ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/kaifang-deng/669

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 9 & Iss: 1

Review

International Trends and Frontiers of Corpus Linguistics Based on BER Topic: Theories, Methods and Prospects

Kaifang Deng*

College of Foreign Studies, Hunan Normal University, Changsha, Hunan, 410081, China

Citation: Deng K. International Trends and Frontiers of Corpus Linguistics Based on BER Topic: Theories, Methods and Prospects. *J Artif Intell Mach Learn & Data Sci 2026* 9(1), 3335-3342. DOI: doi.org/10.51219/JAIMLD/kaifang-deng/669

Received: 02 March, 2026; **Accepted:** 16 March, 2026; **Published:** 18 February, 2026

***Corresponding author:** Kaifang Deng, College of Foreign Studies, Hunan Normal University, Changsha, Hunan, 410081, China, E-mail: 15675195286@163.com

ABSTRACT

In the interwoven world of modern linguistics and computer science, corpus linguistics plays a significant role in applied linguistics. As a research method based on large-scale real language texts, corpus linguistics not only provides a new perspective for language research, but also injects a strong impetus for the development of natural language processing technology, which is worthy of in-depth study and application. Based on the BERTopic automatic topic extraction model, this study collected and screened 220 articles published in 2 authoritative international journals of corpus linguistics from January 2019 to March 2024 and analysed the current international trends in corpus linguistics in detail from theoretical and methodological perspectives. The analysis results show that the current research of corpus linguistics is more about the introduction and application of cutting-edge methods and its research objects are mostly language itself (including syntactic, semantic and pragmatic features). The cutting-edge theories used for interpretation mainly belong to cognitive linguistics, sociolinguistics and cross-cultural communication, etc. At the same time, this study analyzed the current research status, hotspots and clustering situation of advanced theory and methods in corpus linguistics, to provide reference and inspiration for corpus linguistics research.

Keywords: Corpus linguistics, International frontier, Theory, BERTopic model

1. Introduction

Against the background of globalization, corpus linguistics, as an important branch of linguistic research, has increasingly attracted the attention of the international academic community due to its unique research methodology, rich data resources and wide applicability in multiple disciplines and fields. The development of corpus linguistics can be traced back to the middle of the 20th century and it has gone through several stages, including initial germination, silence, revival, development and modern development. Its evolution is closely linked to the development of linguistic theories, technological advances and changes in educational needs[1] With the advent of the big data era, corpus linguistics will be more widely used in the fields of natural language processing, artificial intelligence, data mining, etc., which can not only enhance the performance and efficiency of these technologies, but also provide strong support for language education, cultural exchange, cross-linguistic communication and other social fields. These uniqueness's also make corpus linguistics occupy an important position in linguistic research and make significant contributions to the advancement of linguistic theories and the expansion of application fields.

2. Literature Review

The current research literature on corpus linguistics mainly involves the following aspects:

- From the perspective of language ontology, corpus linguistics uses large-scale real-language text data to reveal the use frequency of vocabulary, syntactic structure collocation and semantic changes and extracts the syntactic, semantic and pragmatic features of vocabulary and phrases by analyzing their use in contexts and then constructs semantic networks or models[2-7].
- From the perspective of language teaching and acquisition, researchers analyse a large number of language facts to find the rules of language use by building a large-scale learner corpus. Teachers can utilize the data in the corpus to understand the actual performance and problems of learners in the process of language learning, to make targeted teaching suggestions and guidance[8-11].
- From the perspective of corpus linguistics and cognitive linguistics, the corpus can provide rich real language data for cognitive linguistics, which can help cognitive linguistics better reveal the cognitive process and mechanism of language, as well as the thinking and psychological changes reflected behind the language. And the theories and methods of cognitive linguistics can also provide corpus linguistics with new research ideas and methodological guidance[12-15].
- From the perspective of application of corpus in the field of natural language processing and artificial intelligence, corpus linguistics provides a large amount of real language text data, linguistic knowledge and strong natural language processing and artificial intelli research results can provide theoretical and me guidance for NLP and promote the continuous of NLP technology[6,16-19].
- From the perspective of corpus linguistics and sociolinguistics, corpus data are used to study social language phenomena, such as language variation, the relationship between language use and social factors, etc. Corpus data help to reveal the close connection between language and society, culture, psychology, etc[20-22].
- From the perspective of corpus linguistics in the field of multilingualism and cross-culturalism, by comparing the text corpus covering multiple languages, by comparing text corpora covering multiple languages, researchers can tap into the patterns and characteristics between different languages, which helps to promote communication and interaction between languages, linguistic diversity and cultural exchange (Van, 2019)[23,24].
- Some scholars also made an overview summary of the current research status and development trend of corpus linguistics[25-27].

