BAHASTRA
Vol. 45 (1) 2025, pp. 73-90
P-ISSN: 0215-4994 | E-ISSN: 2548-4583
http://journal1.uad.ac.id/index.php/BAHASTRA/index

Corpus-driven language learning: A scientometric analysis of contemporary trends and their trajectories for pedagogical innovation within the Indonesian context

Danang Satria Nugraha [a, b, 1, *]
[a] Sanata Dharma University, Yogyakarta, Indonesia
[b] University of Szeged, Hungary

[1] d.s.nugraha@usd.ac.id
* Correspondent author

Received: January 6, 2025 **Revised:** April 20, 2025 **Accepted**: April 22, 2025

KEYWORDS
Corpus-Driven
Language
Learning
Indonesia
Pedagogical
Innovation

ABSTRACT

In the contemporary landscape of language education, the integration of technology and data-driven approaches has emerged as a significant catalyst for pedagogical innovation. This study investigates contemporary trends and trajectories in corpus-driven language learning (CDLL) through a scientometric analysis of 1,118 journal articles indexed in Scopus from 2014 to 2024. This study selected articles that substantively explored CDLL principles and applications using empirical methodologies, excluding those lacking direct CDLL focus or empirical data to capture the most recent decade of research. Employing bibliometric analysis and keywords co-occurrence visualization with VOSviewer, the study identified key themes, prominent actors, and emerging patterns in CDLL research, aiming to inform pedagogical innovation within the Indonesian context. Results revealed a significant growth in CDLL publications, with research concentrated in China, the United States, and the United Kingdom. Keyword analysis identified four distinct thematic clusters: (1) computational linguistics and artificial intelligence (37%), highlighting the integration of deep learning, language modeling, and machine translation in language learning; (2) specialized applications of CDLL (22%), particularly in information management; (3) human-centered language learning (21.5%), emphasizing social interaction, cognitive processes, and technology integration; and (4) foundational principles of CDLL (19.5%), encompassing corpus linguistics, language acquisition, and pedagogical practices. These findings underscore the growing prominence of CDLL in language education and its potential to transform pedagogical practices in Indonesia by leveraging technology, promoting learner autonomy, and integrating authentic language data into diverse learning contexts.

© 2025 The Author(s). Published by Universitas Ahmad Dahlan.
This is an open-access article under the CC-BY-SA license.

Introduction

In the contemporary landscape of language education, the integration of technology and data-driven approaches has emerged as a significant catalyst for pedagogical innovation. Corpus-driven language learning (hereafter: CDLL), an approach that utilizes large collections of authentic language data (corpora) to inform language teaching and learning (Crosthwaite, 2019; Flowerdew & Petrić, 2024; Wicher, 2019), has gained increasing prominence in recent years. CDLL offers numerous benefits, including providing learners with exposure to real-world language use (Crosthwaite & Baisa, 2023; Yao, 2019), facilitating data-

driven insights into language patterns (Chen & Jiao, 2019; Crosthwaite et al., 2021), and enabling personalized learning experiences (Crosthwaite & Stell, 2019; Pawlak & Kruk, 2022). This approach holds relevance for the Indonesian context, where the diverse linguistic landscape and the growing need for effective language education necessitate innovative pedagogical solutions.

The importance of CDLL lies in its potential to bridge the gap between theoretical knowledge and practical application in language learning. By analyzing authentic language data, learners gain a deeper understanding of how language is used in real-world contexts (Charles, 2022), thereby fostering a more nuanced comprehension of linguistic patterns beyond prescriptive grammar rules. Enabling them to develop communicative competence and cultural awareness (Liu & Chen, 2023), this direct engagement with authentic language exposes learners to diverse linguistic expressions and cultural nuances embedded within the data. Furthermore, CDLL facilitates data-driven decision-making in language education, allowing educators to tailor instruction to learner needs, identify common errors, and develop targeted learning materials (Alruwaili, 2024; Bal-Gezegin et al., 2022; Ma et al., 2022), this evidence-based approach ensures that pedagogical interventions are informed by actual language use and learner performance. In the Indonesian context, where language diversity and geographical disparities pose significant challenges to language education, CDLL offers a promising avenue for enriching language learning outcomes and promoting educational equity.

Previous research has explored various facets of CDLL, examining its theoretical underpinnings, pedagogical applications, and empirical evidence of its effectiveness. For instance, scholar have investigated how corpus linguistics can inform pedagogical grammar instruction by providing authentic examples of language use (e.g., Esfahani & Ketabi, 2024; Lyu & Han, 2023; Kızıl, 2023). Studies have investigated the use of corpora in language teaching, the development of corpus-based learning materials, and the integration of CDLL with technology-enhanced learning environments. This includes research on concordance tools and data-driven learning activities that enable learners to explore language patterns and develop their linguistic awareness (e.g., Lin, 2021; Muftah, 2023; Pérez-Paredes, 2022). Furthermore, research has examined the impact of CDLL on learner outcomes, including vocabulary acquisition, grammatical development, and communicative competence. Specifically, studies have shown the positive effects of corpus consultation on learners' lexical knowledge and grammatical accuracy (e.g., Flowerdew, 2022; Yu & Altunel, 2023; Zare & Delavar, 2024). These studies provide a valuable foundation for understanding the potential of CDLL to transform language pedagogy.

However, despite the growing body of research on CDLL, there remains a need for a comprehensive overview of the current trends and trajectories within this field. A scientometric analysis of CDLL research can provide valuable insights into the key themes, prominent actors, and emerging research directions, informing future research endeavors and pedagogical practices. Such an analysis can also shed light on the global landscape of CDLL research and its potential implications for specific contexts, such as Indonesia. In Indonesia, the integration of technology into language education is an area of increasing interest, yet the specific application and impact of corpus-driven approaches remain relatively underexplored. Understanding global trends in CDLL can offer insights for Indonesian educators and researchers seeking to leverage authentic language data for pedagogical innovation and curriculum development. Therefore, this study addresses this gap by conducting a scientometric analysis of CDLL research, aiming to answer the following research questions: (a) what are the main thematic cluster and emerging trends within CDLL research from 2014 to 2024; and (b) what are the implications of these trends and thematic clusters for pedagogical innovation in the Indonesian context?

Method

Design

The present study employed a scientometrics approach to analyze the current trends and trajectory of corpus-driven language learning (CDLL) research, aiming to identify key themes, prominent actors, and potential implications for pedagogical innovation within the Indonesian context. Scientometrics, a quantitative research method for analyzing scientific and technological literature (Henneken & Kurtz, 2019; Sooryamoorthy, 2020; Waltman & van Eck, 2019), allows for a comprehensive and objective assessment of research trends, emerging topics, and influential publication within a specific field. By leveraging bibliometric data and visualization techniques, this study sought to provide a macro-level perspective on the CDLL research landscape, informing future research directions and pedagogical practices in language education.

Material

The primary data source for this study was the Scopus database, a comprehensive abstract and citation database of peer-reviewed literature. Scopus was selected due to its extensive coverage of scientific publications across various disciplines, including linguistics, language education, and computer science, all relevant to CDLL research. The data collection process involved a systematic search of the Scopus database using a combination of keywords related to CDLL, such as "corpus linguistics," "language learning," "data-driven learning," and "natural language processing." The search was limited to publications from 2014 to 2024 to capture the most recent trends and developments in the field. The retrieved data included bibliographic information, abstracts, keywords, author affiliations, and funding sponsor. Accordingly, specific criteria for query search process is as follow.

(1) "(TITLE-ABS-KEY (corpus-driven AND language AND learning) OR TITLE-ABS-KEY (corpus-based AND language AND learning) OR TITLE-ABS-KEY (corpus-informed AND language AND learning) OR TITLE-ABS-KEY (data-driven AND language AND learning) OR TITLE-ABS-KEY (language AND learning AND with AND corpora) OR TITLE-ABS-KEY (corpus AND linguistics AND for AND language AND acquisition) OR TITLE-ABS-KEY (using AND corpora AND in AND language AND teaching) OR TITLE-ABS-KEY (corpus-assisted AND language AND learning)) AND PUBYEAR > 2013 AND PUBYEAR < 2025 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (SUBJAREA , "ARTS") OR LIMIT-TO (SUBJAREA , "SOCI") OR LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (LANGUAGE , "English"))"

Procedure

The collected data under underwent a rigorous analysis procedure involving several stages. Initially, the data was cleaned and standardized to ensure consistency and accuracy. For this stage, a raw data of 4,688 relevant journal articles went to exclusion and inclusion phase. The inclusion criteria primarily focused on articles that substantively addressed the core aspects of CDLL and employed empirical research methodologies. Conversely, exclusion criteria encompassed articles lacking a direct focus on CDLL, such as those primarily discussing broader educational technologies without specific language learning applications, as well as those employing only theoretical-based approaches without empirical data. The criteria have been applied to the stage and it produce only 1,118 articles for the subsequent phase of analysis. A keyword analysis was then conducted with a minimum occurrence threshold of 30, resulting in 200 keywords selected from a total of 4,258. These keywords were further filtered based on the total strength of their co-occurrence links with other keywords, using VOSviewer (1.6.20), a bibliometric visualization software (van Eck & Waltman, 2023). The parameters for this analysis were: type of analysis is co-occurrence; full counting method; and unit of analysis: all keywords. This process generated network visualizations and identified thematic clusters based on keyword co-occurrence, providing a graphical representation of the relationships between different research themes and highlighting the key areas of focus within CDLL research. By combining quantitative analysis with qualitative interpretation of thematic clusters, this study aimed to provide a comprehensive and well-developed understanding of the current state and future trajectories of CDLL research. Research activity can be seen in Fig. 1.

