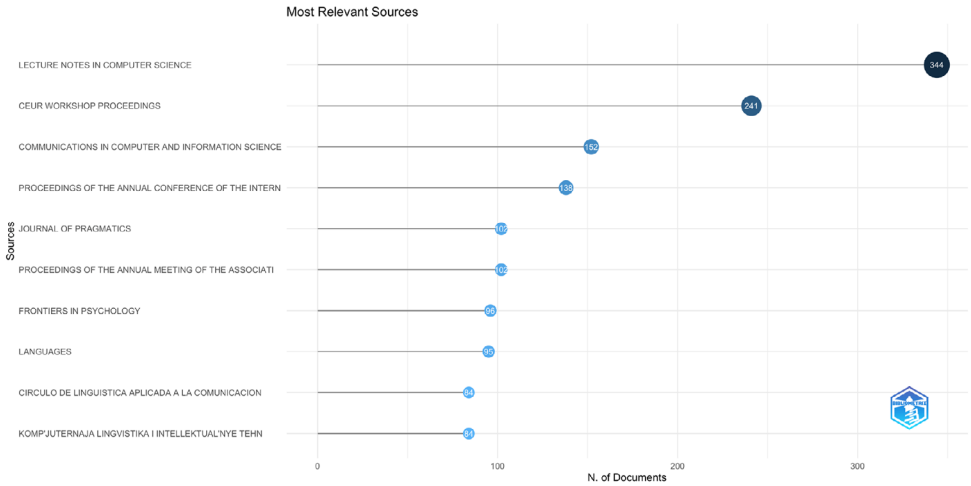
The annual publication dynamics between 2015 and 2024 are summarized in Table 1. In 2015, there were 783 publications, which increased to 939 in 2016. In 2017, the number of publications reached 965, followed by 1,179 in 2018, and 1,230 in 2019. A noticeable growth occurred in 2020, with 1,578 publications, followed by 1,622 in 2021, 1,802 in 2022, and 1,913 in 2023. The highest number of publications was recorded in 2024, totaling 2,216 articles. Overall, the results indicate a consistent year-on-year increase in the number of publications during the studied period (Figure 1).

Figure 1. Annual Publication Dynamics in Corpus Linguistics Based on Scopus Data (2015–2025)



In the research corpus, Figure 2 shows the number of publications by the most productive sources. The source with the highest number of documents is Lecture Notes in Computer Science (344 documents). This is followed by CEUR Workshop Proceedings (241) and Communications in Computer and Information Science (152). Proceedings of the Annual Conference of the Intern… contains 138 publications. Journal of Pragmatics and Proceedings of the Annual Meeting of the Association… recorded 102 documents each. Frontiers in Psychology (98), Languages (95), Círculo de Lingüística Aplicada a la Comunicación (84), and Komp'yuternaâ Lingvistika i Intellektual'nye Tehn… (84) are also among the leading sources (Figure 2).

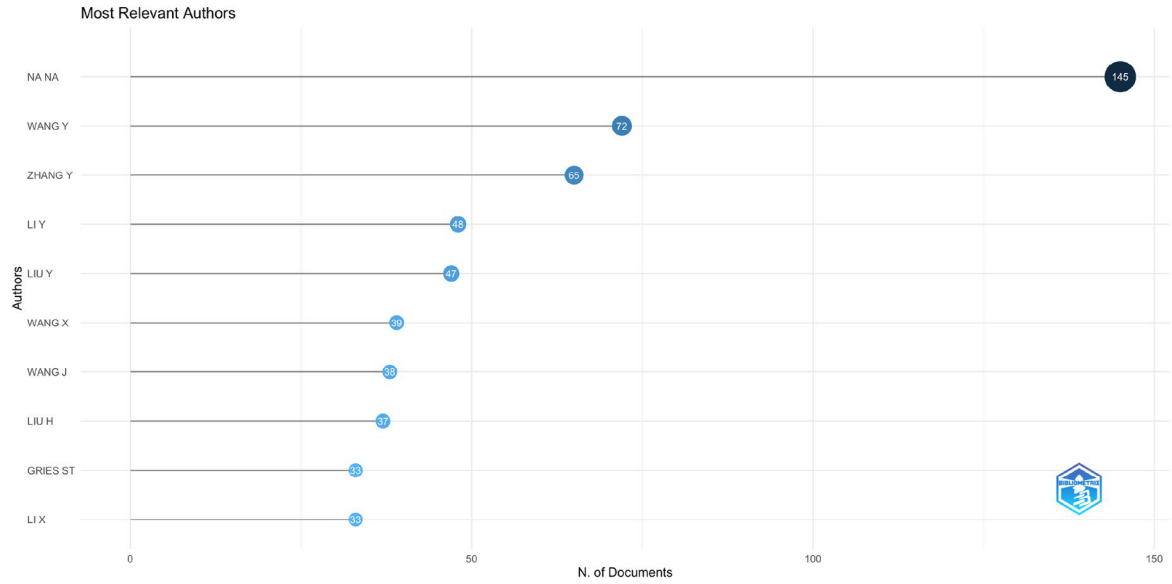
Figure 2. Most Productive Sources Based on the Keyword “Corpus Linguistics”



In Figure 3, the authors with the highest number of publications on the topic of corpus linguistics between 2015 and 2025 are shown. The diagram presents the most frequently occurring author names in the database and the number of documents attributed to them. The highest count is associated with entries labeled as Na Na (145 documents). However, this figure represents bibliographic records where the author name was incomplete or inconsistent. Therefore, this data does not accurately reflect the scientific productivity of an individual author and indicates the need for data cleaning.

Among the clearly identified authors, Wang Y has the highest number of publications (72 documents). This is followed by Zhang Y (65 documents), Li Y (48 documents), and Liu Y (47 documents). Additionally, Wang X (39), Wang J (38), Liu H (37), Gries St (33), and Li X (33) are included in the top ten authors by publication count. Overall, the figure shows that authorship productivity is unevenly distributed, with a limited group of authors contributing a high number of publications. Furthermore, the use of abbreviated or inconsistent author names highlights the issue of author identification in bibliometric datasets (Figure 3).