Reviewing the previous literature, most of the corpus linguistics research is related to language itself (vocabulary, meanings and grammar) or language teaching and acquisition, providing real language examples through corpora to improve language teaching methods and materials. Few studies comprehensively summarize the current cutting-edge theories and methods of international corpus linguistics research, not to mention the use of computerized big data computation and quantitative methods to explore international journals' attention to the theories and research methods. Therefore, based on the BER Topic topic model, this paper collates all the papers published in two specialized international corpus linguistics journals in the past five years, processes and fine-tunes the relevant data, so that the model can automatically identify and extract the current international research topics of the cutting-edge theories and methods of applied linguistics. We can further analyze the research hotspots of advanced theories, methods and the future development trend of corpus linguistics research based on the data results. The main research questions of this paper include:

- What is the international thematic focus of corpus linguistics?
- What are the cutting-edge theories and methods of corpus linguistics?
- What are the research hotspots and future development trends of corpus linguistics?

3. Research Design

This part mainly contains two parts: the general framework introduction and data processing. Regarding the fields involved in the journal titles and the impact factor coefficients of the journals, the authors selected 2 international corpus linguistics journals with high impact, explored the corpus linguistics research topics, cutting-edge theories and methods covered in the papers published in these journals in the past 5 years (from 2019 to 2024) and analyzed their research hotspots and future development trends, the specific steps are shown in **(Figure 1)**.

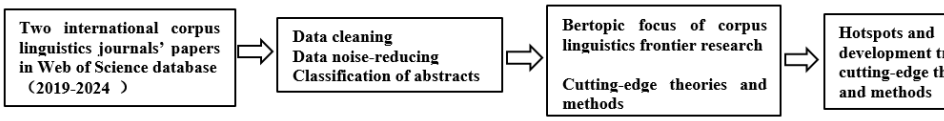

Figure 1: General Research Framework.

- **Data collection:** The corpus of the papers in this study is all from SSCI core journals of the Web of Science database. As of March 6, 2024, we have collected a total of 249 papers from 2 corpus linguistics journals (in the last five years) in the Web of Science database. Relevant information, such as author names, titles, abstracts, keywords and publication dates, is retained.
- **Data processing:** The 249 papers obtained were cleaned, that is, some papers that were repetitive and non-English were deleted. Meanwhile, the remaining data were subjected to English lexical processing and the text was noise-reduced by using the default English stop word list in Python. Finally, combined with manual screening, the "thesis and review papers" in the data were retained and the format of the text was unified, resulting in 220 valid documents. The specific journal data are shown in **(Table 1)**.

Table 1: Information table of 2 international corpus linguistics journals.

Journals	Impact factor	Cite score	IF*CS	Papers
Corpus Linguistics and Linguistic Theory	1.9	2.28	4.33	114
International Journal of Corpus Linguistics	1.8	1.15	2.07	106

- **Automatic identification of theoretical and methodological topics:** Automatic topic recognition is to model all abstracts of papers published in the recent five years in international corpus linguistics journals collected by computer and manually screened. The research topics of international corpus linguistics research are identified by a computer algorithm and then clustered. Finally, the advanced theories and methods are analyzed to make the information more diverse.
- **Comprehensive discussion:** By automatically identifying the topics of all abstracts published in two international corpus linguistics journals in the past five years and classifying their subject words, we get the cutting-edge theories and methods of corpus linguistics. At the same time, the research hotspot and future development trend of cutting-edge theories and methods are further elaborated and analysed. The final analysis results can help people better understand the international dynamic trends of corpus linguistics research.

4. Results and Analysis

BERTopic is a topic extraction model that can efficiently and quickly model a large amount of short text data[28] and thus is also well suited for topic extraction and clustering of short texts such as dissertation abstracts, keyword sets, etc.

4.1. Topic focus of frontier research

Based on the BERTopic model, the themes of all the paper abstracts published in the last five years in 2 international corpus linguistics journals are automatically extracted and the specific results are shown in **(Figures 2-4)**.

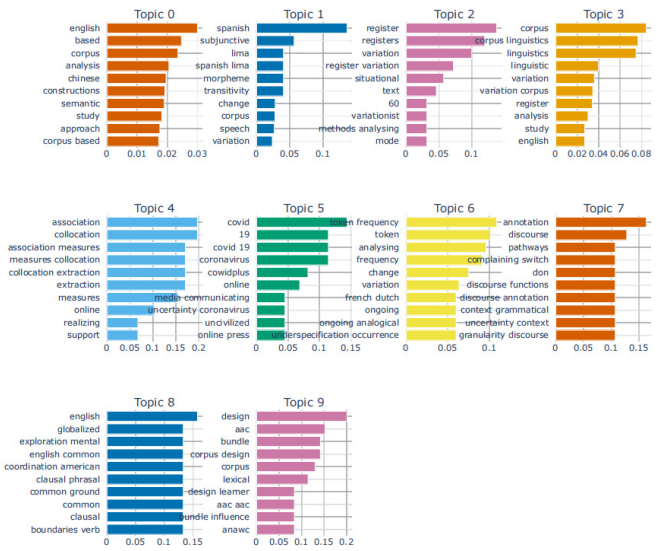

Figure 2: Scores of subject terms in cutting-edge research of corpus linguistics.