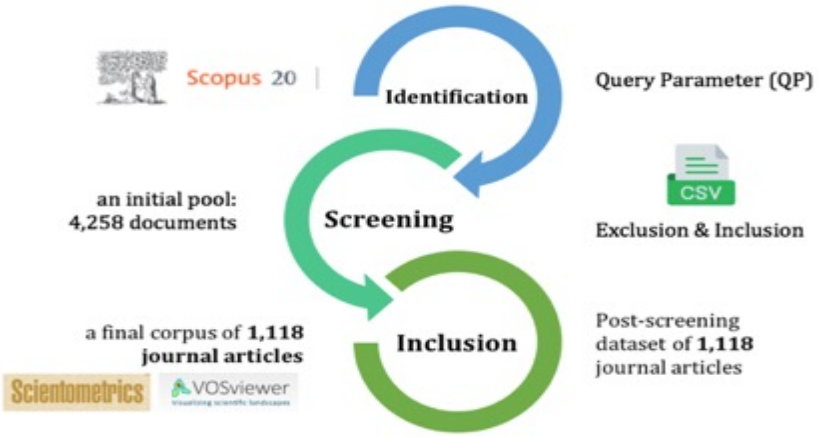

Fig. 1. Research Activity of the Current Study

Results and Discussion

This section presents the key findings derived from the scientometric analysis of corpus-driven language learning research. It outlines the yearly publication trends, highlighting the prominent journals and influential authors that have shaped the field. Furthermore, it explores the distribution of research across various subject areas, affiliations, countries, and funding sponsors. This comprehensive overview provides valuable insights into the current landscape of corpus-driven language learning research and its potential implications for pedagogical innovation within Indonesian context. The subsequent Discussion section will delve deeper into these findings, examining their significance and offering potential avenues for future research and practice.

Documents by Year

Based on our analysis, Figure 2 illustrates a pronounced upward trend in the number of publications related to corpus-driven language learning between 2014 and 2024. The graph depicts a steady increase in yearly output, with notable acceleration in publication frequency from 2020 onwards. This surge suggests a burgeoning interest in the field, potentially attributable to several factors, including the increasing availability of large language corpora, advancements in computational linguistics, and a growing recognition of the pedagogical benefits of data-driven language instruction. The observed trend aligns with broader shifts in language teaching towards more empirical and learner-centered approaches, where authentic language data plays a central role in informing pedagogical practices.

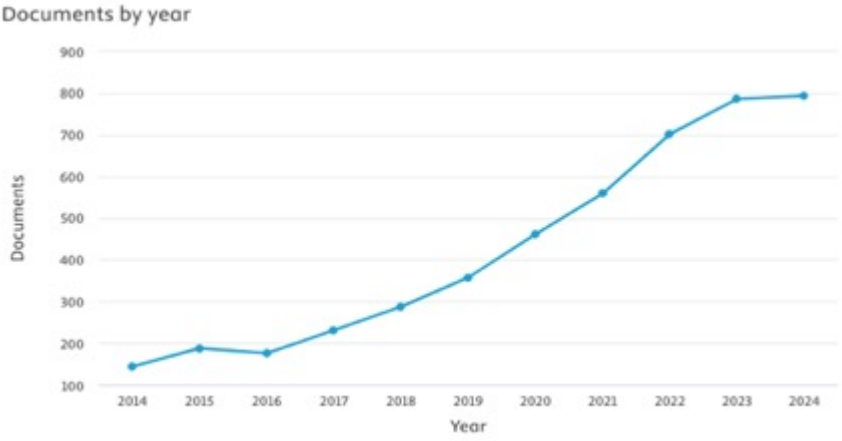

Fig. 2. Publication frequency of CDLL in 2014 – 2024

Furthermore, the sustained growth in publication through 2023 and 2024 indicates that corpus-driven language learning remains dynamic and evolving area of inquiry. This continued momentum underscores the increasing relevance of this approach within the broader landscape of language education, where the integration of technology and data-driven insights holds significant promise for enhancing language acquisition and pedagogical innovation. The exponential growth observed in Figure 2 warrants further analysis into the specific themes and trends driving this expansion (*see* Sub-section: Distribution of the Keywords), as well as their implications for language teaching and learning practices in Indonesia (*see* Sub-section: Discussion).

Documents per Year by Source

According to our analysis, Figure 3 provides a granular view of the distribution of publications across prominent journals within the field of corpus-driven language learning (CDLL) from 2013 to 2024. While several journals contribute to the discourse, "IEEE Access" (Q1 | SJR: 0.96) and "Expert Systems with Applications" (Q1 | SJR: 1.88) demonstrate a clear pattern of growth, particularly after 2020. This suggests an increasing focus on technological applications and computational approaches within CDLL, reflecting the broader trend of integrating technology into language education. Interestingly, "ACM Transactions on Asian and Low Resource Language Information Processing" (Q2 | SJR: 0.54) exhibits a more sporadic publication pattern, indicating a potentially niche focus with CDLL, likely related to specific contexts and computational challenges.

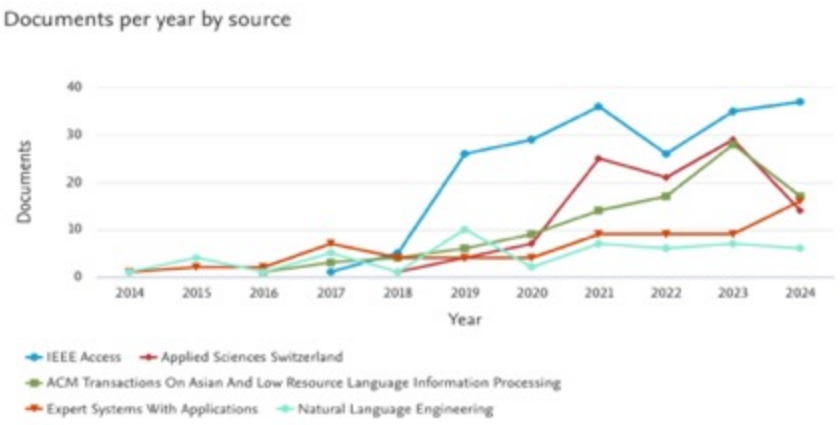

Fig. 3. Journal identification for the topic of CDLL in 2014 – 2024

The figure further reveals a notable increase in publications within "Natural Language Engineering" (Q1 | SJR: 0.66) from 2022 onwards, coinciding with the surge in overall CDLL publications observed in Figure 2. This suggests a growing synergy between corpus linguistics and natural language processing, with implications for the development of sophisticated language learning tools and resources. Conversely, "Applied Sciences Switzerland" (Q2 | SJR: 0.51) shows a relatively stable output, pointing to a consistent, albeit less pronounced, contribution to the field. This analysis of journal-specific publication trends illuminates the diverse avenues through which CDLL research is disseminated and highlights the evolving landscape of scholarly communication within this domain.

Documents by Author

Regarding the leading author(s), Figure 4 presents a visual representation of the most prolific authors contributing to the corpus-driven language learning (CDLL) literature between 2014 and 2024. This author-centric analysis reveals a diverse range of scholars actively shaping the field with contributions spanning various theoretical perspectives, methodological approaches, and language contexts. The prominence of these authors underscores their significant influence in advancing CDLL research and promoting its application in language pedagogy. Furthermore, the figure highlights the collaborative nature of CDLL research, as evidenced by the substantial body of work produced by these leading figures. This collaborative spirit fosters intellectual exchange and accelerates the development of innovative approaches to language teaching and learning informed by corpus data.

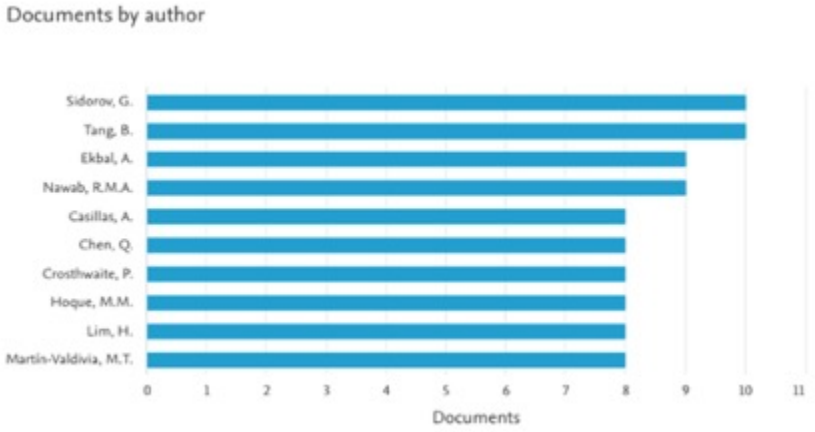

Fig. 4. Notable authors of CDLL in 2014 – 2024

A closer examination of Figure 4 reveals that Sidorov, G. and Tang, B. emerge as particularly influential figures, each with over 10 publications in the analyzed corpus. Their substantial contributions suggest a sustained commitment to CDLL research and a significant impact on the field's trajectory. Following closely are Ekbal, A. and Nawab, R.M.A., whose prolific publication records further underscore the dynamic and evolving nature of CDLL scholarship. Interestingly, the remaining authors, while contributing a slightly

smaller number of publications, still demonstrate a considerable commitment to this research domain. This distribution suggests a healthy and active research community with a diverse range of scholars contributing to the advancement of CDLL.

