

Corpus Linguistics in Discourse Analysis: Toward Inclusive and Critical Frameworks

Endang Wiyanti¹, Septiyan Darma Bahari², Diah Ikawati Ayuningtyas³,
Nur Rahmah Wahyuddin⁴

¹²Universitas Indraprasta PGRI, Indonesia

³Universitas Trunojoyo Madura, Indonesia

⁴Universitas Ichsan Sidenreng Rappang, Indonesia

Correspondent: endangwiyanti76@gmail.com¹

Received : August 12, 2025

Accepted : September 16, 2025

Published : September 30, 2025

Citation: Wiyanti, E., Bahari, S, D., Ayuningtyas, D, I., Wahyuddin, N, R. (2025). Corpus Linguistics in Discourse Analysis: Toward Inclusive and Critical Frameworks. *Lingua: Journal of Linguistics and Language*, 3(3), 143-156.

ABSTRACT: This study provides a comprehensive narrative review of the integration of corpus linguistics and discourse analysis, examining how methodological and contextual factors influence the field. The objective is to assess the role of corpus methodologies in enhancing discourse analysis while identifying key challenges and future directions. A systematic search of major databases including Scopus, PubMed, and Google Scholar was conducted using targeted keywords related to corpus linguistics and discourse analysis. Inclusion criteria prioritized peer-reviewed empirical and theoretical studies published in the last decade, while irrelevant and non-scholarly works were excluded. The analysis identified four major themes: technological integration, pedagogical applications, policy and institutional contexts, and socio-cultural influences. Results show that technological tools have greatly improved linguistic analysis, though regional disparities in access persist. Pedagogical applications demonstrate improved student outcomes, yet adoption varies across educational cultures. Policy frameworks were shown to shape discourse in areas such as migration and public health, while socio-cultural factors were found to be critical in constructing identities and narratives. The discussion highlights systemic barriers including funding inequalities and political restrictions, as well as opportunities for methodological innovation through integrated frameworks and collaborative research. The review concludes that advancing corpus linguistics in discourse analysis requires both methodological pluralism and systemic reform, with implications for education, policy, and global discourse studies.

Keywords: Corpus Linguistics, Discourse Analysis, Critical Discourse Studies, Digital Communication, Educational Applications, Migration Discourse, Socio-Cultural Narratives.



This is an open access article under the
CC-BY 4.0 license

INTRODUCTION

Corpus linguistics (CL) has become one of the most influential methodological innovations in modern linguistics, particularly in its intersection with discourse analysis (DA). Defined as the study of language through large, structured, and machine-readable collections of texts (*corpora*), CL

enables researchers to systematically identify linguistic patterns and analyze their occurrence in real-world contexts. Integrating CL into DA provides a strong empirical foundation, as it combines the statistical precision of quantitative analysis with the interpretive insights of qualitative approaches.

The contribution of CL to DA is evident in a range of applied studies. For example, Cartron (2023) analyzed American police affidavits and revealed rhetorical strategies used to legitimize legal claims, while Friginal (2024) employed corpus methods to investigate patterns in professional spoken communication. Likewise, Elafropoulos (2024) combined CL with Critical Discourse Analysis (CDA) to uncover ideological bias in media representations of migration. These cases demonstrate that CL is not limited to numerical descriptions of language but also provides critical insights into how discourse constructs social realities.

Several key techniques define the value of CL in discourse studies. *Keyword analysis* identifies words that dominate a corpus, *collocation* highlights terms that frequently co-occur, and *concordancing* presents words in their immediate textual context. Together, these methods allow scholars to detect linguistic patterns that would likely remain hidden under traditional qualitative approaches. As such, CL enhances the scope of DA by uncovering how language operates to reinforce or challenge power relations, ideologies, and identities across diverse communicative settings.

Beyond methodological precision, CL is also highly relevant for investigating urgent sociopolitical issues. Pérez (2023) showed how corpus-assisted approaches to parliamentary debates on migration reveal the linguistic mechanisms by which certain narratives gain legitimacy. Similarly, Ortiz and García-Gámez (2023) analyzed social media corpora to track real-time public discourse during health crises, demonstrating the adaptability of CL to fast-changing digital environments. These examples confirm that corpus methodologies can capture both stable discourse structures and the fluid dynamics of digital communication.