A total of 10 topics were obtained after automatic extraction of the themes from the abstract text **(Figure 2)**. The fields involved "language ontology", "language acquisition and teaching", "sociolinguistics and intercultural communication", "corpus methods and computer technology". The overall topic distribution is shown in **(Table 2)**:

(Table 2) shows the automatic classification of the topics of articles published in the last 5 years in 2 international journals of applied linguistics, which are grouped into 4 major fields according to the subject terms of each topic. The order of the number of topics in different fields is: language ontology (including language acquisition and teaching) > corpus methods and computer technology > sociolinguistics and intercultural communication.

Table 2: Information table of automatic classification of journal paper topics.

Topics	Frequency	Specific subject terms
language ontology	19%	topic0、topic3、topic2、topic6
Language acquisition and teaching	23%	topic1
sociolinguistics and intercultural communication	16%	topic4、topic8
corpus methods and computer technology	31%	topic5、topic7、topic9

It is important to note that these topic classifications are not absolute and a topic may involve multiple domains with different focuses and preferences. For example, some analyzed language rules in language ontology research would be applied to language teaching and acquisition. In addition to the establishment of social media, multilingual or parallel corpora to analyze the differences in language between different social backgrounds or cultures and to explore the inner language rules, social and cultural factors behind them, the researches on sociolinguistics and cross-cultural communication will also refer to the corpus methodology and computer technology used. The literature focusing on corpus methods and computer technology will also be supported by a large number of real language data or social discourse corpora to further improve the accuracy and efficiency of corpus methods and computer models. Therefore, the division of corpus linguistics topic classifications is not absolute and they are closely related, including a variety of subject areas.

The clustering trend of different topics in corpus linguistics is more intuitively reflected in **(Figure 3)**.

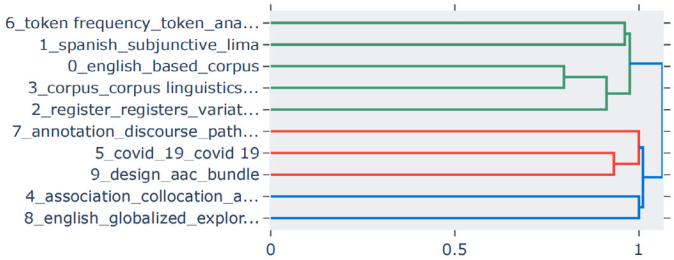

Figure 3: Topic (subject terms) clustering of corpus linguistics cutting-edge research

(Figure 3) visualizes the connections among topics by hierarchical clustering. Several clusters formed after model calculation, with different colors representing text proximity of certain topics, showing the connections among topics at different levels. For example, the green clusters include the research on language ontology (topic0, topic3, topic2, topic6) and language teaching and acquisition (topic1) in corpus linguistics. The red clusters focus on the combination of corpus methods and computer technology (topic5, topic9, topic7) with some sociolinguistics. The blue clusters (topic4, topic8) are mainly about sociolinguistics and intercultural communication **(Figure 4)**.

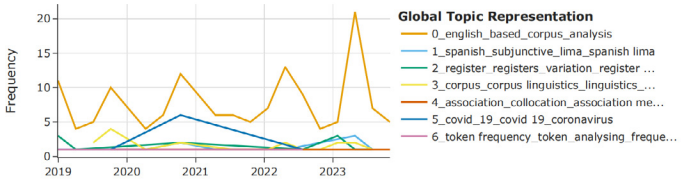

Figure 4: Dynamic evolution of topics in corpus linguistics cutting-edge research.

The dynamic evolution of the topics of corpus linguistics **(Figure 4)** shows that:

The amount of corpus linguistics research is on the rise as a whole and has attracted more and more attention in recent years.

Language ontology is the main body of corpus linguistics and the number of documents accounts for the majority of the whole; language acquisition and teaching are also the focus of the research.

The focus of corpus linguistics research will also be restricted by external environmental factors. For example, due to the global spread of the COVID-19 from 2019 to 2021, databases such as media discourse and journal paper discourse on the COVID-19 have been established one after another to analyze discourse expressions (semantic and pragmatic features) in special periods, revealing the importance of scientific argumentation in times of intense social anxiety[29].

With the globalization and economic integration, there has been an increasing interest in multilingual and cross-linguistic corpus research, so the sociolinguistic and cross-linguistic cultural research using the corpus method has gradually increased.

4.2. Cutting-edge theories and research methods

By automatically extracting the topics of abstracts of the papers in the current frontiers of corpus linguistics, the data results are further mined to explore the cutting-edge theories and research methods. The specific results are shown in **(Table 3)**.

Table 3: International cutting-edge theories and methods in corpus linguistics.