Documents by Affiliation

Going deeper to analyze the authorship, Figure 5 offers insights into the institutional landscape of corpus-driven language learning (CDLL) research by showcasing the affiliations of notable authors publishing in this domain between 2014 and 2024. The figure highlights a diverse range of institutions contributing to the advancement of CDLL, spanning various geographical regions and academic traditions. This global representation underscores the widespread interest in CDLL and its potential to enhance language pedagogy across diverse contexts. Moreover, the presence of both established and emerging institutions in the figure suggests a dynamic and evolving research landscape, with contributions from both traditional centers of scholarship and institutions at the forefront of innovation in language education.

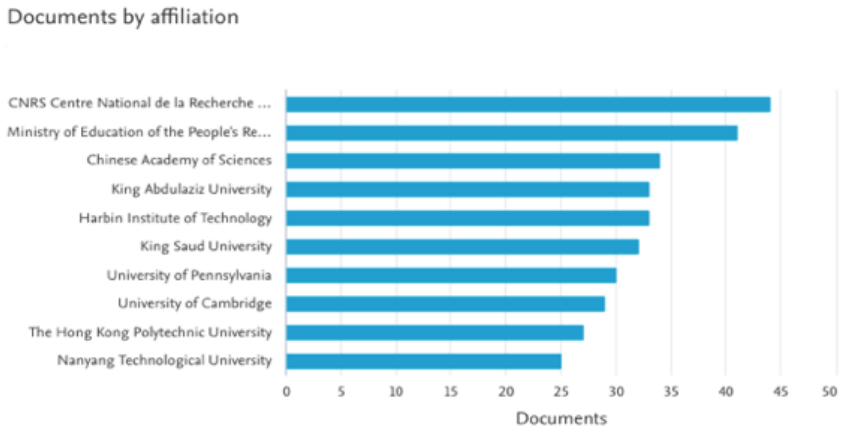

Fig. 5. Notable authors' affiliations of CDLL in 2014 – 2024

A careful scrutiny of Figure 5 reveals a concentration of research activity within specific institutions. Notably, the CNRS (Centre National de la Recherche Scientifique) and the Ministry of Education of the People's Republic of China emerge as leading contributors to CDLL scholarship, indicating a strong emphasis on research in these regions. Furthermore, the presence of several prominent universities, such as King Abdul-Aziz University, Harbin Institute of Technology, and King Saud University, highlights the significant role of higher education institutions driving CDLL research. Interestingly, the figure also reveals contributions from institutions in North America, Europe, and Asia, including the University of Pennsylvania, University of Cambridge, The Hong Kong Polytechnic University, and Nanyang Technological University, underscoring the global reach and collaborative nature of CDLL research. This diverse institutional landscape fosters a rich exchange if ideas and contributes to the development of innovative pedagogical approaches informed by corpus data.

Documents by Subject Area

Figure 6 provides a comprehensive overview of the disciplinary distribution of corpus-driven language learning (CDLL) research from 2014 to 2024, categorized by subject area. The diversity of disciplines represented in the figure underscores the interdisciplinary nature of CDLL, drawing upon insights and methodologies from various fields to advance language pedagogy. This cross-disciplinary fertilization enriches the theoretical foundations of CDLL and promotes innovative approaches to language teaching and learning. Furthermore, the figure highlights the potential of CDLL to bridge the gap between theoretical research and practical applications, as evidenced by the contributions from fields such as computer science, social sciences, and arts and humanities, all of which have direct relevance to language education.

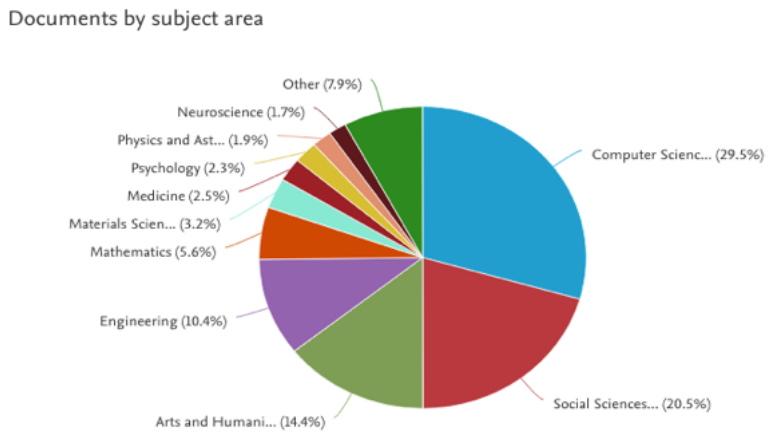

Fig. 6. Notable of subject area for the topic of CDLL in 2014 – 2024

A deeper exploration of Figure 6 reveals that computer science constitutes the largest subject area, accounting 29.5% of the publication analyzed. This finding underscores the growing importance of computational approaches and technological tools in CDLL research and practice. Social sciences follow closely with 20.5% reflecting the significant role of sociological, psychological, and educational perspectives in understanding language learning processes and informing pedagogical interventions. Arts and humanities, encompassing linguistics, literature, and language studies, contribute 14.46%, highlighting the continued importance of core disciplinary knowledge in shaping CDLL research. Interestingly, the figure also reveals contributions from fields such as engineering, mathematics, and even neuroscience, albeit to a lesser extent. This diverse disciplinary landscape reflects the nature of CDLL and its potential to integrate insights from various field to enhance language pedagogy.

Documents by Funding Sponsor

Regarding the funding sponsor of the CDLL research and publication, Figure 7 provides a revealing glimpse into the funding landscape of corpus-driven language learning (CDLL) research, highlighting the key sponsors who have supported investigations in this domain between 2014 and 2024. The diversity of funding sources represented in the figure underscores the growing recognition of CDLL's importance and its potential to drive innovation in language education. The presence of prominent national and international funding bodies signals a commitment to advancing knowledge and promoting research in this field. Furthermore, the figure reflects a global interest in CDLL, with support emanating from various regions and reflecting diverse research priorities. This diversified funding landscape fosters a vibrant and dynamic research environment, enabling scholars to explore a wide range of topics and contribute to the advancement of CDLL.

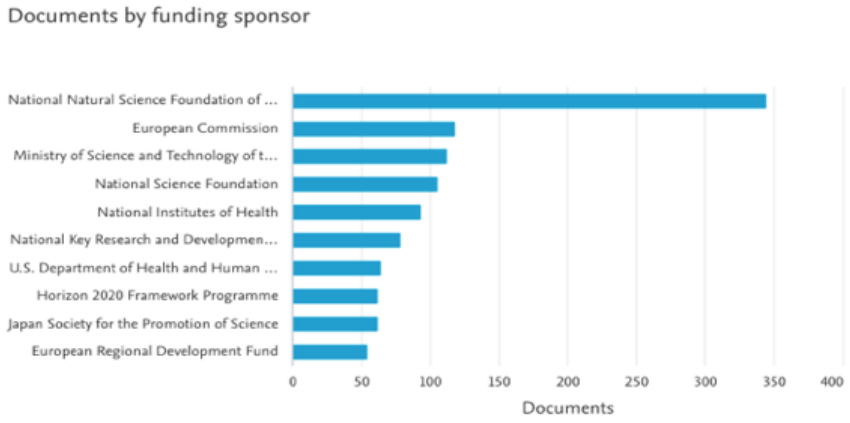

Fig. 7. Notable sponsor for research funding on CDLL in 2014 – 2024

A close reading of Figure 7 reveals that the National Natural Science Foundation of China emerges as the leading funding sponsor for CDLL research, indicating a substantial investment in this area within China. This finding aligns with the prominent role of Chinese institutions and researchers observed in previous figures, highlighting China's commitment to advancing CDLL scholarship. The European Commission also plays a significant role in supporting CDLL research, reflecting a broader European interest in this field. Furthermore, the figure shows contributions from other prominent funding bodies, including the Ministry of Science and Technology of the People's Republic of China, the National Science Foundation, and the National Institutes of Health, underscoring the global significance of CDLL research and its potential to impact language education worldwide. This analysis of funding sponsors provides valuable insights into the priorities and trends shaping CDLL research, while also highlighting the importance of continued investment in this promising field.