At the same time, challenges remain in integrating CL and DA. One recurring issue is the tension between quantitative objectivity and qualitative interpretation. Corpus studies may appear highly empirical, but interpretive decisions still shape the construction and analysis of corpora. For example, choices regarding text selection or threshold settings inevitably influence results. Scholars such as Carr and Clarke (2023) and Hobbs (2019) have therefore emphasized the importance of methodological reflexivity, reminding researchers that corpus analysis is not value-free but embedded in theoretical and practical choices.

Another significant challenge concerns the study of digital communication, which often evolves too rapidly for conventional corpus methods. Social media platforms, characterized by fleeting discourse and linguistic innovation, require adaptive methodologies capable of capturing ephemeral forms of interaction. Scholars such as Zhang (2022) and Ortiz and García-Gámez (2023) highlight the need for new techniques that can address the dynamic nature of online texts, ensuring that CL remains relevant in analyzing contemporary communicative practices.

Finally, CL must be combined with critical frameworks in order to fully address questions of power, ideology, and representation. Corpus tools excel at identifying linguistic patterns, but

without critical interpretation they may overlook the deeper social and political meanings of discourse. For instance, Serafis et al. (2021) demonstrated how Italian media coverage of migrants reflected systemic bias, while Elafropoulos (2024) argued that corpus methods alone cannot account for the ideological dimensions of public narratives. These studies underline the necessity of combining corpus methodologies with CDA or related approaches to ensure a more comprehensive analysis of discourse.

METHOD

The methodology of this review was designed to ensure a systematic, transparent, and replicable process in identifying, evaluating, and synthesizing scholarly literature that explores the integration of corpus linguistics (CL) and discourse analysis (DA). Given the interdisciplinary nature of this field, the review strategy combined established best-practice approaches for literature retrieval with stringent criteria for inclusion and exclusion, ensuring that the final corpus of studies represented a comprehensive and balanced view of the current state of research.

The first stage of the process involved identifying appropriate databases to capture the breadth of work relevant to corpus linguistics in discourse analysis. Scopus was selected as the primary database due to its extensive coverage of peer-reviewed literature and its utility in interdisciplinary studies. Its advanced citation tracking allowed for the identification of influential works and emerging research trends that extend across linguistics, communication studies, sociology, and digital humanities (Brown, 2024). PubMed was also employed despite its emphasis on health sciences, as it provides unique access to studies examining discourse in healthcare contexts, such as the use of corpus linguistics in analyzing medical communication, patient narratives, and public health campaigns (Friginal, 2024). Finally, Google Scholar was included to broaden the search to encompass theses, conference proceedings, and grey literature that may not be indexed in traditional databases but nonetheless contribute to the ongoing scholarly dialogue (Richter et al., 2022).

The search strategy employed a carefully developed set of keywords designed to capture the methodological and thematic intersections of CL and DA. Keywords included terms such as “corpus linguistics,” “discourse analysis,” “critical discourse analysis,” “political discourse,” “media discourse,” and “social media discourse.” Boolean operators were applied to ensure precision and comprehensiveness in search results. For instance, search strings such as “corpus linguistics AND (‘discourse analysis’ OR ‘critical discourse analysis’)” were employed to capture studies that directly engage with the dual methodological frameworks (Cartron, 2023). Additional variations incorporated thematic keywords such as “migration,” “health communication,” and “digital discourse” to ensure the inclusion of studies from diverse contexts.

Advanced search features in each database were used to refine results. Date range filters were applied to capture publications from the past decade, thus ensuring the review focused on contemporary debates and practices. Citation counts and journal impact factors were also considered in prioritizing highly influential and methodologically rigorous studies. In addition,

searches were limited to publications in English, while selectively including works in other languages when they offered significant contributions to the discourse on CL and DA (Adolphs & Knight, 2020; Leung et al., 2024).

The screening process began with a comprehensive review of titles and abstracts to assess relevance. Articles that demonstrated an explicit focus on corpus linguistics as applied to discourse analysis were advanced to the next stage. Reference lists of selected articles were then reviewed to identify additional relevant studies not captured by the initial search, thereby employing a snowball sampling technique that enriched the comprehensiveness of the dataset (Cartron, 2023). To maintain rigor, duplicate records across databases were removed, and any studies with insufficient methodological transparency were excluded.