	Advanced theories	Advanced research methods
language ontology	Prototype theory, Metaphor theory, Corpus-based Distributional Semantic Space, Boundary Permeability Hypothesis, Rhetorical Structure theory (RST), Conceptual Metaphor Theory, Assessing theory, Construction grammar (CxG), analysis of Discourses (CBADs), etc.	Aspectual-Semantic classification model, Context Specificity measure, Word2vec approach，Cognitive Representation model, Constant Rate Hypothesis (CRH), Multidimensional Scaling (MDS), Semantic Map, Behavioral Profile (BP) approach, Hierarchical Agglomerative Cluster analysis (HAC), Multiple Correspondence analysis, Conditional Inference Trees, Conditional Random Forests, Hierarchical Configural Frequency Analysis (HCFA), Convolutional Neural Networks (CNNs), Dependency Profiles (DPs), Data-driven method, Collexeme analysis, DISCOver, Distributional Semantic Model, Collocation Strength, Multifactorial approach, Multi-dimensional (MD) analysis, Corpus-assisted Multimodal Discourse analyses, Four Dispersion measure, Automated Vocabulary analysis tool, Bibliometric analysis, etc.
language acquisition and teaching	Second Language Acquisition theory, Usage-based Language Acquisition theory, etc.	Usage-based approach, Cross-corpus approach, Multifactorial model, Mixed-effects Logistic, Regression, Unidirectional Association Score Delta P, Mixed-effects negative binomial regression analyses, Hierarchical Cluster Analyses, Verb-particle Constructions (VPC), Logistic Regression, Naive Discriminative Learning, etc.
sociolinguistics and cross-cultural communication	Speech Act theory, Rhetorical Structure Theory (RST), Segmented Discourse Representation Theory, Polarity Shift, Language Contact Theory, Standard theories of Sociolinguistic Variation, Theory of Moral Politics (TMR), etc.	Multi-Feature statistical model, Theory of Mind (ToM), Social Distance measure, Information-theoretic Metric, Kolmogorov Complexity, model of Loan-use, Keymorph analysis (KMA), Iinductive (Key POS-tags') method, Discourse Dynamics approach, Annotation approach, etc.
corpus methods and computer technology	Asymmetric Priming Hypothesis, Pooling Strategy, etc.	Multidimensional Scaling (MDS), Semantic Map, Null-hypothesis Significance Testing, Recurrent Neural Network Computational Learning model, Semantic Vector Spaces, Unidirectional Association Score Delta P, Wilcoxon Rank Sum test, Text-dispersion-based measure, Nonprogressive Alternation, Generalized Linear Mixed Methods-tree analysis, Unified Dimension (UniDim) approach, Convolutional Neural Networks (CNNs), DISCOver, Computational N-gram Language mode, Structural Equation modeling, Measured Variable Path model, Keyness, Electronic Supplement analysis (ESA), Annotation approach, Bayesian Language Variation Analysis (BLaVA), Log Likelihood Ratio, Cubic Mutual Information (MI3), Concordance analysis, etc.

(Table 3) summarizes the cutting-edge theories and research methods in different fields of corpus linguistics in the past five years. Combined with the data in **(Figures 2-4)**, it can be concluded that in current corpus linguistics, language ontology (the study of syntax, semantics and pragmatic features) is still the focus of research and most of the the theories used are related to cognitive linguistics, such as Conceptual Metaphor theory, Prototype theory, Construction grammar (CxG), etc. More research methods are also involved, among which the Behavioural Profile (BP) approach, Hierarchical Agglomerative Cluster analysis (HAC) and Multiple Correspondence analysis have been utilised a lot in corpus linguistics. In addition, due to the vigorous development of information technology such as computer and multimedia in recent years, more and more researches on language ontology, language teaching and acquisition and sociolinguistics have applied computer artificial intelligence methods, such as Semantic Map, Null-hypothesis Significance Testing, Recurrent Neural Network Computational Learning model, Semantic Vector Spaces, Unidirectional Association Score Delta P, etc. It is worth noting that the corpus linguistics research mostly involves technology and methodology and there are not many theoretical interpretations used to explain the rules and phenomena of language, which is a research direction worth expanding and deepening in the future of corpus linguistics.

5. Discussion

Based on **(Table 3)**, it can be seen that the current corpus

linguistics research mainly focuses on the application and discussion of methods and the interpretation of linguistic rules is mostly related to cognitive linguistics theory. Corpus linguistics provides a large amount of real language data, which provides an empirical basis for the cognitive linguistics' construction and validation. The theoretical framework and methodology of cognitive linguistics also provide a new perspective and tool for corpus linguistics research. The two complement each other and their combination helps us to understand the nature and rules of language more comprehensively and also provides important theoretical support and empirical basis for other fields such as artificial intelligence or NLP.