Figure 3. Top 10 Most Productive Authors in the Field of Corpus Linguistics (2015–2025)

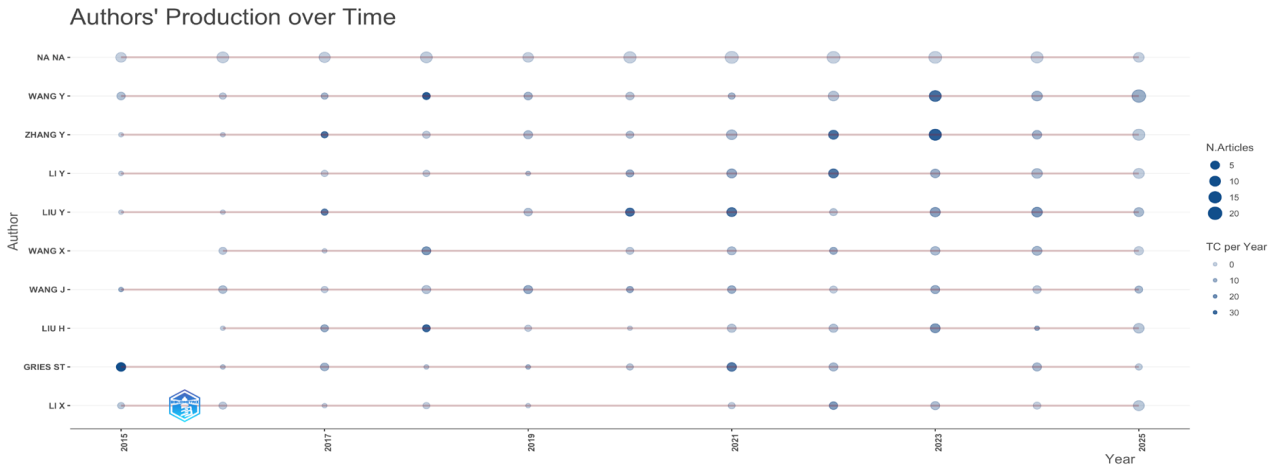


The figure below shows the publication activity and citation indicators of the most productive authors based on the keyword “corpus linguistics” between 2015 and 2025. The dotted points along the straight line in the diagram represent the number of publications in a given year, while the color intensity indicates the number of citations in that year (TC/year). The red lines indicate the period of active publication for each author.

The graph shows that the publications of Gries St, Liu H, Wang J, Wang X, Liu Y, Li Y, Zhang Y, Wang Y, and Li X are unevenly distributed over the years. Some authors had high publication counts and citation indicators in certain years, while in other years these metrics were low. A number of authors show more pronounced activity after 2020.

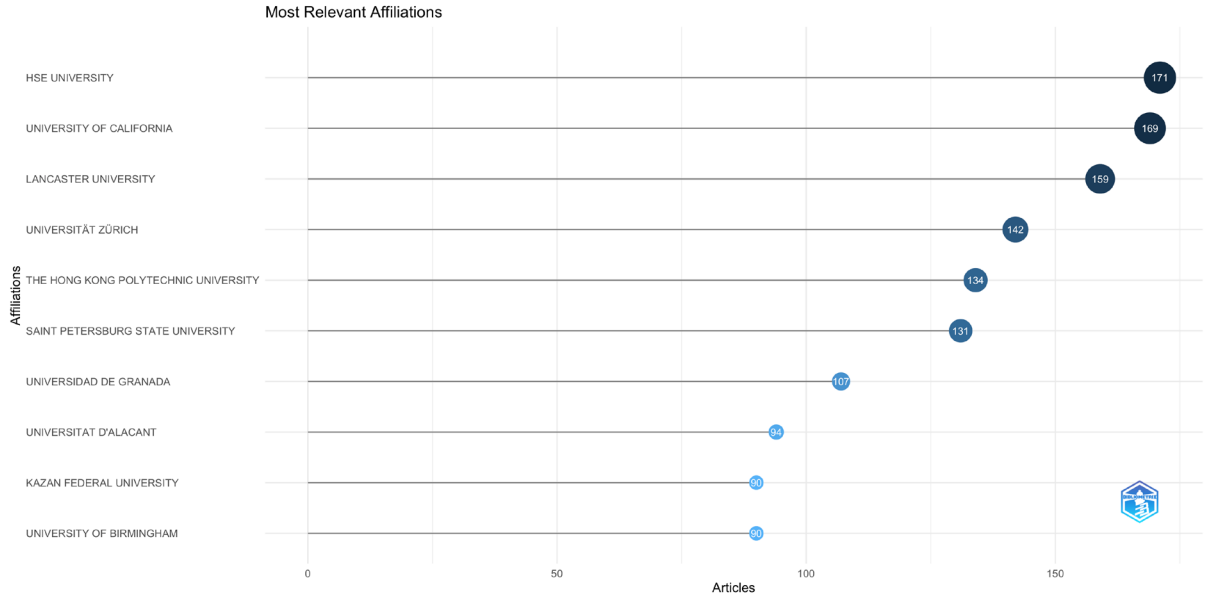
Additionally, entries labeled as “Na Na” appear across all years, but their citation indicators are very low or zero. This data represents bibliographic records where author information was incomplete or not disambiguated, rather than reflecting the productivity of an individual author. Overall, the graph illustrates the temporal dynamics of publication activity and citation counts for leading authors during the studied period (Figure 4).

Figure 4. Author Productivity Over Time



The diagram below shows that the highest number of publications in the studied dataset belongs to HSE University (171). This is followed by the University of California (169) and Lancaster University (159). Universität Zürich (142), The Hong Kong Polytechnic University (134), and Saint Petersburg State University (131) also stand out with high publication counts. Among other institutions, Universidad de Granada (107), Universitat d'Alacant (94), Kazan Federal University (90), and University of Birmingham (90) are shown (Figure 5).

Figure 5. Top 10 Universities Based on the Keyword "Corpus Linguistics" (2015–2025)



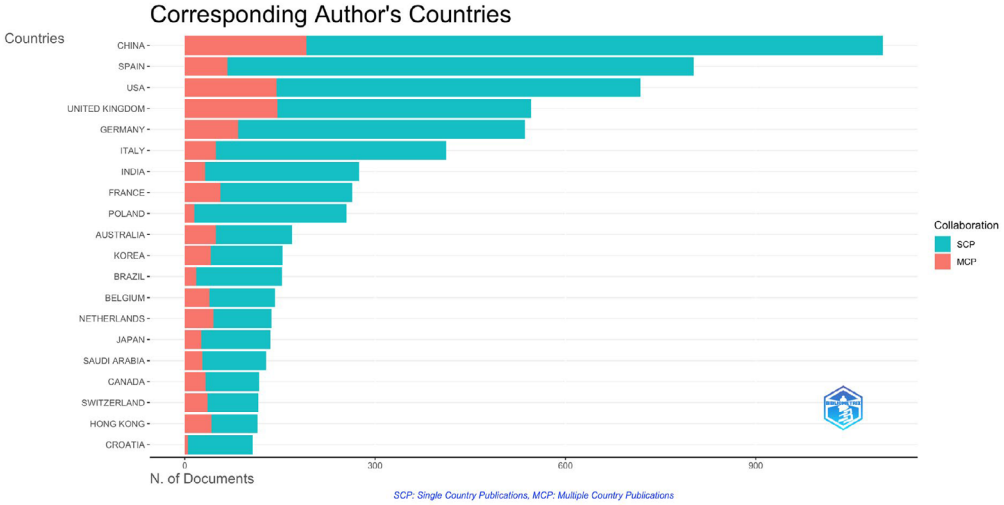
The data shown in the figure are based on the SCOPUS database. Here, the productivity of publications and the level of international collaboration are presented according to the country affiliation of corresponding authors. According to the analysis, the highest number of publications belongs to China (1,100 articles), followed by Spain (802) and the USA (718). The top five are completed by the United Kingdom (546) and Germany (536).