Documents by Country

Concerning the nation or country, Figure 8 provides a geographical overview of corpus-driven language learning (CDLL) research, highlighting the countries that have made significant contributions to the field between 2014 and 2024. This figure illustrates a diverse range of countries actively engaged in CDLL research, underscoring the global reach and relevance of this approach to language pedagogy. This widespread participation suggests that CDLL is not confined to specific linguistic or cultural context but rather holds universal appeal for enriching language learning across diverse educational settings. Furthermore, the figure points to a growing international community of scholars dedicated to advancing CDLL research and promoting its application in language classroom worldwide.

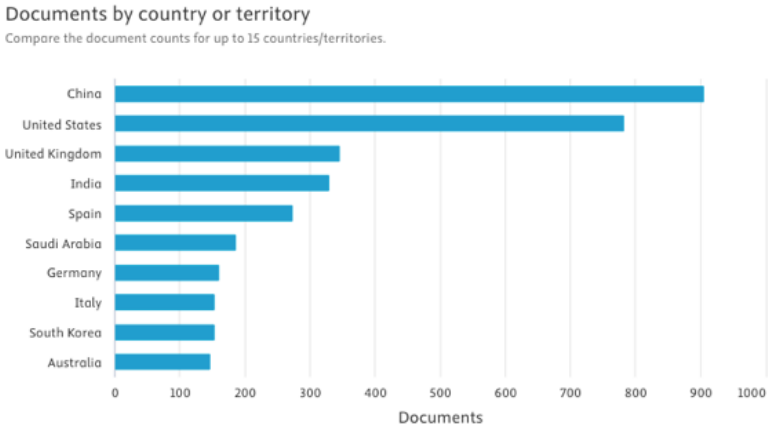

Fig. 8. Notable country for research on CDLL in 2014 – 2024

A detailed examination of Figure 8 reveals that China emerges as the leading contributor to CDLL research, with a substantial number of publications originating from this country. This finding aligns with the prominence of Chinese institutions and funding agencies observed in previous figures, reinforcing the notion that China is at the forefront of CDLL scholarship. The United States follows closely, indicating a string research presence in this domain. The United Kingdom and India also demonstrate significant contributions to CDLL research, further highlighting the global nature of this field. Interestingly, the figure also reveals contributions from countries such as Spain, Saudi Arabia, Germany, Italy, South Korea, and Australia, underscoring the widespread interest in CDLL across diverse linguistic and cultural contexts. This geographical distribution of research activity fosters a rich exchange of ideas and promotes cross-cultural collaboration in the pursuit of innovative pedagogical approaches informed by corpus data.

Documents by the Keywords

Another important analysis of the current study is about the distribution of the keywords. Figure 9 presents a visually compelling representation of the main thematic clusters within corpus-driven language learning (CDLL) research from 2014 to 2024, generated through keyword co-occurrence analysis. This visualization illuminates the complex and interconnected nature of CDLL scholarship, revealing a network of interrelated themes that contribute to the field's dynamism. The distinct clusters, represented by different colors, highlight key areas of focus within CDLL ranging from computational linguistics and natural language processing to language learning and pedagogy. The interconnectedness of these clusters underscores the nature of CDLL, drawing upon insights and methodologies from various fields to advance

language education. This visualization provides a valuable overview of the thematic landscape of CDLL research, revealing the key areas of inquiry that have shaped the field's trajectory over the past decade.

A micro-level examination of Figure 9 reveals four distinct clusters, each representing a major thematic focus within CDLL research. The red cluster centered around "natural language processing" and "deep learning," highlights the growing importance of computational approaches and artificial intelligence in language learning. The green cluster, with keywords such as "electronic health record" and "unified medical language system," suggest an emerging trend of applying CDLL in specialized domain like healthcare. The blue cluster, focused on "human," "language," and "learning," emphasizes the core principles of language acquisition and pedagogy, while the yellow cluster, with terms like "linguistics," "corpus linguistics," and "semantics," reflects the foundational linguistic knowledge underpinning CDLL research. The varying sizes of these clusters and their interconnections reflects the relative prominence and interrelationships of these themes within the CDLL research landscape. This analysis provides valuable insights into the key areas of inquiry driving CDLL research and their potential implications for pedagogical innovations.

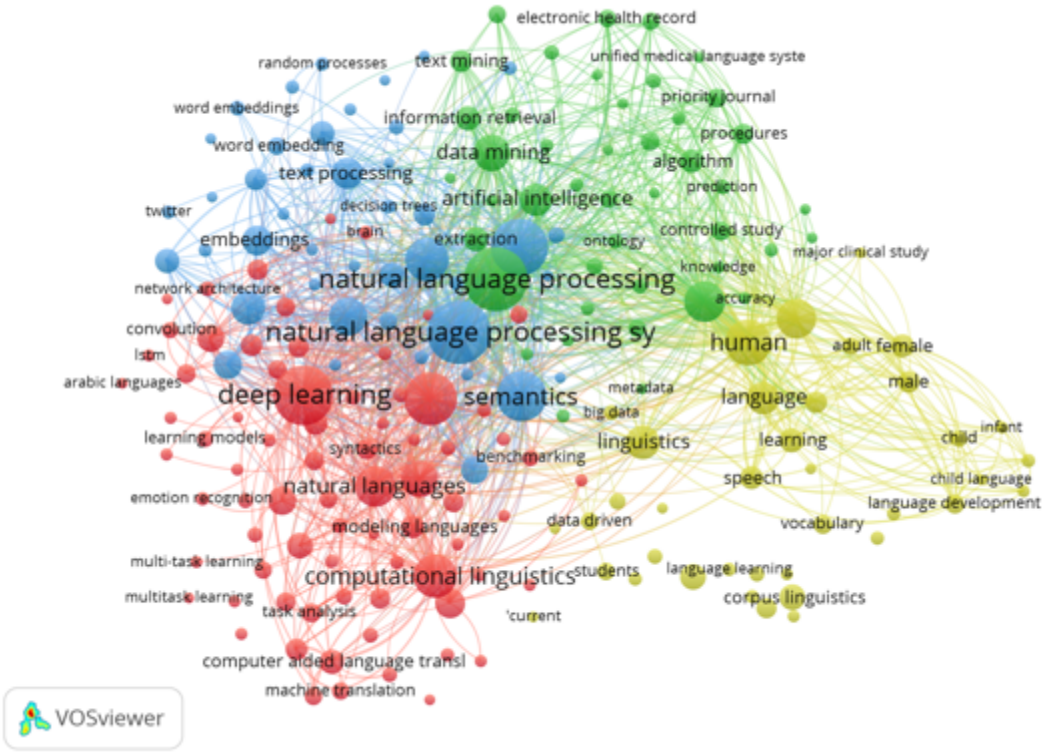

Fig. 9. Main clusters for research on CDLL in 2014 – 2024

In detail, Table 1 presents a detailed breakdown of the keywords constituting Cluster 1 (Red Cluster in Figure 9), which represents a prominent thematic focus within CDLL research. The keywords within this cluster predominantly revolve around computational linguistics and artificial intelligence, highlighting the growing influence of these fields on language learning research and pedagogy. This emphasis in technology-driven approaches reflects a broader trend in education towards leveraging computational tools and resources to enhance language acquisition and instruction. The comprehensive list of keywords provides valuable insights into the specific areas of inquiry within this cluster, ranging from natural language processing and deep learning to machine translation and neural networks. This granular perspective allows for a deeper understanding of the technological advancements driving innovation in CDLL and their potential implications for language pedagogy.

Further analysis of the keywords reveals several distinct sub-clusters within Cluster 1, each representing a more specialized area of focus within the broader theme of computational linguistics and artificial intelligence. One sub-cluster centers around deep learning (1.1), encompassing keywords such as "deep learning," "deep neural network," "deep neural networks," "learning models," "multilayer neural networks," and "neural networks." This sub-cluster highlights the growing importance of deep learning techniques in CDLL research, particularly for tasks such as natural language processing, machine translation, and speech recognition. Another sub-cluster (1.2) focuses on language modeling, with keywords like "language model," "language processing," "large language model," "modeling (languages),"

and "natural languages." This sub-cluster reflects the increasing interest in developing sophisticated language models that can accurately represent and generate human language, with implications for language learning tools and resources. Additionally, keywords such as "machine translation," "machine translations," "neural machine translation," "computer aided language translation," and "translation (languages)" form a sub-cluster related to machine translation (1.3), underscoring the role of technology in facilitating cross-lingual communication and language learning. These sub-clusters, along with others related to specific techniques and applications, provide a nuanced understanding of the diverse research directions within the broader theme of computational linguistics and artificial intelligence in CDLL.