Therefore, this section focuses on the interpretive cognitive theories and advanced methods used in "corpus techniques and cognitive linguistics", making a comprehensive analysis towards them. Exploring the hotspots and trends of cutting-edge theories and methods will help to promote innovative development, enhance the depth of the discipline and cultivate innovative thinking and thinking ability.

5.1. Corpus linguistics and cognitive linguistics

Corpus linguistics pays more attention to the data collection and analysis of data and reveals the rules of language through statistics and computation[30]. Cognitive linguistics emphasizes the cognitive properties and functions of language, focuses on explaining how language works from the perspective of the human mind[31]. This perspective helps dig deeper into the cognitive mechanism behind language data and reveals the mental processes and cognitive strategies of language use. The combination of corpus linguistics and cognitive linguistics makes linguistic research both empirical and able to delve into the cognitive aspects of language, providing more comprehensive and in-depth insights.

5.2. Research hotspots of corpus methodology and cognitive linguistics

Through keywords, we can quickly understand the main research content of an article, which is the essence of the article, the concentration of the content and the direction of research. Based on the keyword co-occurrence and clustering function in CiteSpace 6.2. R6, we get a total of 238 documents in the field of corpus methodology and cognitive linguistics in the last five years, exploring their hotspots and frontier trends. The top 10 keywords are selected and sorted according to the two indices of keyword frequency and centrality, as shown in **(Table 4)**.

(Table 4) shows that the top ten keywords in both frequency and centrality are almost consistent, which further indicates the consistency of validity and reliability of data results. Keywords with high centrality (Centrality >= 0.1) can be easily regarded as the inflection point of research trend, which represents the research hotspot in this field to a certain extent[32]. Therefore, according to Table 5, it can be seen that the current focus of corpus methodology and cognitive linguistics is mainly on "Language ontology" ("English" "constructions" "cognitive metaphor"), "Language teaching and acquisition" ("acquisition") and "Language comparison" ("comparative correlatives"). The centrality of "English" "constructions" and "cognitive metaphor" is 0.73, 0.62 and 0.53, respectively, which effectively supports the semantic network **(Figure 5)**.

Table 4: Top 10 keywords ranked by frequency and centrality.

Frequency	Year	Keywords	Centrality	Year	Keywords
72	2019	cognitive linguistics	0.73	2019	English
23	2019	corpus linguistics	0.62	2019	constructions
16	2019	language	0.53	2020	cognitive metaphor
15	2019	English	0.48	2019	corpus
12	2020	metaphor	0.46	2019	corpus linguistics
12	2019	corpus	0.38	2019	cognitive linguistics
8	2019	conceptual metaphor	0.32	2019	acquisition
7	2020	construction grammar	0.27	2019	language
7	2019	discourse	0.24	2020	metaphor
7	2019	acquisition	0.18	2019	comparative correlatives

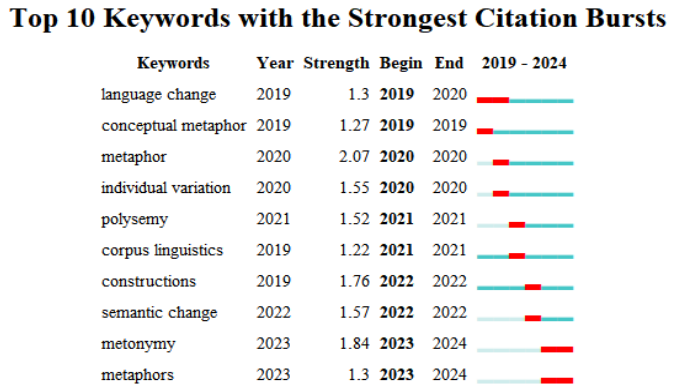

Figure 5: Keywords burstiness mapping of corpus methodology and cognitive linguistics.

The Keyword burstiness mapping can demonstrate the sudden decrease or increase of the cited frequency of documents, which in turn reflects the major shift of research hotspots. To track the inflection, point of hotspots between corpus methodology and cognitive linguistics, we used the Burstness function of CiteSpace to detect the burstiness of keywords and 10 bursty keywords were obtained **(Figure 5)** and each research hotspot showed an outburst trend in a short time. Over time, the research hotspots have shifted from language ontological research (syntax, semantics and pragmatic features) or application of theories ("language change" "conceptual metaphor" "constructions") in 2019 to the semantic metaphors' dynamic evolution and the further expansion of metonymy theory in the past two years ("semantic change" "metonymy" "metaphors"). It can be seen that the theoretical interpretation of cognitive linguistics is more and more closely integrated with corpus methodology and technology. For example, Krieken & Sanders (2019) constructed a semantic network for the German preposition hinter ('behind') based on the theoretical framework of principled polysemy and a self-built corpus; they also presented six senses of the preposition hinter, hinting at the polysemous nature of prepositions more generally. What's more, conceptual metaphor theory was applied to account for metaphorical extensions of hinter to more abstract domains of embodied experience.

The development trend of corpus methodology and cognitive linguistics research.