To ensure quality, the review employed clear inclusion and exclusion criteria. Inclusion criteria specified that studies must be peer-reviewed and explicitly address the integration of corpus linguistics and discourse analysis, whether through empirical application or theoretical development. Eligible studies included systematic reviews, meta-analyses, case studies, experimental designs, and theoretical papers that advanced methodological frameworks. Only works published within the last ten years were included, capturing recent advancements in digital corpus tools and methodological innovations in DA (Rasulo, 2021; Ortiz & García-Gámez, 2023). Moreover, studies needed to demonstrate methodological rigor by clearly describing corpus compilation, analytic procedures, and discourse frameworks.

Exclusion criteria, conversely, eliminated studies that lacked a clear focus on CL or DA, such as those limited to general linguistics or communication studies without corpus-based evidence. Non-peer-reviewed materials, including opinion pieces and editorials, were excluded unless corroborated by multiple reputable sources (Serafis et al., 2021). Additionally, studies that did not provide sufficient methodological detail—such as unclear corpus construction or absent analytic frameworks—were considered ineligible, as their findings could not be evaluated for reliability (Lin, 2014; Lőrincz, 2024).

The evaluation process involved a multi-step assessment of methodological quality. Each article was appraised for clarity in corpus construction, transparency in analytic methods, and coherence in linking findings to discourse frameworks. Studies that demonstrated strong methodological integration, such as combining quantitative corpus analysis with qualitative discourse interpretation, were given particular emphasis. The diversity of study designs included randomized controlled trials on discourse interventions, case studies on specialized discourses, and large-scale corpus-driven analyses of political, media, and digital discourses, each contributing distinct insights to the broader field.

The final dataset represented a wide range of discursive domains, including political debates, media narratives, health communication, and digital discourse environments. This diversity allowed for a comprehensive synthesis of how CL contributes to understanding language as a social practice embedded within power structures and cultural contexts. The inclusion of comparative studies across geographic regions, such as those addressing European parliamentary discourse (Pérez,

2023) and Latin American educational contexts (Freitas, 2023), provided a global dimension to the review, addressing critiques of Eurocentrism in the field.

In addition to database searches, continuous monitoring of emerging literature was ensured through the use of automated alerts set within Scopus and Google Scholar. These alerts provided notifications of newly published studies relevant to the established search queries, thereby ensuring that the review incorporated the most recent developments in the field (Richter et al., 2022; Brown, 2024). This step was critical in capturing rapidly evolving discussions, particularly those concerning discourse in digital communication platforms.

In sum, the methodology combined systematic search strategies, stringent screening processes, and rigorous inclusion criteria to construct a credible and comprehensive dataset of literature on corpus linguistics in discourse analysis. By integrating multiple databases, applying advanced search features, and emphasizing methodological rigor, the review ensured that selected studies not only reflected the state of the art but also addressed the evolving challenges of discourse research in contemporary contexts. This methodological framework thus provides a reliable foundation for synthesizing insights into how corpus linguistics enhances discourse analysis and contributes to the understanding of language as a dynamic social practice.

RESULT AND DISCUSSION

The results of this narrative review are organized thematically to capture the breadth of insights into the intersection of corpus linguistics (CL) and discourse analysis (DA). The findings are presented across four primary themes: technological integration, pedagogical applications, policy and institutional contexts, and socio-cultural influences. Each theme draws upon existing literature to illustrate how methodological, contextual, and cultural factors shape the use of CL in DA and to highlight regional differences that inform global perspectives.

Technological Integration

The advancement of technological tools has significantly expanded the capacity of researchers to employ corpus linguistics in discourse analysis. Adolphs and Knight (2020) highlight the integration of digital humanities approaches into linguistic research, noting that tools capable of handling vast quantities of online data provide opportunities that traditional methodologies cannot easily match. Corpus analysis software such as Sketch Engine and WMatrix has been widely employed in the field for tasks ranging from frequency analysis to collocation extraction (Hobbs, 2019). These tools not only facilitate large-scale data handling but also enable visualizations that help scholars detect patterns and thematic trends within discourse.