Table 1. Keywords within the four main clusters for research on CDLL in 2014 – 2024

Cluster	*(%)*	*Itemization*
Cluster 1 - Red (74 items)	37	attention mechanism; attention mechanisms; automatic speech recognition; benchmarking; bert; brain; character recognition; codes (symbols); computational linguistics; computer aided language translation; contrastive learning; convolution; convolution neural network; convolution neural networks; cross-lingual; data augmentation; decoding; deep learning; deep neural network; deep neural networks; emotion recognition; errors; feature extraction; features extraction; forecasting; hidden markov models; job analysis; knowledge management; language model; language processing; large dataset; large language model; learn+; learning approach; learning models; learning systems; learning techniques; long short-term memory; low resource languages; lstm; machine translation; machine translations; modeling languages; multi-task learning; multilayer neural networks; multitask learning; natural language processing (nlp); natural languages; network architecture; neural machine translation; neural network; neural networks; neural-networks; parallel corpora; performance; pre-trained language model; pre-training; question answering; recurrent neural network; recurrent neural networks; reinforcement learning; representation learning; sign language; signal encoding; speech processing; speech recognition; state of the art; syntactics; task analysis; transfer learning; transformer; transformers; translation (languages); unsupervised learning
Cluster 2 - Green (44 items)	22	accuracy; active learning; algorithm; algorithms; article; artificial intelligence; artificial neural network; attention; automation; classification; classifier; controlled study; data mining; databases-factual; diagnosis; diseases; electronic health record; electronic health records; embedding; extraction; factual database; information extraction; information processing; information retrieval; information storage and retrieval; knowledge; knowledge base; knowledge based system; learning algorithm; metadata; natural language processing; neural networks; ontology; prediction; priority journal; procedures; quality control; relation extraction; search engines; short term memory; software; support vector machine; text mining; unified medical language
Cluster 3 - Blue (43 items)	21.5	behavioral research; classification; conditional random field; covid-19; data handling; decision making; decision trees; embeddings; graphic methods; information retrieval systems; learning algorithms; machine learning; machine learning approaches; machine learning method; machine learning model; machine learning techniques; named entity recognition; natural language processing; nlp; opinion mining; random processes; semantic similarity; semantics; semi-supervised learning; sentiment analysis; sentiment classification; social media; social networking; statistics; supervised learning; supervised machine learning; support vector machines; text classification; text processing; topic modeling; twitter; vector spaces; word embedding; word embeddings; word representations; word sense disambiguation; word2vec
Cluster 4 - Yellow (39 items)	19.5	academic writing; adult; big data; child; child language; computer aided instruction; corpora; corpus; corpus analysis; corpus linguistics; data driven; data-driven learning; e-learning; education; female; human; human experiment; humans; infant; language; language acquisition; language development; language learning; learner corpus; learning; linguistic features; linguistics; major clinical study; male; multilingualism; preschool child; second language acquisition; speech; students; teaching; vocabulary

Moreover, the sub-cluster identified within Cluster 1 represent specific areas focus within the broader theme of computational linguistics and artificial intelligence in CDLL. Examining these sub-clusters through targeted research questions can further illuminate their role in advancing language pedagogy. As for the sub-cluster 1.1 (deep-learning sub-cluster), one can pay attention on (a) to what extent can deep learning models improve the accuracy and efficiency of automated feedback mechanisms in language learning platforms; (b) how can deep learning be leveraged to personalize language learning experiences and adapt to individual learner needs and preferences; and (c) what are ethical considerations surrounding the use of

deep learning algorithms in evaluating and assessing language proficiency? Next, as for the sub-cluster 1.2 (language modelling sub-cluster), one should focus on (a) how can large language models be used to create more authentic and engaging language learning materials and activities; (b) can language models be effectively employed to stimulate real-life language use and provide learners with opportunities for interactive practice, and (c) what are the limitations of current language models in capturing the nuances and complexities of human language, and how can these limitations be addressed in the context of CDLL? Lastly, regarding the sub-cluster 1.3 (machine translation sub-cluster), one may put a close attention on (a) how can machine translation technologies be integrated into language learning classrooms to support learners in comprehending and producing texts in their target language; (b) to what extent can machine translation be used to facilitate cross-lingual communication and collaboration among language learners; and (c) what are the pedagogical implications of using machine translation in language learning, and how can educators effectively address potential challenges and misconceptions?

In addition, Table 1 also presents detailed analysis results on Cluster 2, depicted as the Green Cluster in Figure 9. This cluster represents a distinct thematic focus within CDLL research, characterized by its emphasis on applications in specialized domains. The keywords reveal a convergence of CDLL with fields such as medical informatics and natural language processing, highlighting the potential of corpus-based approaches to enhance information extraction, knowledge management, and decision-making in healthcare settings. This interdisciplinary focus underscores the versatility of CDLL and its capacity to address real-world challenges beyond traditional language education contexts. The table's comprehensive list of keywords offers valuable insights into the specific areas of inquiry within this cluster, including electronic health records, natural language processing, information retrieval, and knowledge-based systems, among others.

A deeper exploration of the keywords reveals several discernible sub-clusters within Cluster 2, each representing a more specialized area of investigation within the broader theme of CDLL application in healthcare. One prominent sub-cluster (2.1) centers around electronic health records (EHRs), encompassing keywords such as "electronic health record," "electronic health records," "information extraction," "information retrieval," and "information storage and retrieval." This sub-cluster highlights the growing interest in leveraging CDLL techniques to extract meaningful information from EHRs, potentially for task such as clinical decision support, patient monitoring, and disease surveillance. Another sub-cluster focuses on knowledge management and representation (2.2), with keywords like "knowledge," "knowledge base," "knowledge-based system," "ontology," and "unified medical language." This sub-cluster reflects the importance of organizing and structuring medical knowledge in a way that can be effectively utilized by both humans and machines, with potential applications in medical education, clinical research, and patient care. Furthermore, keywords such as "natural language processing," "text mining," "data mining," and "algorithms" form a sub-cluster (2.3) related to the computational analysis of medical language data, underscoring the role of natural language processing techniques in extracting insights from clinical texts and facilitating knowledge discovery in healthcare. These sub-clusters, along with others related to specific applications and methodologies, provide a nuanced understanding of the diverse research directions within the broader theme of CDLL application.

Furthermore, the sub-clusters identified within Cluster 2 (Green Cluster) point towards specialized application of CDLL, particularly in information management and knowledge extraction, with implications for language learning beyond the healthcare domain. Exploring these sub-clusters through focused research questions can help uncover their potential for pedagogical innovation. As for the first sub-cluster (2.1) (EHRs sub-cluster), the questions can be (a) how can CDLL techniques be utilized to automatically extract and organize key linguistic features from authentic texts for the development of language learning materials; (b) can CDLL-based information retrieval systems be employed to personalize the selection of learning resources based on individual learner needs and preferences; and (c) how can the analysis of learner-generated texts, analogous to EHRs, inform the development of adaptive language learning systems that respond to individual learning patterns and difficulties? Next, for the second sub-cluster (2.2) (knowledge management sub-cluster), one can pay focus on (a) how can CDLL be used to create comprehensive and interconnected knowledge bases of linguistic information that can be accessed and utilized by language learners; (b) can ontologies, informed by corpus data, be developed to represent complex linguistic concepts and relationships in a way that is accessible and meaningful to language learners; and (c) how can CDLL contribute to the development of intelligent tutoring systems that provide personalized feedback and guidance based on a deep understanding of linguistic structures and patterns? Lastly, as for the third sub-cluster (2.3) (computational analysis of language data sub-cluster), the research questions can be (a) how can text mining and natural language processing techniques be employed to analyze learner language and identify patterns of errors and difficulties; (b) can CDLL-based computational analysis reveal hidden relationships between linguistic features and learner proficiency, leading to more

effective language assessment and feedback; and (c) what are the ethical considerations surrounding the use of algorithms and automated analysis of learner language data in educational settings?

Furthermore, concerning Cluster 3, Table 1 also delineates the keywords associated with the blue cluster in Figure 3. This cluster encapsulates a core thematic focus within CDLL research, emphasizing the human aspects of language learning and the integration of technology to enhance pedagogical practices. The keywords encompass a wide range of concepts related to language acquisition, cognitive processes, social interaction, and technological applications in education. This perspective underscores the importance of considering the unique interplay between human cognition, social dynamics, and technological advancements in the design and implementation of effective CDLL interventions. The comprehensive list of keywords provides valuable insights into the specific areas of inquiry within this cluster, including machine learning, natural language processing, sentiment analysis, social media, and various learning algorithms and techniques. Overlay visualization can be seen in Fig. 10.