Keyword clustering plays a vital role in clarifying the research direction, referring to the potential research problems existing in the field, which can summarize the similarity between nodes of keywords, cluster nodes with obvious co-occurrence relationship into a class according to data operation and represent the trend of a cluster changing over time[33]. It is also conducive to the researcher's observation and analysis of the development and change of the research field. Therefore, we cluster the keywords of corpus methodology and cognitive linguistics **(Figure 6)**.

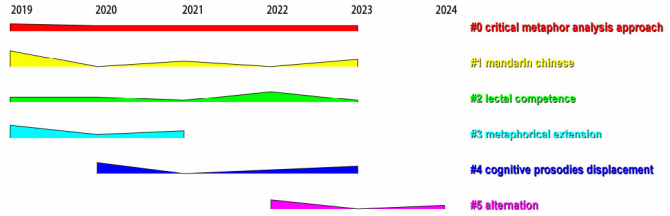

Figure 6: Clustered peaks of corpus methodology and cognitive linguistics.

5.3. Cutting-edge corpus methodology - Behavioral profile (BP)

Behavioural Profile (BP) in corpus linguistics is a cutting-edge approach to describe how language users use language. It covers several aspects of language behaviour and has wide application and research value in linguistics, psychology, natural language processing, human-computer interaction and other fields. By analysing behavioural profiles features, we can better understand the nature of language and human behaviour[34,35]. This section organises the documents on Behavioural Profile (BP), a cutting-edge corpus linguistics methodology, in the WOS database and explores its hotspots and development trends.

5.4. Keywords of behavioral profile (BP) research

According to the two indicators of keyword frequency and centrality, we can get the research hotspots in the related fields and the sudden decrease or increase in the citation frequency of documents in different periods, i.e., the significant shift of research hotspots. As shown in **(Table 5 and Figure 7)**:

Table 5: Table of keywords ranking top 10 in frequency and centrality.

Frequency	Year	Keywords	Centrality	Year	Keywords
18	2006	Behavioral Profile	0.86	2006	Behavioral Profile
4	2015	lexical semantics	0.34	2006	cluster analysis
3	2006	cluster analysis	0.19	2015	lexical semantics
2	2013	converging evidence	0.17	2018	construction grammar
2	2021	semantic change	0.14	2013	converging evidence
2	2023	conceptual structure	0.11	2021	semantic change
2	2010	internal semantic structure	0.07	2015	perception verbs
2	2023	syntax	0.04	2023	correspondence analysis
2	2018	construction grammar	0.00	2023	conceptual structure
2	2015	perception verbs	0.00	2023	syntax

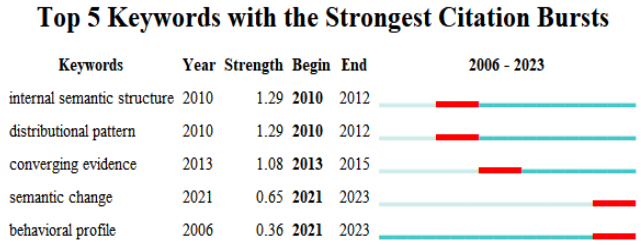

Figure 7: Keyword burst mapping of Behavioural Profile (BP).

(Table 6) shows that in the research of corpus frontier methodology, Behavioral Profile (BP), "Behavioural Profile" "c luster analysis" "lexical semantics" and "construction grammar" are the important support points of the keyword co-occurrence network and their centrality is 0.86, 0.34, 0.19 and 0.17, respectively. The keyword hotspots are mainly related to the application of corpus methodology ("Behavioral Profile" "cluster analysis" "correspondence analysis" "converging evidence"), syntactic, semantic and pragmatic features' static and dynamic evolution of language ontology ("lexical semantics" "semantic change" "conceptual structure" "internal semantic structure" "syntax" "construction grammar" "perception verbs"), etc.

(Figure 7) shows that with the booming development of science, technology and big data, the research content of Behavioural Profile (BP) is more inclined to the dynamic evolution of linguistic features or to explore the interaction of different linguistic features. Liu explored the diachronic semantic changes of the Chinese temperature word "re/heat" by using the corpus-based behavioral profile analysis method[36], combined with multi-factor feature analysis and frequency-based quantitative analysis. The methodology extends the traditional BP method of hierarchical coalescent clustering analysis and employs multiple correspondence analysis to explain and visualize the effectiveness of lexical and semantic changes. It also demonstrates the important role of socio-cultural factors in semantic change.

5.5. Behavioral profile (BP) research trends

To visually present the evolution and development trends of keywords in different periods, we visualize the time zone of keywords in the study of Behavioral Profile (BP), a cutting-edge corpus method, to explore its research mainstream in each stage **(Figure 8)**.