The level of international collaboration is measured using the Multiple Country Publication (MCP) indicator. The highest MCP shares are observed in Hong Kong (36.5%), the Netherlands

(32.8%), Switzerland (31%), Australia (29%), Canada (28.2%), and Belgium (27.5%). Among high-productivity countries, the United Kingdom (26.7%) and the USA (20.2%) show relatively high levels of international collaboration.

In some countries, publications conducted within a single country (Single Country Publications, SCP) dominate. For example, Croatia (4.7%), Poland (5.9%), and Spain (8.4%) are characterized by a low proportion of internationally co-authored publications. This indicates that scientific leadership in these countries is primarily national. Overall, the diagram illustrates differences in scientific productivity and international collaboration between countries, as well as the relationship between national and global scientific interactions (Figure 6).

Figure 6. Level of International Collaboration Based on the Keyword "Corpus Linguistics" (2015–2025)



The Most Global Cited Documents analysis in the Bibliometrix software allows identifying the most influential scientific works in the research corpus based on citation metrics. This serves to show the scientific evolution of the research direction, its theoretical foundations, and the vector of technological development. Each column presented in the table describes a specific bibliometric measure. The "Paper" column provides the full bibliographic description of the publication, clarifying the scientific context of the work through the list of authors, publication year, and the name of the conference or journal. DOI (Digital Object Identifier) ensures direct access to the original publication as a permanent digital identifier. The Total Citations (TC) metric indicates how many times the work has been cited in databases, quantitatively reflecting its scientific impact. TC per Year shows the average number of citations per year, allowing assessment of whether the work has maintained its relevance over time or received particular attention during a certain period. The Normalized TC metric adjusts the citation count relative to the publication year, allowing a fair comparison of the scientific impact of works published in different years.

Content and quantitative analysis of the results showed that the highest citation metrics belong to works in natural language processing (NLP), neural language models, deep learning, semantic representation, and automatic processing of linguistic resources. In particular, the work by Peters et al. (2018), *Deep Contextualized Word Representations (ELMo)*, leads in all metrics, demonstrating the formation of a new NLP paradigm through contextual word vectors. This work is subsequently considered one of the theoretical foundations for large language

models such as BERT and GPT. Additionally, Petroni et al. (2019) showed that language models can retain hidden semantic knowledge, while the works of Warstadt et al. (2019) address the issue of evaluating grammatical knowledge through language models. Talat & Hovy (2016) examine the societal and ethical aspects, investigating the influence of NLP on society.

Applied research directions also show high citation metrics. Cheng & Lapata (2016) achieved significant results in neural text summarization; Clark et al. (2020) in multilingual question-answering systems (TyDi QA); Nguyen & Grishman (2015) in automatic relation extraction; Habibi et al. (2017) in biomedical named entity recognition. These works demonstrate the effective application of NLP technologies in various practical areas. Furthermore, studies such as Straka et al. (2016) introducing the UDPipe tool, Zeldes (2017) describing the GUM corpus, Pechenick et al. (2015) analyzing the Google Books corpus, and Kulkarni et al. (2015) on statistical detection of language change, illustrate the scientific significance of corpus linguistics and language resources. Atkins & Rundell (2008), as a fundamental work in lexicography, also shows high Normalized TC values, demonstrating that traditional linguistics remains influential.

In addition, works covering education, social media, genre classification, language change, and dependency length minimization are also represented in the table. This indicates that linguistics develops not only theoretically but also in close connection with social, educational, cognitive, and technological aspects. In particular, the study by Mizumoto & Eguchi (2023) on the use of artificial intelligence for automatic essay scoring highlights the recent integration of AI technologies into the education sector.

Overall, the Most Global Cited Documents results show that interdisciplinary works at the intersection of artificial intelligence, deep learning, NLP, corpus resources, and applied linguistics have particularly high scientific impact. The dynamics of citation metrics clearly illustrate the evolutionary development of research in linguistics, from classical lexicography and corpus studies to neural language models, and further to large language models and multilingual systems. These data indicate that the research direction is fully consistent with current scientific trends and that these technological directions will continue to play a leading role in the future (Table 1).

Table 1. Most Cited Research Publications

Paper	DOI	Total Citations	TC per Year	Normalized TC
Peters, Neumann, Iyyer, Gardner, Clark, Lee, & Zettlemoyer (2018).	https://doi.org/10.18653/v1/n18-1202	7299	811,00	400,91
Petroni, Rocktäschel, Riedel, Lewis, Bakhtin, Wu and Miller (2019)	https://doi.org/10.18653/v1/D19-1250	1678	209,75	151,73
Talat and Hovy . (2016)	https://doi.org/10.18653/v1/n16-2013	1490	135,45	92,15
Warstadt, Singh and Bowman (2019).	https://doi.org/10.1162/tacl_a_00290	819	102,38	74,05
Su (2020).	NA	514	73,43	52,89
Cheng & Lapata (2016).	https://doi.org/10.18653/v1/P16-1046	500	45,45	30,92
Clark, Choi, Collins, Garrette, Kwiatkowski, Nikolaev, & Palomäki (2020).	https://doi.org/10.1162/tacl_a_00317	461	65,86	47,44