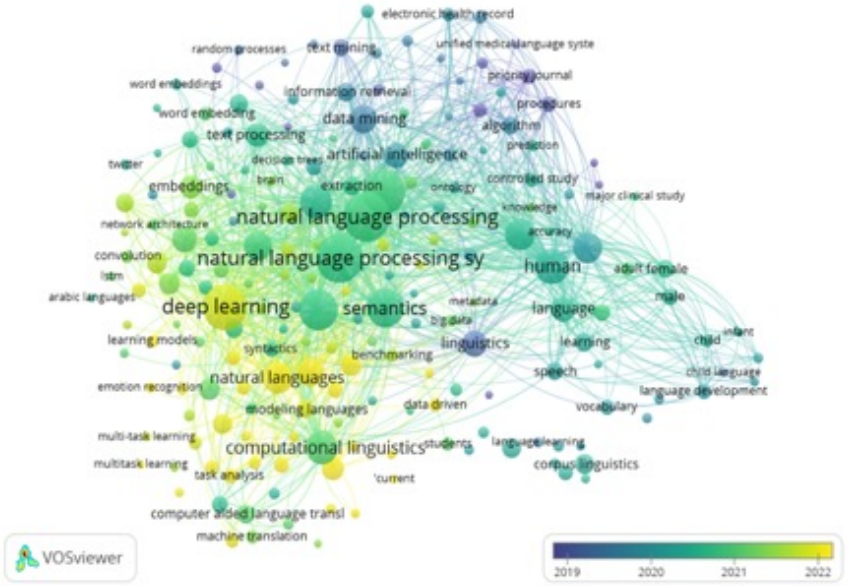

Fig. 10. Overlay visualization of the main clusters for research on CDLL in 2014 – 2024

A close reading of the keywords reveals several distinct sub-clusters within Cluster 3, each representing a more specialized area of focus within the broader theme or human-centered language learning and its applications in education. One sub-cluster centers around machine learning and its applications in education (3.1), encompassing keywords such as “machine learning,” “machine learning approaches,” “machine learning method,” “machine learning model,” “machine learning techniques,” “supervised learning,” “supervised machine learning,” and “learning algorithms.” This sub-cluster highlights the growing interest in leveraging machine learning algorithms to personalize learning experiences, provide adaptive feedback, and optimize instructional strategies. Another sub-cluster focuses on social and interactive aspects of language learning (3.2), with keywords like “social media,” “social networking,” “twitter,” and “behavioral research.” This sub-cluster reflects the increasing recognition of the role of social interaction and online communication in language acquisition and the potential of CDLL to harness these dynamics of pedagogical purposes. Additionally, keywords such as “natural language processing,” “nlp,” “text processing,” “sentiment analysis,” “semantic similarity,” and “semantics” form a sub-cluster (3.3) related to the computational analysis of human language, underscoring the importance of natural language processing techniques in understanding and facilitating language learning. These sub-clusters, along with others related to specific learning techniques and cognitive processes, provide a nuanced understanding of the diverse research directions within the broader theme of human-centered language learning and technology integration in CDLL.

Furthermore, the sub-clusters within Cluster 3 (Blue Cluster) reveal notable research directions focused on the human aspects of language learning and the integration of technology in pedagogical practices. Formulating targeted research questions for each sub-cluster can further illuminate their potential contributions to CDLL and pedagogical innovation. As for the sub-cluster 3.1 (machine learning in education sub-cluster), one may put attention on the (a) how can machine learning algorithms be employed to identify and address individual learner needs and learning styles in CDLL contexts; (b) can machine learning be used to develop adaptive language learning platforms that dynamically adjust the

difficulty and content of learning materials based on learning performance; and (c) what are the ethical considerations surrounding the use of machine learning algorithms in making decisions about learner progress and assessment in CDLL? Next, as for the sub-cluster 3.2 (social and interactive language learning sub-cluster), one can put attention on the (a) how can CDLL be integrated with social media platforms and online communities to foster authentic communication and collaboration among language learners; (b) can the analysis of learner interactions on social media inform the design of more effective CDLL activities that promote engagement and motivation; and (c) what are the challenges and opportunities associated with using social media data in CDLL research, particularly concerning privacy and ethical considerations? Lastly, as for the sub-cluster 3.3 (computational analysis of human language sub-cluster), one can focus on (a) how can natural language processing be used to analyze learner language and provide personalized feedback on linguistic accuracy and complexity; (b) can sentiment analysis and opinion mining techniques be employed to gauge learner engagement and motivation in CDLL activities; and (c) to what extent can computational analysis of language data reveal underlying cognitive processes involved in language acquisition and inform the design of more effective CDLL interventions?

As the last analysis results, Table 1 also enumerates the keywords associated with Cluster 4, visually represented as the yellow cluster in Figure 9. This cluster encapsulates a fundamental thematic focus within CDLL research, emphasizing the core principles of linguistics, language acquisition, and pedagogical applications. The keywords encompass a broad spectrum of concepts related to language development, language learning processes, learner characteristics, and pedagogical approaches. This comprehensive perspective underscores the importance of grounding CDLL research in a solid understanding of linguistic theory, empirical findings on language acquisition, and effective pedagogical practices. The Table 1's detailed list of keywords offers valuable insights into the specific areas of inquiry within this cluster, including corpus linguistics, language acquisition, language development, learner corpora, and various learner demographics and learning contexts. Density visualization can be seen in Fig. 11.

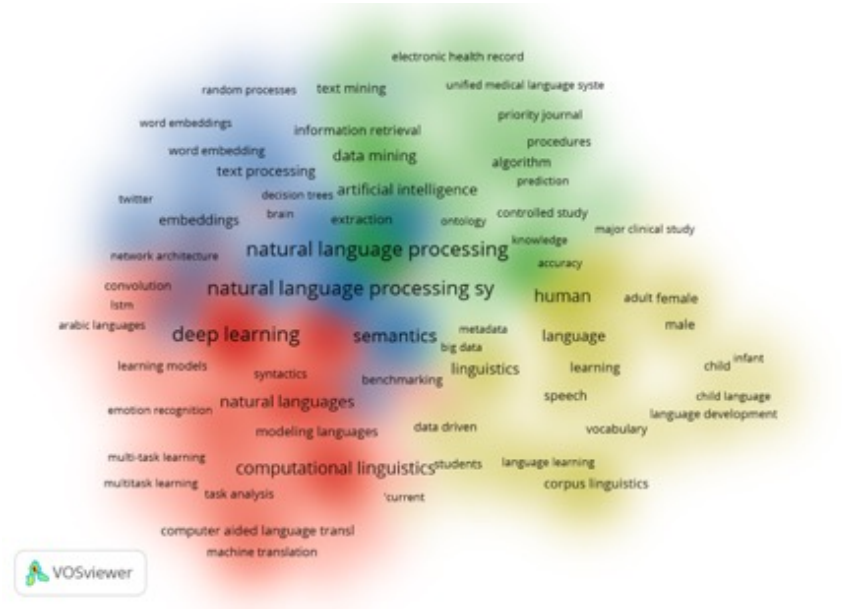

Fig. 11. Density visualization of the four main clusters for research on CDLL in 2014 – 2024

A critical examination of the keywords reveals distinct sub-clusters within Cluster 4, each representing a more specialized area of focus within the broader theme of linguistics, language acquisition, and pedagogy. One sub-cluster (4.1) centers around corpus linguistics and its applications in language learning, encompassing keywords such as "corpora," "corpus," "corpus analysis," "corpus linguistics," "data-driven," "data-driven learning," and "linguistic features." This sub-cluster highlights the importance of utilizing corpus data to inform language teaching and learning, providing authentic examples of language use and insights into language patterns and variations. Another sub-cluster (4.2) focuses on language acquisition and development, with keywords like "language acquisition," "language development," "second language acquisition," "child," "child language," "infant," and "preschool child." This sub-cluster reflects the ongoing investigation into the processes and stages of language learning across different age groups and developmental contexts. Additionally, keywords such as "learning," "learner corpus," "students," "teaching," and "computer-aided instruction" form a sub-cluster (4.3) related to pedagogical practices and learner

characteristics, emphasizing the importance of learner-centered approaches and the integration of technology in language education. These sub-clusters, along with others related to specific linguistic concepts and learner demographics, provide a nuanced understanding of the diverse research directions within the broader theme of linguistics, language acquisition, and pedagogy in CDLL.

Therefore, the sub-cluster within Cluster 4 underscore the foundational role of linguistic theory, language acquisition research, and pedagogical practices in CDLL. Examining these sub-clusters through specific research questions can further illuminate their contribution to advancing language pedagogy. As for the sub-cluster 4.1 (corpus linguistics and language learning sub-cluster), one may put attention on (a) how can corpus analysis be used to identify and prioritize linguistic features that are most relevant for learners at different proficiency levels; (b) can data-driven learning approaches, based on corpus evidence, enhance learner understanding of language variation and register; and (c) how can corpus linguistics inform the development of language learning materials that reflect authentic language use and address learner needs? Beside, as for the sub-cluster 4.2 (language acquisition and development sub-cluster), one may pay focus on (a) how can insights from child language acquisition research inform the design of CDLL interventions for young learners; (b) what are the similarities and differences in the language acquisition trajectories of first and second language learners, and how can CDLL address these specific needs; and (c) can the analysis of learners corpora reveal developmental patterns in language acquisition and inform personalized feedback and instruction? Lastly, as for the sub-cluster 4.3 (pedagogical practices and learner characteristics sub-cluster), one may pay a close attention to (a) how can CDLL be integrated with computer-aided instruction to create engaging and effective language learning environments; (b) what are the optimal pedagogical strategies for utilizing corpus data and technology to enhance language learning outcomes; and (c) how can learner characteristics, such as motivation, learning styles, and prior knowledge, be considered in the design and implementation of CDLL interventions?