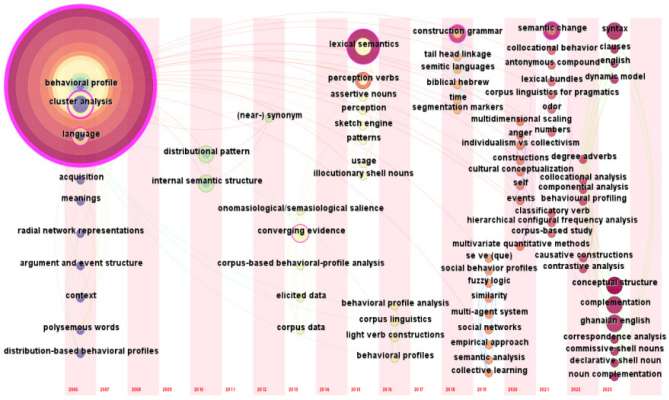

Figure 8: Keyword time zone map of Behavioral Profile (BP).

(Figure 8) shows that the development of Behavioral Profile (BP) research can be divided into three stages: the first stage is from 2006 to 2012 and Behavioral Profile (BP) is mainly

applied to the language ontological, such as analyzing semantic features of certain words, distribution of syntactic structures or pragmatic functions or explore the interaction of different language features, which belongs to the static study of language as a whole. In the second stage from 2013 to 2019, the research scope of Behavioral Profile (BP) is further broadened. Although the research object still focuses on the language ontology, the research scope involves different disciplines, such as language acquisition and teaching, sociolinguistics, etc., to explore the diversity of the linguistic dimension in different fields. The third stage is from 2020 to the present. With the development of computer technology, the Behavioral Profile (BP) method is not only applied to the dynamic evolution of language, but also combines a variety of different corpus methodological tools or computer-assisted functions, which makes its analysis and interpreting functions more powerful. The study of the target language is more thorough, logical and persuasive. These methods and tools can also help linguistics researchers gain a deeper understanding of the complexity and dynamics of language.

6. Conclusion

Based on the BERTopic topic model, this study explores the focused topics, cutting-edge theories and methods of international corpus linguistics, as well as its research hotspots and future development trends, by organising all the papers published in 2 international corpus linguistics journals from 2019 to 2024. The results show that:

The current research in corpus linguistics mainly focuses on the language ontology, followed by language acquisition and teaching, sociolinguistics and cross-cultural communication, computers and multimodal technology, etc. With the development of science and technology, the corpus methods and tools are getting more and more perfect and the interpretation is becoming more powerful.

The cutting-edge theories used for interpretation in corpus linguistics mainly exist in cognitive linguistics, sociolinguistics and cross-cultural communication, such as conceptual metaphors, construction grammar theory, prototype category theory, speech act theory, etc. The cutting-edge corpus methods mainly come from the fusion of various corpus methodologies and technologies or the combination of computer modelling to make the application of corpus methods and tools more extensive.

The cutting-edge theory and methodology of corpus linguistics, such as the advanced cognitive linguistic theory, Behavioral Profile (BP) approach, etc., not only provide a brand-new research perspective and methodology for corpus linguistics research, but also promote the innovative development of linguistic research.

In short, the emergence of corpus linguistics has transformed the study from a traditional model based on intuition and experience to a science based on evidence and data. Through fine-grained data analysis and contextual interpretation, corpus linguistics can reveal the deep rules and characteristics of language and help us understand the nature of language more deeply. The empirical-based research method not only improves the objectivity and accuracy of linguistic research, but also injects new vitality into linguistic research. At the same time, corpus linguistics is of great significance in promoting the progress of NLP, serving language teaching and language learning, as well as enhancing the construction of multilingual corpora.