Nguyen, & Grishman (2015, June).	https://doi.org/10.3115/v1/w15-1506	461	38,42	24,49
Habibi, Weber, Neves, Wiegandt, & Leser (2017).	https://doi.org/10.1093/bioinformatics/btx228	457	45,70	32,63
Li, Zhao, Hu, Li, Liu, & Du (2018).	https://doi.org/10.18653/v1/p18-2023	429	47,67	23,56
Rohrbach, Rohrbach, Tandon & Schiele (2015).	https://doi.org/10.1109/CVPR.2015.7298940	408	34,00	21,67
Atkins, & Rundell (2008).	https://doi.org/10.1093/oso/9780199277704.001.0001	400	100,00	106,59
Mizumoto, & Eguchi, (2023).	https://doi.org/10.1016/j.rmal.2023.100050	395	98,75	105,26
Straka, Hajič, & Straková (2016).	NA	373	33,91	23,07
Khashabi, Khot, Sabharwal, & Roth (2018).	NA	363	40,33	19,94
Zappavigna (2015).	https://doi.org/10.1080/10350330.2014.996948	355	29,58	18,86
Fischer, Pardos, Baker, Williams, Smyth, Yu & Warschauer (2020).	https://doi.org/10.3102/0091732X20903304	345	49,29	35,50
Kulkarni, Al-Rfou, Perozzi & Skiena (2015, May).	https://doi.org/10.1145/2736277.2741627	323	26,92	17,16
Feng, Xiang, Glass, Wang & Zhou (2015, December).	https://doi.org/10.1109/ASRU.2015.7404872	306	27,82	18,92
Hovy & Spruit (2016).	https://doi.org/10.18653/v1/p16-2096	304	27,64	18,80
Zeldes (2017).	https://doi.org/10.1007/s10579-016-9343-x	291	29,10	20,78
Futrell, Mahowald & Gibson (2015).	https://doi.org/10.1073/pnas.1502134112	290	24,17	15,41
Onan (2018).	https://doi.org/10.1177/0165551516677911	284	31,56	15,60
Pechenick, Danforth & Dodds (2015).	https://doi.org/10.1371/journal.pone.0137041	281	23,42	14,93

The Tree Map visualization created in the Bibliometrix software clearly demonstrates the thematic priorities within the research corpus. The most frequent term is “linguistics” (2046), indicating that the studies are generally considered within the broader field of linguistics. Additionally, the high frequency of terms such as “computational linguistics” (1572), “natural language processing systems” (1440), “semantics” (1280), and “corpus linguistics” (1231) shows that corpus linguistics at the current stage is developing in close integration with computational linguistics and natural language processing (NLP).

Among the medium-frequency terms are “natural language processing” (770), “corpus” (617), “classification (of information)” (573), “natural languages” (557), “language processing” (464), “syntactics” (461), and “language model” (436), which indicates that issues such as automated text processing, language modeling, and structural analysis occupy an important place in these studies. Additionally, the occurrence of terms like “deep learning” (351), “machine learning” (349), “artificial intelligence” (308), “sentiment analysis” (311), and “speech recognition” (414) highlights the connection of corpus linguistics with modern artificial intelligence technologies.

Overall, the Tree Map structure indicates that corpus linguistics research has shifted toward a scientific paradigm that is data-driven, automated, algorithmic, and model-oriented rather than traditional linguistic analysis. At the same time, the presence of technical or indexing-related terms such as “na,” “human,” “humans,” and “article” suggests the necessity of preliminary data cleaning (Figure 7).

Figure 7. Tree Map Visualization of Dominant Terms in Corpus Linguistics Publications (Scopus, 2015–2025) Based on Bibliometrix Analysis

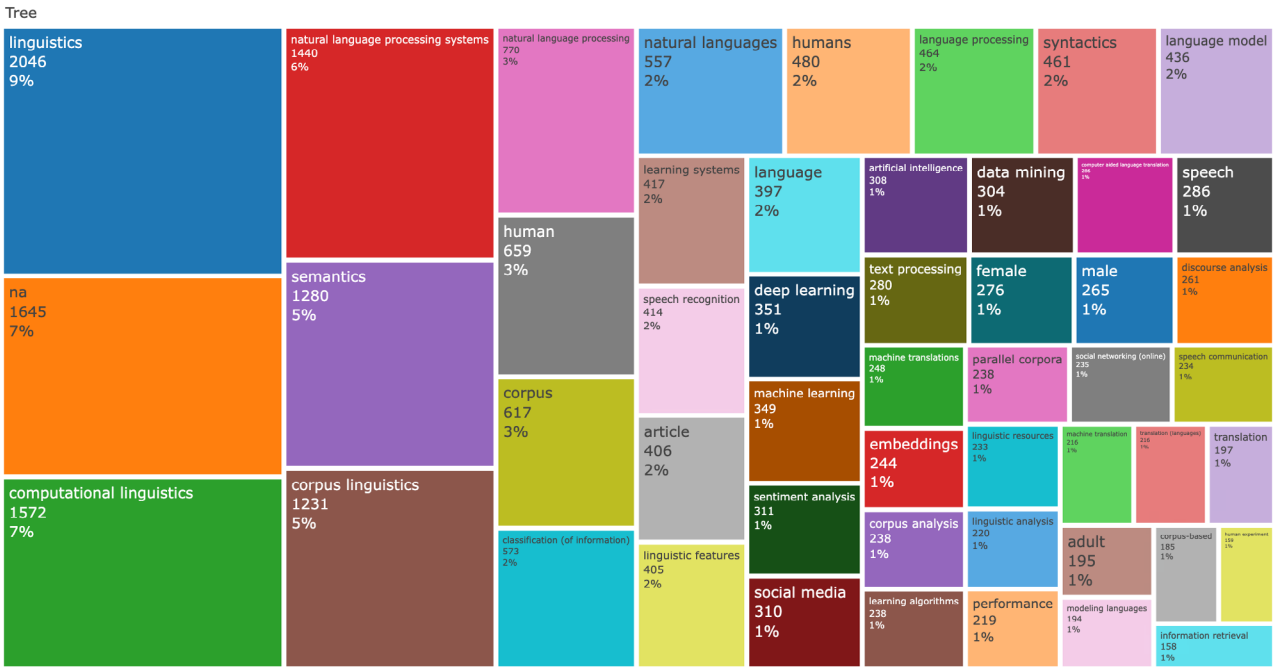
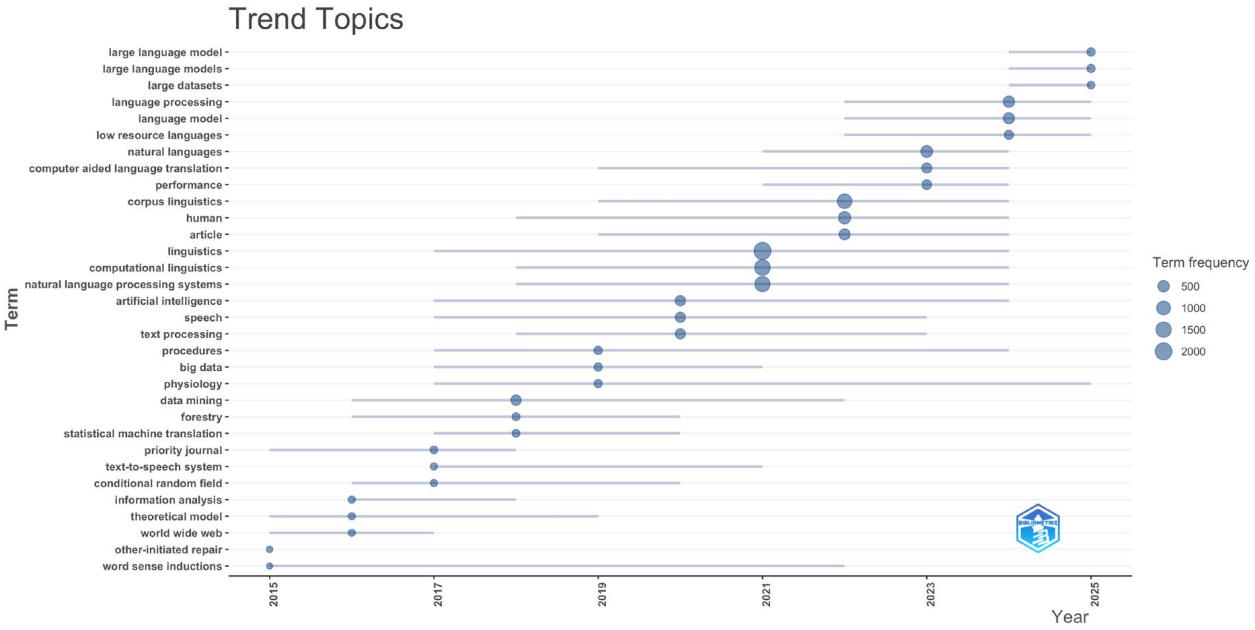