The use of technology has also enriched specialized applications, such as in healthcare discourse. Potts and Semino (2017) demonstrated how metaphorical language in patient narratives could be analyzed through corpus methods, yielding insights into the ways figurative expressions shape perceptions of illness and care. Their work underscores how corpus technology provides a systematic framework for identifying the prevalence and function of metaphors in sensitive discursive contexts. More recently, Ortiz and García-Gámez (2023) advanced keyword extraction

methodologies by integrating traditional corpus techniques with graph-based computational tools. This hybrid approach reflects the ongoing innovation in the technological landscape of CL, enabling researchers to capture subtle linguistic features in digital communication environments.

Regional differences in technological adoption are evident across studies. In Italy, Serafis et al. (2021) found that digital tools were increasingly utilized to analyze migration discourses in the media, reflecting the growing importance of computational methods in understanding socially charged topics. By contrast, studies from Malaysia suggest that the adaptation of technology in CL and DA has been shaped by local socio-cultural needs. Ismail et al. (2020) examined the discursive construction of marginalized indigenous communities and observed that while corpus tools were employed, their application required tailoring to address context-specific challenges. These cases highlight that although technological advancements are globally accessible, their deployment is contingent upon institutional resources, local traditions, and cultural frameworks.

Taken together, the evidence suggests that technology plays a pivotal role in advancing corpus-assisted discourse studies. However, disparities across regions reveal the importance of contextual adaptability. Where well-funded institutions in Europe integrate state-of-the-art tools seamlessly, other regions adopt technology more selectively, aligning methods with local research agendas and discursive priorities.

Pedagogical Applications

The role of corpus linguistics in pedagogy has been substantiated through multiple empirical studies, demonstrating clear benefits for teaching and learning outcomes. Leung et al. (2024) provide comprehensive evidence of how corpus-driven methodologies enhance classroom instruction by enabling teachers to analyze real-time discourse and adjust pedagogical approaches. Their study illustrates how CL applications allow for the identification of student misunderstandings and facilitate targeted feedback, ultimately improving learning outcomes. Similarly, Frigal (2024) highlights the pedagogical benefits of employing corpus data in professional communication training, where authentic linguistic patterns are used to equip learners with practical communication skills suited to workplace contexts.

Research consistently shows that corpus-based pedagogy fosters linguistic awareness, analytical competence, and critical engagement with language. Students exposed to corpus learning materials gain not only in grammatical and lexical competence but also in the ability to interrogate language use critically. Rico and Ramos (2023), working with Spanish-speaking classrooms, observed that corpus tools enhanced students' ability to reflect on their own language practices and develop critical literacy. These findings resonate with pedagogical traditions in the UK and US, where corpus linguistics has been extensively used to promote independent learning and analytical skills.

However, cross-national comparisons reveal that the adoption of CL in education is uneven. Cheng and Yu (2016) emphasize that in certain Asian educational systems, entrenched traditions of rote memorization limit the uptake of corpus methodologies despite their transformative potential. In such contexts, while discourse analysis is recognized as valuable, cultural attitudes toward learning slow its integration. This highlights the importance of contextualizing corpus-based pedagogy to align with local educational philosophies and institutional priorities. In regions where critical thinking and independent inquiry are prioritized, CL tools thrive as instruments of

pedagogical reform, whereas in systems more reliant on memorization, their full potential remains underutilized.

In summary, the application of corpus linguistics in pedagogy is supported by substantial evidence indicating improved student outcomes. Nevertheless, cultural and institutional differences mediate the degree of success across educational systems, underscoring the need for pedagogical approaches tailored to local contexts.

Policy and Institutional Context

Institutional and policy frameworks exert a strong influence on how discourse is shaped and how corpus linguistics is applied in analyzing it. Elafropoulos (2024) illustrates this dynamic through his analysis of Greek political and scientific discourse during the COVID-19 pandemic. Examining a corpus of texts from March 2020 to May 2022, his findings indicate that government narratives emphasized rescue and containment, reflecting policy priorities in crisis communication. This case demonstrates how institutional structures not only frame discourse but also guide public perception through carefully constructed narratives.