7. References

1. Park H, Nam D. Corpus linguistics research trends from 1997 to 2016: A co-citation analysis. Linguistic Research, 2017;34: 427-457.
2. Jansegers M, Gries ST. Towards a dynamic behavioural profile: A diachronic study of polysemous sentir in Spanish. Corpus Linguistics and Linguistic Theory, 2017;13: 145-187.
3. Hwang H. Syntactico-semantic realizations of pronouns in the English transitive construction: A corpus-based analysis. Corpus Linguistics and Linguistic Theory, 2022;18: 115-143.
4. Liesenfeld A, Liu MC, Huang CR. Profiling the Chinese causative construction with rang, shi and ling using frame semantic features. Corpus Linguistics and Linguistic Theory, 2022;18: 263-306.
5. Deng KF, Deng YH. A study on the thematic focus of Chinese characteristic discourse based on the complex adaptive systems theory. Social Sciences in Hunan, 2024;2024: 148-154.
6. Deng KF, Deng YH. The diachronic evolution of the verbalized Hong/Red + Object construction and its cognitive motivations. Journal of Xi'an International Studies University, 2024;32; 55-60.
7. Deng YH, Deng KF. Towards a dynamic behavioural profile of the Chinese verbalized color words: A case study of "Hong/Red" "Bai/White" and "Hei/Black." Foreign Languages in China, 2023;2023: 38-47.
8. Stamatovic MV. Vocabulary complexity and reading and listening comprehension of various physics genres. Corpus Linguistics and Linguistic Theory, 2020;16: 487-514.
9. Callies M, Simsek T. Corpus linguistics for English teachers: New tools, online resources and classroom activities. International Journal of Corpus Linguistics, 2020;25: 230-234.
10. Deng YH, Xu QA. An empirical study on the data-driven acquisition of the English "colour word + object" construction. Technology Enhanced Foreign Language Education, 2022;2022: 59-65.
11. Xu JJ, Kang H. Critical contingency competition in L2 clause positioning acquisition: The case of concessive clause by Chinese EFL learners. Corpus Linguistics and Linguistic Theory, 2023;19.
12. Wu SQ. A corpus-based study of the time orientation of qian front and hou back in Chinese. Corpus Linguistics and Linguistic Theory, 2022;18: 447-475.
13. Deng YH, Cheng LY, Xu QA. A diachronic study of the semantic and pragmatic tendencies of Chinese and English psychological adjective + object constructions. Foreign Language Teaching and Research, 2023;55: 176-188.
14. Tsakuwa M, Wen X, Lamido I. A chained metonymic approach to ido 'eye' constructional metonymies in Hausa. Cognitive Linguistics, 2023;34: 165-196.
15. Toth M. A case for metonymic synesthesia. Review of Cognitive Linguistics, 2023;21.
16. Gries ST. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. Corpus Linguistics and Linguistic Theory, 2020;16: 617-647.
17. Hilpert M, Saavedra DC. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. Corpus Linguistics and Linguistic Theory, 2020;16: 393-424.
18. Zimmermann R. An improved test of the constant rate hypothesis: Late Modern American English possessive have. Corpus Linguistics and Linguistic Theory, 2023;19: 323-352.
19. Deng YH, Xu QA, Luo J. Automatic recognition and feature research of discourse with Chinese characteristics based on T5 language model. Foreign Languages in China, 2024: 58-67.

20. Zhu L. Kinship metaphors in the Chinese construction A shi B zhi fu/mu: Biology and culture as conceptual basis. Current Studies in Chinese Language and Discourse: Global Context and Diverse Perspectives, 2019;10: 199-219.
21. Podhorodecka J. Real-life pseudo-passives: The usage and discourse functions of adjunct-based passive constructions. Poznan Studies in Contemporary Linguistics, 2021;57: 33-57.
22. El Shami THS, Shuaibi JA, Zibin A. The function of metaphor modality in memes on Jordanian Facebook pages. Sage Open, 2022;12.
23. Zibin A, Abdullah AD. The conceptualization of tolerance in the UAE press media: A case study of 'The Year of Tolerance.' Open Linguistics, 2019;5: 405-420.
24. Wnuk E, Ito Y. The heart's downward path to happiness: Cross-cultural diversity in spatial metaphors of affect. Cognitive Linguistics, 2021;32: 195-218.
25. Gries ST. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). International Journal of Corpus Linguistics, 2019;24: 385-412.
26. Crosthwaite P, Luciana WD. Exploring language teachers' lesson planning for corpus-based language teaching: A focus on developing TPACK for corpora and DDL. Computer Assisted Language Learning, 2023;36: 1392-1420.
27. Grieve J. Register variation explains stylometric authorship analysis. Corpus Linguistics and Linguistic Theory, 2023;19: 47-77.
28. Egger R, Yu JN. A topic modelling comparison between LDA, NMF, Top2Vec and BERTopic to demystify Twitter posts. Frontiers in Sociology, 2022;7: 1-15.
29. Curry N, Perez-Paredes P. Stance nouns in COVID-19 related blog posts. International Journal of Corpus Linguistics, 2021;26: 469-497.
30. Gries ST. Corpus-based methods and cognitive semantics: The many senses of to run. In S. T. Gries & A. Stefanowitsch (Eds.), Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis, 2006: 57-99.
31. Evans V. Cognitive linguistics: A complete guide. Edinburgh University Press, 2019.
32. Chen J, Meng S, Zhou W. The exploration of fuzzy linguistic research: A scientometric review based on CiteSpace. Journal of Intelligent & Fuzzy Systems, 2019;36: 3655-3669.
33. Lu X, Zhou L, Zhang A, et al. Application of deep learning and intelligent sensing analysis in smart home. Sensors, 2024;24: 1-13.
34. Gries ST. Behavioural profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. The Mental Lexicon, 2010;5: 323-346.
35. Levshina N. How to do linguistics with R: Data exploration and statistical analysis. John Benjamins, 2015.
36. Liu ML. Towards a dynamic behavioural profile of the Mandarin Chinese temperature term re. Corpus Linguistics and Linguistic Theory, 2023;19: 289-321.
37. Kermer F. Semantic network of the German preposition HINTER. Review of Cognitive Linguistics, 2021;19: 403-428.