Figure 8. The Trend Topics graph created in Bibliometrix (Biblioshiny) clearly illustrates the thematic evolution of research in corpus linguistics from 2015 to 2025. The graph shows the top three most frequent keywords per year (N. of words per year: 3), highlighting the dominant trends that captured the scientific community’s attention during specific periods. Between 2015 and 2017, relatively diverse and low-frequency terms such as word sense induction, theoretical model, statistical machine translation, text-to-speech system, and information analysis appeared, reflecting research focused on classical, modeling, and algorithmic issues in computational linguistics. From 2018 to 2020, the research focus gradually shifted toward data-driven approaches, with increased frequency of terms like data mining, big data, text processing, speech, and especially artificial intelligence.

From 2021 onwards, a significant thematic shift is observed: terms such as computational linguistics, linguistics, natural language processing systems, and corpus linguistics become highly frequent, indicating that linguistic research is increasingly integrated with NLP systems. The most prominent trends occur in 2023–2025, with terms like large language model(s), language model, language processing, large datasets, and low resource languages dominating. This demonstrates that corpus linguistics research has transitioned to a new scientific paradigm based on artificial intelligence, particularly large language models. Additionally, the appearance of the term low resource languages indicates growing interest in applying language technologies to underrepresented languages.

Overall, the Trend Topics graph depicts a clear evolution from traditional corpus and computational linguistics toward research grounded in big data, artificial intelligence, and large language models.

Figure 8. Evolution of Thematic Trends in Corpus Linguistics Research (2015–2025) Based on Scopus Data and Bibliometrix Analysis



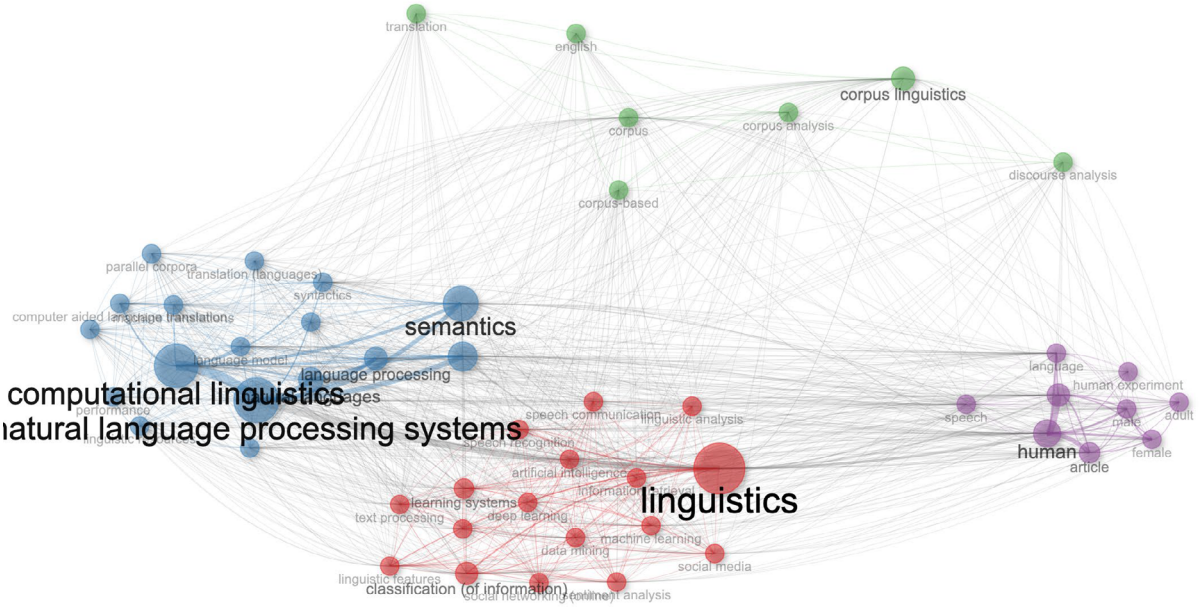
Co-occurrence Network analysis clearly illustrates the interrelationships between keywords and the internal thematic structure of the research field. Two main clusters are evident in the network structure. The first cluster (Cluster 1) is centered around the term linguistics and is closely connected with concepts such as machine learning, deep learning, artificial intelligence, data mining, sentiment analysis, social media, speech recognition, classification (of information), and text processing. This cluster describes the modern direction of corpus linguistics, which is data-driven, algorithmic, and integrated with artificial intelligence technologies.

The second cluster (Cluster 2) is represented by the terms computational linguistics, natural language processing systems, semantics, natural language processing, language processing, language model, and syntactics, reflecting the theoretical and applied foundations of computational linguistics and NLP systems.