Discussion

The scientometric analysis presented in this study provides a comprehensive overview of the current landscape of corpus-driven language learning (CDLL) research, highlighting key trends, prominent actors, and thematic foci within this field (*see also* Figure 10 and Figure 11). The observed growth in publications, the diversity of contributing institutions and countries, and the interdisciplinary nature of the research underscore the increasing prominence of CDLL in language education. This expanding interest in CDLL aligns with the broader shift towards data-driven and learner-centered approaches in language pedagogy (Lusta et al., 2023), where authentic language data plays a central role in informing instructional practices (cf. Chen & Flowerdew, 2018; Gardner, 2024). The findings of this analysis have significant implications for language education in Indonesia, where the integration of CDLL holds considerable promise for enhancing language acquisition and promoting pedagogical innovation. Specifically, these global trends in CDLL research offer a foundation for Indonesian scholars and educators to explore and adapt corpus-based methodologies to address the unique challenges and opportunities within the Indonesian educational context, such as its multilingualism and diverse learning environments (Nugraha et al., 2025).

The prominence of computational linguistics and artificial intelligence in CDLL research, as evidenced by Cluster 1 (Red Cluster), signals the growing importance of technology in language learning. In line with Alruwaili (2024) and Crosthwaite (2017), the application of deep learning, language modeling, and machine translation techniques offers exciting possibilities for creating personalized and adaptive learning experiences, automating feedback mechanisms, and facilitating cross-lingual communication. Within the Indonesian context, embracing these technological advancements could revolutionize language education by providing learners with access to sophisticated language learning tools and resources, particularly in geographically dispersed or resource-constrained settings. However, successful implementation necessitates careful consideration of the existing digital infrastructure and the digital literacy levels of both educators and learners across the diverse Indonesian demographic area. For instance, AI-powered language learning platforms could offer personalized instruction and feedback to students in remote areas with limited access to qualified language teachers. Furthermore, the development of Indonesian language corpora and natural language processing tools could facilitate the creation of culturally relevant and engaging learning materials, catering to the diverse linguistic landscape of the archipelago. Therefore, strategic investment in technological infrastructure and targeted training programs for educators are crucial to effectively harness the potential of AI and computational linguistics to enhance CDLL in Indonesia. This would not only enhance the quality of language education but also promote educational equity by bridging the gap between urban and rural learners.

Cluster 2 (Green Cluster) highlights the potential of CDLL to extend beyond traditional language education contexts and address real-world challenges in specialized domains. As noted by Hu et al. (2016),

Mamta et al. (2022), and Saeed et al. (2022), the focus on information extraction, knowledge management, and computational analysis of language data has implications for various fields, including healthcare, legal, and business sectors. In the Indonesian context, this suggests opportunities for integrating CDLL into vocational training programs and professional development initiatives, equipping learners with the skills to analyze and utilize language data in their respective fields. However, the successful adoption of CDLL in these specialized domains within Indonesia necessitates the development of relevant Indonesian language corpora and the adaptation of existing analytical tools to specific linguistic characteristic of Indonesian language (e.g., Nugraha, 2021, 2024a, 2024b). For instance, tourism professionals could benefit from training in corpus-based analysis of online reviews to understand tourist perceptions and tailor services accordingly. Similarly, legal professionals could utilize CDLL techniques to analysis legal documents and extract relevant information for case preparation. Therefore, collaborative efforts between language education specialists and professionals in various sectors are crucial to identify specific needs and develop targeted CDLL applications and training modules relevant to the Indonesian professional landscape. By incorporating CDLL into vocational training, Indonesia can cultivate a workforce equipped with advanced language skills and data analysis capabilities, enhancing their competitiveness in the global market.

Moreover, Cluster 3 (Blue Cluster) emphasizes the human aspects of language learning, recognizing the importance of social interaction, cognitive processes, and affective factors in language acquisition. According to Balouchzahi et al. (2023), Liu et al. (2015), and Nugraha (2024c), the integration of machine learning with social media analysis and natural language processing techniques offers new avenues for understanding learner behavior, personalizing feedback, and fostering engagement. For Indonesian language education, this underscores the need for pedagogical approaches that not only leverage technology but also cultivate learner autonomy, collaboration, and authentic communication. However, the effective integration of these human-centered approaches within the Indonesian context necessitates careful consideration of cultural norms and communication styles prevalent across its diverse regions. Encouraging learners to actively participate in online language learning communities, for instance, can foster collaborative learning and expose them to diverse language varieties within Indonesia. Therefore, pedagogical strategies should be adapted to promote culturally sensitive online interactions and leverage digital tools to facilitate meaningful communication that respects local customs and values. Furthermore, incorporating project-based learning activities that require authentic communication, such as creating podcasts or collaborating on digital storytelling projects, can enhance motivation and provide opportunities for meaningful language use. By integrating these pedagogical approaches, Indonesian language educators can empower learners to take ownerships of their language development and become active participants in the digital world.

Finally, Cluster 4 (Yellow Cluster) reinforces the foundational role of linguistic theory, language acquisition research, and pedagogical practices in CDLL. In line with Emir & Yangın-Ekşi (2023), Ihrmark (2023), and Sun & Hu (2023), the emphasis on corpus linguistics, language development, and learner-centered approaches highlights the importance of grounding technological innovation in a solid understanding of language and learning. For Indonesian language educators, this means embracing CDLL not merely as a technological add-on but as a pedagogical framework that integrates authentic language data, research-informed practices, and learner-centered approaches to foster effective language acquisition (cf. Nugraha, 2024d, 2025). However, the effective application of these foundational principles within the Indonesian context requires the development and accessibility of representative corpora of Indonesian Language, encompassing its various registers and regional dialects. This might involve incorporating corpus-based activities into lesson plans, such as analyzing authentic texts to identify common collocations or grammatical structures. Furthermore, sustained professional development initiatives are essential to equip Indonesian language educators with the necessary skills and knowledge to effectively utilize these corpora and integrate CDLL principles into their diverse teaching contexts. Educators can utilize learner corpora to diagnose common errors and tailor instruction to address specific learning needs. By actively engaging with CDLL principles and incorporating them into their teaching practices, Indonesian language educators can create dynamic and engaging learning environments that promote language acquisition and cultural understanding.

Conclusion

This scientometric analysis, in accordance with the first research question, has provided a detailed overview of the current state and trajectory of CDLL research, revealing key trends and thematic clusters that illuminate the field's evolution and potential. The exponential growth in publications, the diverse range of contributing actors, and the interdisciplinary nature of the research underscore the growing significance of CDLL in language education. The findings highlight the increasing prominence of computational

approaches, the potential for CDLL applications beyond traditional language teaching contexts, and the importance of grounding technological innovation in a solid understanding of language acquisition and pedagogical principles. Moreover, in the light of second research question, this study contributes valuable insights to the ongoing discourse on CDLL and its implications for language pedagogy, particularly within the Indonesian context, where the integration of CDLL holds significant promise for enhancing language learning and promoting educational equity.

Despite its comprehensive scope, this study acknowledges certain limitations. The analysis primarily relied on bibliometric data from the Scopus database, which, while extensive, may not capture the full breadth of CDLL research published in other venues or in language other than English. Furthermore, the focus on keywords analysis and thematic clustering, while providing valuable insights into research trends, may not fully capture the nuances and complexities of individual studies. Future research could expand the scope of analysis to include other databases, grey literature, and qualitative research methods to provide a more holistic understanding of the CDLL landscape. Moreover, future research could delve deeper into specific thematic areas within CDLL, exploring the pedagogical applications of emerging technologies such as deep learning and natural language processing in greater detail. Investigating the effectiveness of CDLL interventions in diverse learning contexts, including Indonesian language classrooms, would provide valuable empirical evidence to guide pedagogical innovation. Furthermore, exploring the ethical implications of integrating AI and machine learning in language education, particularly concerning data privacy and algorithmic bias, is crucial to ensure responsible and equitable implementation of CDLL. By addressing these research gaps, the field can further advance its understanding of CDLL and its potential to transform language pedagogy in the digital age.

Declarations

Author contribution	:	The present study was conceived and executed in its entirety by Danang Satria Nugraha, who assumed sole responsibility for all aspects of the investigation.
Funding statement	:	The research presented herein was conducted independently and did not receive any financial support or grants from external funding agencies or institutions.
Conflict of interest	:	The author of this study declares no conflicts of interest, financial or otherwise, that could potentially influence the interpretation or presentation of the research findings.
Ethics Approval	:	In adherence to the ethical guidelines established by BAHASTRA in 2023, this research, which does not involve human subjects or sensitive data, does not require formal ethics approval or informed consent procedures.
Additional information	:	No further information pertaining to the research design, methodology, or findings is available beyond what is presented in this manuscript.