According to network centrality metrics, the linguistics node has the highest betweenness centrality value (64.835) and a high PageRank score (0.067), indicating that it serves as the main bridge connecting all thematic directions. Additionally, the terms semantics (22.906), computational linguistics (14.194), and natural language processing systems (11.370) also exhibit high betweenness centrality, demonstrating their important role in the network structure.

The closeness centrality values are roughly similar (≈0.02), indicating that the keywords are densely interconnected, and the research topics are highly integrated. Overall, this network structure highlights that corpus linguistics research has shifted from traditional linguistic analysis toward a comprehensive scientific paradigm based on computational methods, NLP technologies, and artificial intelligence (Figure 9).

Figure 9. Co-occurrence Network of Keywords and Thematic Clusters in Corpus Linguistics Publications (2015–2025) Constructed Using Bibliometrix Based on Scopus Data



The Thematic Map analysis constructed in Bibliometrix identifies the development level (density) and relevance (centrality) of research topics based on keywords, dividing the thematic structure of the corpus linguistics field into four regions. In the upper-right quadrant, the Motor themes area, terms such as computational linguistics, natural language processing systems, and semantics are concentrated. These topics have high centrality and high development levels, indicating that they are the main driving directions of the current research field. Their high frequency (1572; 1440; 1280) and PageRank scores further confirm their leading role in the scientific discourse.

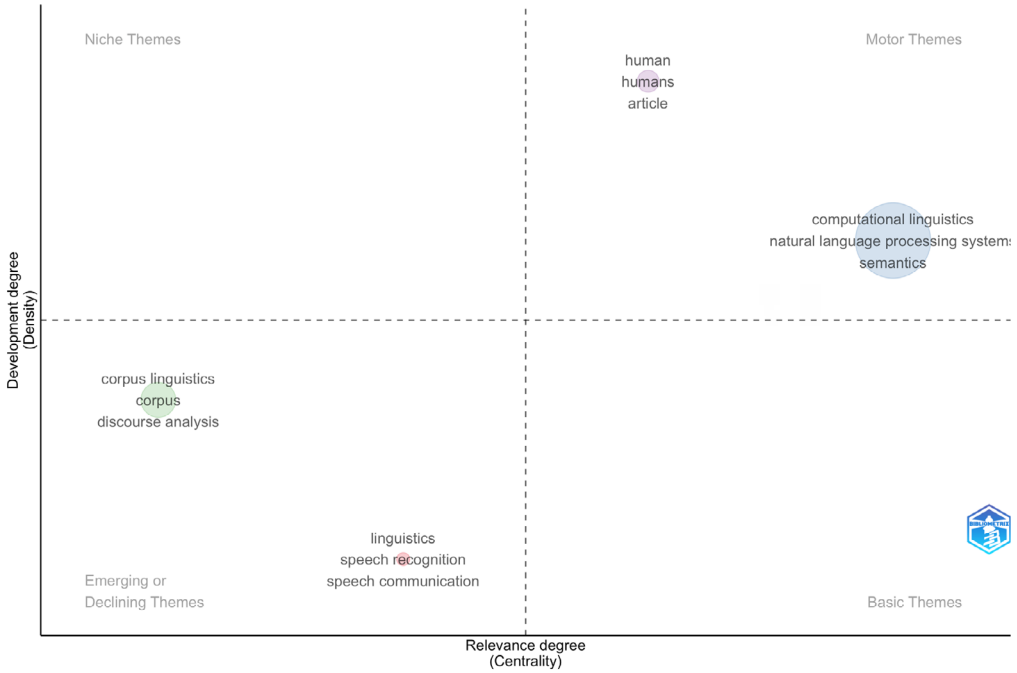
In the lower-left quadrant, the Emerging or Declining themes area, terms like linguistics, speech recognition, and speech communication are presented. These topics have low centrality and weak development levels, representing either emerging directions or areas whose relevance is gradually declining. Nevertheless, terms such as speech recognition (betweenness = 56.476), speech communication (131.616), and code-switching (168.639) have high betweenness centrality, indicating their role as connectors within the network. In the upper-left quadrant, Niche themes, terms such as corpus linguistics, corpus, and discourse analysis are located, showing that while these topics have high internal development, their overall centrality within the network is relatively low. This indicates that these topics are deeply studied within a narrow scope but have limited connections with other directions.

Overall, the Thematic Map results demonstrate that the main driving force of corpus linguistics research is concentrated in computational linguistics, NLP systems, and semantic analysis, while speech technologies and traditional linguistic issues remain relatively peripheral or transitional (Figure 10).

The Countries' Collaboration World Map visualization, created using Bibliometrix, clearly illustrates the broad international scientific collaboration in the field of corpus linguistics. On the map, joint publications between countries are represented by arcs, with the thickness of each connection reflecting its frequency. The results indicate that the geography of scientific

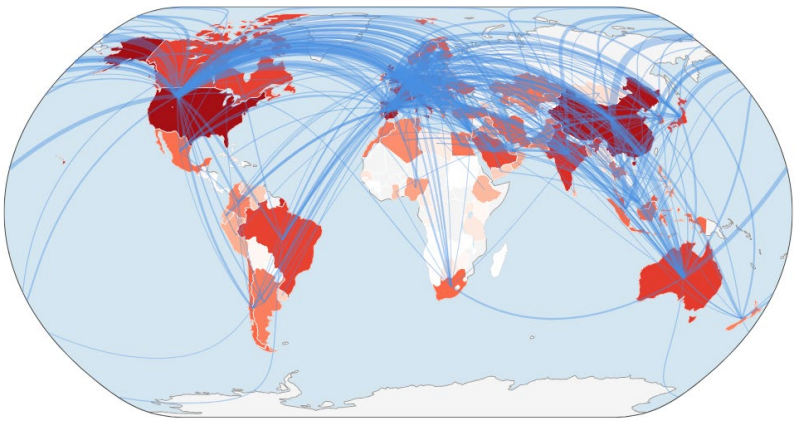
collaboration extensively covers Europe, North America, Asia, and Australia. In particular, the USA, the United Kingdom, China, Australia, Canada, and Western European countries appear as major hubs in the international research network.

Figure 10. Thematic Map of Corpus Linguistics Publications (2015–2025)



Tabular data show the specific frequency of collaborations between countries. For example, Australia has active scientific collaborations with Canada (9), Hong Kong (9), Indonesia (8), Malaysia (5), and Ireland (3). This indicates that Australia acts as a scientific bridge between the Asia-Pacific region and Western countries. Additionally, regional collaborations such as Argentina–Colombia (3), Algeria–Qatar (3), and Albania–North Macedonia (2) are also observed, demonstrating that corpus linguistics research is conducted not only in developed countries but across diverse regions. Overall, this visualization shows that corpus linguistics research has a clear international character, a multi-centric scientific exchange network, and well-established academic connections between countries (Figure 11).

Figure 11. Countries' International Scientific Collaboration Map in Corpus Linguistics Research Based on Bibliometrix Analysis (2015–2025, Scopus)



Conclusion

In conclusion, the bibliometric analysis of 16,402 documents on the keyword "corpus linguistics" from 2015 to 2025, based on the SCOPUS database, demonstrates that the research objectives have been fully achieved. The annual average growth rate of publications at 10.76%, the dissemination of documents across 2,667 sources, registration of 25,788 authors, and an average of 7.709 citations per document quantitatively confirm the field's stable and intensive development in the international scientific landscape. Analysis of annual publication dynamics from 2015 to 2024 showed a continuous increase in output, highlighting the growing relevance of corpus linguistics. Identification of the most productive sources and universities, leading authors, and their publication activity over time made it possible to describe the institutional and authorship structure of the research field, while also revealing the need for author name standardization and data cleaning.

Analysis of international collaboration results indicated that corpus linguistics research has a clear global character, and the MCP and SCP indicators specifically reflect the level of scientific interaction between countries. The Most Global Cited Documents analysis showed that the scientific evolution of the field has shifted from traditional corpus studies to neural language models, artificial intelligence, and NLP technologies. Tree Map, Trend Topics, Co-occurrence Network, and Thematic Map visualizations demonstrated that research topics are developing in close integration with computational linguistics, semantics, NLP systems, artificial intelligence, deep learning, and language models, while traditional areas such as corpus linguistics and discourse analysis have taken on a relatively niche character. The Countries' Collaboration World Map results confirmed the multi-centric structure of scientific cooperation and the establishment of stable international academic networks. Thus, the obtained results made it possible to comprehensively describe, based on empirical data, the structure of scientific productivity in the field of corpus linguistics, its thematic evolution, institutional and geographic distribution, as well as the technological vector of future research directions.

Conflict of interests, acknowledgements and funding information

Published within the framework of the Project AP25794379 "The study of the formation of lexical skills using corpus linguistics in teaching the Kazakh language and the development of its methodology".

The article contains no conflict of interests.

Contributions of the authors. A.N. Oraz compiled the materials and wrote the main content of the article. K.T. Malikov defined the core idea and conceptual framework of the study, conducted the analysis of the research materials, and critically reviewed and revised the manuscript. N.S. Khalikova was responsible for data collection and formal analysis.

References

Alamri, B. (2022). The Role of Corpus Linguistics in Grammar Instruction: A Review of Literature. International Journal of Linguistics, 14(6), pp. 158–167. https://doi.org/10.5296/ijl.v14i6.20500

Atkins, B. S., & Rundell, M. (2008). The Oxford guide to practical lexicography. Oxford University Press. https://doi.org/10.1093/oso/9780199277704.001.0001

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative science studies, 1(1), 377-386. https://doi.org/10.1162/qss_a_00019

Cheng, J., & Lapata, M. (2016). Neural summarization by extracting sentences and words. In K. Erk & N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 484–494). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1046

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomäki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics, 8, 454–470. https://doi.org/10.1162/tacl_a_00317

Crosthwaite, P., Ningrum, S., & Schweinberger, M. (2023). Research trends in corpus linguistics: A bibliometric analysis of two decades of Scopus-indexed corpus linguistics research in arts and humanities. International Journal of Corpus Linguistics, 28(3), 344-377. https://doi.org/10.1075/ijcl.21072.cro

Gapporov Baxriddin Baxtiyor ugli. (2025). The Role Of Corpus Linguistics In The Study Of Collocations. International Journal Of Literature And Languages, 5(12), 261–263. https://doi.org/10.37547/ijll/Volume05Issue12-70

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 33(14), i37–i48. https://doi.org/10.1093/bioinformatics/btx228

Haeruddin, H., Fitriani, F., & Zuhriah, Z. (2025). Modern Linguistics: Bibliometric Analysis based on Scopus and Google Scholar (2019-2023) using VosViewer. Jurnal Sastra Indonesia, 14(2), 206-220 https://orcid.org/0000-0001-6790-0706

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on chinese morphological and semantic relations. arXiv preprint arXiv:1805.06504. https://doi.org/10.18653/v1/p18-2023

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Nguyen, T. H., & Grishman, R. (2015, June). Relation extraction: Perspective from convolutional neural networks. In Proceedings of the 1st workshop on vector space modeling for natural language processing (pp. 39-48). https://doi.org/10.3115/v1/w15-1506

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. A. Walker, H. Ji, & A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018 (Vol. 1, pp. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-1202

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019, November). Language models as knowledge bases?. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)(pp. 2463-2473). https://doi.org/10.18653/v1/D19-1250

Plungian, V. A. (2024). Corpus linguistics nowadays. Herald of the Russian Academy of Sciences, 94(9), 787–794. https://doi.org/10.31857/S0869587324090018

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3202-3212). https://doi.org/10.1109/CVPR.2015.7298940

Roslim, N., Hakimi Tew Abdullah, M., Faathinah Mohammad Roshdan, N., Jin Ng, Y., & Ali Resvani Kalajahi, S. (2023). Learner Corpus Research: A Bibliometric Analysis. International Research in Education, 11(2), pp. 76–93. https://doi.org/10.5296/ire.v11i2.21396

Su, W. (2020). [мақала атауы]. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020). International Conference on Learning Representations.

Talat, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93). https://doi.org/10.18653/v1/n16-2013

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 7, 625-641. https://doi.org/10.1162/tacl_a_00290

А.Н. Ораз[1]*, Қ.Т. Маликов[2], Н.С. Халикова[3]
[1,3]Ө. Жәнібеков атындағы Оңтүстік Қазақстан педагогикалық университеті, Шымкент, Қазақстан
[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

CORPUS LINGUISTICS КІЛТ СӨЗІ БОЙЫНША ҒЫЛЫМИ ЖАРИЯЛАНЫМДАРДЫҢ БИБЛИОМЕТРИЯЛЫҚ ТАЛДАУЫ: SCOPUS ДЕРЕКТЕРІ НЕГІЗІНДЕ (2015–2025)