References

Alruwaili, A. K. (2024). Exploring language teachers' perceptions of corpus literacy skills at pre-tertiary level. *International Journal of Computer-Assisted Language Learning and Teaching*, *14*(1). https://doi.org/10.4018/IJCALLT.352064

Bal-Gezegin, B., Akbaş, E., & Başal, A. (2022). Corpus made my job easier: Preservice language teachers' corrective feedback practices in writing with corpus consultation. In *English Language Education, 30*. https://doi.org/10.1007/978-3-031-13540-8_14

Balouchzahi, F., Sidorov, G., & Gelbukh, A. (2023). PolyHope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, *225*. https://doi.org/10.1016/j.eswa.2023.120078

Charles, M. (2022). The gap between intentions and reality: Reasons for EAP writers' non-use of corpora. *Applied Corpus Linguistics*, *2*(3). https://doi.org/10.1016/j.acorp.2022.100032

Chen, M., & Flowerdew, J. (2018). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, *23*(3), 335–369. https://doi.org/10.1075/ijcl.16130.che

Chen, Z., & Jiao, J. (2019). Effect of the blended learning approach on teaching corpus use for collocation richness and accuracy. In *Communications in Computer and Information Science, 1048*. https://doi.org/10.1007/978-981-13-9895-7_6

Crosthwaite, P. (2017). Retesting the limits of data-driven learning: feedback and error correction. *Computer Assisted Language Learning*, *30*(6), 447–473. https://doi.org/10.1080/09588221.2017.1312462

_____. (2019). Data-driven learning and younger learners: Introduction to the volume. In *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners*. https://doi.org/10.4324/9780429425899-1

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, *3*(3). https://doi.org/10.1016/j.acorp.2023.100066

Crosthwaite, P., Luciana, & Schweinberger, M. (2021). Voices from the periphery: Perceptions of Indonesian primary vs secondary pre-service teacher trainees about corpora and data-driven learning in the L2 English classroom. *Applied Corpus Linguistics*, *1*(1). https://doi.org/10.1016/j.acorp.2021.100003

Crosthwaite, P., & Stell, A. (2019). It helps me get ideas on how to use my words: Primary school students' initial reactions to corpus use in a private tutoring setting. In *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners*. https://doi.org/10.4324/9780429425899-9

Emir, G., & Yangın-Ekşi, G. (2023). Corpus used as a data-driven learning tool in L2 academic writing: Evidence from Turkish contexts. *Teflin Journal*, *34*(2), 209–225. https://doi.org/10.15639/teflinjournal.v34i2/209-225

Esfahani, M. J. B., & Ketabi, S. (2024). The effect of corpus-assisted language teaching on academic collocation acquisition by Iranian advanced EFL learners. *Journal of Applied Research in Higher Education*, *16*(4), 1188–1213. https://doi.org/10.1108/JARHE-05-2023-0199

Flowerdew, L. (2022). Using corpora for writing instruction (Second edition). In *The Routledge Handbook of Corpus Linguistics*. https://doi.org/10.4324/9780367076399-31

Flowerdew, L., & Petrić, B. (2024). A critical review of corpus-based pedagogic perspectives on thesis writing: Specificity revisited. *English for Specific Purposes*, *76*, 1–13. https://doi.org/10.1016/j.esp.2024.05.003

Gardner, S. (2024). Corpus approaches to discourse and second language research. In *The Routledge Handbook of Second Language Acquisition and Discourse*. https://doi.org/10.4324/9781003177579-14

Henneken, E. A., & Kurtz, M. J. (2019). *Usage bibliometrics as a tool to measure research activity*. https://doi.org/10.1007/978-3-030-02511-3_32

Hu, B., Tang, B., Chen, Q., & Kang, L. (2016). A novel word embedding learning model using the dissociation between nouns and verbs. *Neurocomputing*, *171*, 1108–1117. https://doi.org/10.1016/j.neucom.2015.07.046

Ihrmark, D. (2023). Revisiting the computer as informant from a teacher-mediated perspective: Suggested implementation of an automated language diagnostics tool. *NJES Nordic Journal of English Studies*, *22*(1), 42–67. https://doi.org/10.35360/njes.794

Kızıl, A. S. (2023). Data-driven learning: English as a foreign language writing and complexity, accuracy and fluency measures. *Journal of Computer Assisted Learning*, *39*(4), 1382–1395. https://doi.org/10.1111/jcal.12807

Lin, M. H. (2021). Effects of data-driven learning on college students of different grammar proficiencies: A preliminary empirical assessment in EFL classes. *SAGE Open*, *11*(3). https://doi.org/10.1177/21582440211029936

Liu, S., Tang, B., Chen, Q., & Wang, X. (2015). Effects of semantic features on machine learning-based drug name recognition systems: Word embeddings vs. Manually constructed dictionaries. *Information (Switzerland)*, *6*(4), 848–865. https://doi.org/10.3390/info6040848

Liu, T., & Chen, M. (2023). An investigation into learners' cognitive processes in data-driven learning: Case studies of six learners of Chinese. *Chinese Journal of Applied Linguistics*, *46*(4), 544–561. https://doi.org/10.1515/CJAL-2023-0404

Lusta, A., Demirel, Ö., & Mohammadzadeh, B. (2023). Language corpus and data driven learning (DDL) in language classrooms: A systematic review. *Heliyon*, *9*(12). https://doi.org/10.1016/j.heliyon.2023.e22731

Lyu, Y., & Han, Z. (2023). Applying data-driven learning in self-translation of academic discourse: A case study of a Chinese medical student. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1071123

Ma, Q., Tang, J., & Lin, S. (2022). The development of corpus-based language pedagogy for TESOL teachers: A two-step training approach facilitated by online collaboration. *Computer Assisted Language Learning*, *35*(9), 2731–2760. https://doi.org/10.1080/09588221.2021.1895225

Mamta, Ekbal, A., & Bhattacharyya, P. (2022). Exploring multi-lingual, multi-task, and adversarial learning for low-resource sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(5). https://doi.org/10.1145/3514498

Muftah, M. (2023). Data-driven learning (DDL) activities: Do they truly promote EFL students' writing skills development? *Education and Information Technologies*, *28*(10), 13179–13205. https://doi.org/10.1007/s10639-023-11620-z

Nugraha, D. S. (2021). Morphosemantic features of derivational affix {Me(N)-} in the Indonesian denumeral verb constructions. *Sirok Bastra*, *9*(2). https://doi.org/10.37671/sb.v9i2.317

_____. (2024a). Analyzing prefix /me(N)-/ in the Indonesian affixation: A corpus-based morphology. *Theory and Practice in Language Studies*, *14*(6), 1697–1711. https://doi.org/10.17507/tpls.1406.10

_____. (2024b). A morphological analysis of the Indonesian suffixation: A look at the different types of affixes and their semantic changes. *GEMA Online® Journal of Language Studies*, *24*(4), 109–132. https://doi.org/10.17576/gema-2024-2404-07

_____. (2024c). Navigating challenges and opportunities: Incorporating multimodal analysis into corpus linguistics for social media research. In *Corpora for Language Learning: Bridging the Research-Practice Divide*.

_____. (2024d). Quantitative analysis within language studies: An analytical views based on the bibliometrics method. *Script Journal: Journal of Linguistics and English Teaching*, *9*(2), 16–34. https://doi.org/10.24903/sj.v9i2.1695

_____. (2025). Complex word formation in contemporary syntactic frameworks: Scientometric investigation and its relevance to grammar pedagogy. *JOALL (Journal of Applied Linguistics and Literature)*, *10*(1), 283–315. https://doi.org/10.33369/joall.v10i1.40034

Nugraha, D. S., Widharyanto, W., Setyaningsih, Y., & Rahardi, R. K. (2025). *Linguistik edukasional: Telaah masalah pendidikan bahasa*. Sanata Dharma University Press.

Pawlak, M., & Kruk, M. (2022). Individual differences in computer assisted language learning research. In *Individual differences in Computer Assisted Language Learning Research*. https://doi.org/10.4324/9781003240051

Pérez-Paredes, P. (2022). How learners use corpora. In *The Routledge Handbook of Corpora and English Language Teaching and Learning*. https://doi.org/10.4324/9781003002901-31

Saeed, A., Nawab, R. M. A., & Stevenson, M. (2022). Investigating the feasibility of deep learning methods for urdu word sense disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(2). https://doi.org/10.1145/3477578

Sooryamoorthy, R. (2020). *Scientometrics for the humanities and social sciences*. Routledge. https://doi.org/10.4324/9781003110415

Sun, X., & Hu, G. (2023). Direct and indirect data-driven learning: An experimental study of hedging in an EFL writing class. *Language Teaching Research*, *27*(3), 660–688. https://doi.org/10.1177/1362168820954459

van Eck, N. J., & Waltman, L. (2023). *VOSviewer* (1.6.20). Universiteit Leiden.

Waltman, L., & van Eck, N. J. (2019). *Field normalization of scientometric indicators*. https://doi.org/10.1007/978-3-030-02511-3_11

Wicher, O. (2019). Data-driven learning in the secondary classroom: A critical evaluation from the perspective of foreign language didactics. In *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners*. https://doi.org/10.4324/9780429425899-3

Yao, G. (2019). Vocabulary learning through data-driven learning in the context of Spanish as a foreign language. *Research in Corpus Linguistics*, *7*, 18–46. https://doi.org/10.32714/ricl.07.02

Yu, X., & Altunel, V. (2023). Data-driven learning for foreign and second language education in diverse contexts. In *New Approaches to the Investigation of Language Teaching and Literature*. https://doi.org/10.4018/978-1-6684-6020-7.ch005

Zare, J., & Delavar, K. A. (2024). Enhancing English learning materials with data-driven learning: A mixed-methods study of task motivation. *Journal of Multilingual and Multicultural Development*, *45*(9), 4011–4027. https://doi.org/10.1080/01434632.2022.2134881