Аңдатпа. Бұл зерттеу corpus linguistics саласындағы ғылыми жарияланымдардың даму үрдістерін, құрылымдық ерекшеліктерін және негізгі зерттеу бағыттарын анықтауға бағытталған. Зерттеудің мақсаты – SCOPUS деректер қоры негізінде 2015–2025 жылдар аралығында жарияланған corpus linguistics тақырыбындағы еңбектерге библиометриялық талдау жүргізу арқылы аталмыш саланың ғылыми өнімділігін, жарияланымдар динамикасын, жетекші авторлар мен елдерді, халықаралық ғылыми ынтымақтастық үлгілерін, сондай-ақ басым тақырыптық трендтерді айқындау. Зерттеуде библиометриялық талдау негізгі әдіс ретінде қолданылды. Деректер SCOPUS деректер қорынан “corpus linguistics” кілт сөзі арқылы 23 073 құжат табылып, уақыттық сүзгі нәтижесінде (2015-2025) 16 475 құжат жүктеліп алынып, Bibliometrix бағдарламасының нәтижесінде 16 402 ғылыми құжат талдауға енгізілді. Библиографиялық деректер RStudio бағдарламалау ортасында Bibliometrix пакеті мен Biblioshiny интерфейсі арқылы өңделді. Талдау барысында жарияланымдардың жылдық динамикасы, авторлық және елдер арасындағы ынтымақтастық, ең жиі қолданылатын кілт сөздер мен тақырыптық желілер қарастырылды. Зерттеу нәтижелері corpus linguistics саласында соңғы жылдары ғылыми жарияланымдар санының жоғары екенін, халықаралық ынтымақтастықтың біртіндеп кеңейіп келе жатқанын және зерттеу фокусының есептеу лингвистикасы мен тілдік модельдерге қарай бағытталғанын көрсетеді. Бұл зерттеу corpus linguistics саласындағы қазіргі ғылыми үрдістерді жүйелі түрде түсінуге мүмкіндік беріп, болашақ зерттеулер үшін әдіснамалық және мазмұндық негіз қалыптастырады.

Түйін сөздер: corpus linguistics; библиометриялық талдау; ғылыми жарияланымдар; SCOPUS деректер қоры; зерттеу үрдістері; тақырыптық эволюция; халықаралық ғылыми ынтымақтастық.

А.Н. Ораз[1]*, Қ.Т. Маликов[2], Н.С. Халикова[3]
[1, 3] Южно-Казахстанский педагогический университет имени О. Жанибекова, Шымкент, Казахстан
[2] Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан

БИБЛИОМЕТРИЧЕСКИЙ АНАЛИЗ НАУЧНЫХ ПУБЛИКАЦИЙ ПО КЛЮЧЕВОМУ СЛОВУ CORPUS LINGUISTICS НА ОСНОВЕ ДАННЫХ SCOPUS (2015–2025)

Аннотация. Данное исследование направлено на выявление тенденций развития, структурных особенностей и основных направлений научных исследований в области corpus linguistics. Цель исследования заключается в проведении библиометрического анализа публикаций по теме corpus linguistics, опубликованных в период 2015–2025 гг. на основе данных базы SCOPUS, с целью определения научной продуктивности данной области, динамики публикаций, ведущих авторов и стран, моделей международного научного сотрудничества, а также доминирующих тематических трендов. В исследовании библиометрический анализ использовался в качестве основного метода. По ключевому слову “corpus linguistics” в базе SCOPUS было выявлено 23 073 документа, после применения временного фильтра (2015–2025) было выгружено 16 475 документов, из которых в результате очистки в программе Bibliometrix к анализу было допущено 16 402 научных документа. Библиографические данные обрабатывались в среде программирования RStudio с использованием пакета Bibliometrix и интерфейса Biblioshiny.

В ходе анализа были рассмотрены годовая динамика публикаций, авторское и межстрановое сотрудничество, наиболее часто используемые ключевые слова и тематические сети. Результаты исследования показывают, что в последние годы в области corpus linguistics наблюдается высокий рост количества научных публикаций, международное сотрудничество постепенно расширяется, а фокус исследований смещается в сторону вычислительной лингвистики и языковых моделей. Данное исследование позволяет системно понять современные научные тенденции в области corpus linguistics и формирует методологическую и содержательную основу для будущих исследований.

Ключевые слова: corpus linguistics; библиометрический анализ; научные публикации; база данных SCOPUS; исследовательские тенденции; тематическая эволюция; международное научное сотрудничество.

Information about the authors:

Oraz Aidana Nurmakhankyzy – corresponding author, PhD, Senior Lecturer, O. Zhanibekov South Kazakhstan Pedagogical University, Shymkent, Kazakhstan. E-mail: Aydana.oraz.1997@mail.ru, ORCID: https://orcid.org/0000-0002-4283-9483

Malikov Kuanishbek Turarbekuly – Candidate of Philology, Associate Professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. E-mail: k.malikov@mail.ru, ORCID: https://orcid.org/0000-0002-1563-1927

Khalikova Nurila Satybaldykyzy – Candidate of Philology, Senior Lecturer, O. Zhanibekov South Kazakhstan Pedagogical University, Shymkent, Kazakhstan. E-mail: khalikova_nurila1978@mail.ru, ORCID: https://orcid.org/0000-0002-1360-0131

Ораз Айдана Нурмаханқызы – хат-хабар үшін автор, PhD, аға оқытушы, Ө. Жәнібеков атындағы Оңтүстік Қазақстан педаогогикалық университеті, Шымкент, Қазақстан. E-mail: Aydana.oraz.1997@mail.ru, ORCID: https://orcid.org/0000-0002-4283-9483

Маликов Қуанышбек Тұрарбекұлы – филология ғылымдарының кандидаты, доцент, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан. E-mail: k.malikov@mail.ru, ORCID: https://orcid.org/0000-0002-1563-1927

Халикова Нурила Сатыбалдықызы – филология ғылымдарының кандидаты, аға оқытушы, Ө. Жәнібеков атындағы Оңтүстік Қазақстан педаогогикалық университеті, Шымкент, Қазақстан. E-mail: khalikova_nurila1978@mail.ru, ORCID: https://orcid.org/0000-0002-1360-0131

Ораз Айдана Нурмаханкызы – автор для корреспонденции, PhD, старший преподаватель, Южно-Казахстанский педагогический университет имени О. Жанибекова, Шымкент, Казахстан. E-mail: Aydana.oraz.1997@mail.ru, ORCID: https://orcid.org/0000-0002-4283-9483 .

Маликов Куанышбек Турарбекович – кандидат филологических наук, доцент, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан. E-mail: k.malikov@mail.ru, ORCID: https://orcid.org/0000-0002-1563-1927

Халикова Нурила Сатыбалдиевна – кандидат филологических наук, старший преподаватель, Южно-Казахстанский педагогический университет имени О. Жанибекова, Шымкент, Казахстан. E-mail: khalikova_nurila1978@mail.ru, ORCID: https://orcid.org/0000-0002-1360-0131