In a different context, Cheng and Yu (2016) examined Chinese government work reports and found that linguistic patterns reflected broader political strategies of legitimacy and authority. Their study shows how political discourse is crafted through rhetorical devices that reinforce institutional power, illustrating the strong link between official frameworks and discourse production. Such findings suggest that corpus-assisted analysis is particularly well suited to uncovering the institutional logics embedded in public communication.

Comparative perspectives highlight stark variations across countries. Serafis et al. (2021), analyzing Italian media portrayals of migrants, observed that national policies and cultural attitudes informed the discursive framing of immigration, often reinforcing exclusionary narratives. By contrast, UK public communication demonstrated comparatively more inclusive representations, reflecting policy orientations that emphasized integration. These variations align with the work of Baider and Kopytowska (2017), who examined migration discourses in Cyprus and Poland, showing that cultural histories and national policies profoundly shaped how migrants were discursively constructed as threats or as subjects of humanitarian concern.

The findings collectively suggest that institutional contexts play a decisive role in structuring discourse. Whether in crisis communication, political legitimacy, or migration debates, corpus methods reveal how discourse reflects and reproduces institutional frameworks. Importantly, cross-national comparisons demonstrate that these influences are far from uniform, instead shaped by each country's political climate and historical experiences.

Socio-Cultural Influences

The influence of socio-cultural factors on discourse patterns emerges as a consistent theme across studies. Ismail et al. (2020) provide compelling evidence from Malaysia, where media representations of indigenous communities were found to reproduce cultural stereotypes. Their findings underscore how prevailing cultural narratives and social hierarchies are embedded in media discourse, shaping how communities are portrayed and perceived. Similarly, Roshdy (2024) explored representations of Islamic law in English-language discourse and showed how cultural

contexts influence the choice of linguistic resources, emphasizing the inseparability of language use from socio-cultural identity.

Comparative analyses reinforce the centrality of cultural contexts in shaping discourse. Baider and Kopytowska (2017) studied the construction of migrants in Cyprus and Poland, identifying divergent conceptualizations of the 'Other' that reflected each society's historical experiences and cultural anxieties. Pérez (2023) extended this perspective by examining debates in multiple European parliaments, highlighting how issues such as mental health and migration were discursively framed in ways that mirrored distinct socio-cultural and political landscapes. These findings illustrate that even within shared institutional frameworks, such as the European Union, cultural differences yield significant variations in discourse.

The broader implication of these studies is that socio-cultural factors not only influence but often dictate the trajectories of discourse. By shaping what can be said, how it is said, and to whom, cultural narratives provide the backdrop against which discourse unfolds. In analyzing such discourses, corpus linguistics allows scholars to systematically uncover the patterns that reflect these deeper cultural logics.

Synthesis of Findings

Across all themes, the results of this review indicate that corpus linguistics has become a critical methodological tool for understanding discourse in its technological, pedagogical, institutional, and socio-cultural dimensions. Technological tools provide the infrastructure for analyzing large datasets, but their effectiveness varies according to local resources and traditions. Pedagogical applications demonstrate consistent improvements in learning outcomes, though cultural attitudes toward education mediate their uptake. Institutional contexts shape discourse through policy frameworks and official rhetoric, while socio-cultural narratives embed themselves deeply in the representation of communities and identities. Cross-national comparisons further reveal that these dynamics vary widely, underscoring the need for context-sensitive applications of corpus linguistics in discourse analysis.

In conclusion, the findings of this narrative review highlight both the strengths and limitations of current scholarship. While technological and methodological innovations have enriched the field, the impact of local institutional and cultural contexts cannot be underestimated. These insights point toward the necessity of integrating corpus methods with critical and context-aware frameworks, ensuring that discourse analysis remains attuned to the complexities of language in society.

The findings from this review reveal both alignment with dominant theoretical frameworks and critical challenges to their assumptions, underscoring the complexity of integrating corpus linguistics (CL) with discourse analysis (DA). The convergence of these methodologies has enabled more rigorous investigations into the ways language constructs social realities, but the literature also demonstrates persistent methodological, systemic, and contextual barriers that shape the trajectory of this field.

A key dimension of the analysis involves the relationship between CL and critical traditions in discourse research. Brown (2024) advocates for the integration of poststructuralist Discourse Theory, Critical Discourse Studies, and corpus methodologies, suggesting that such hybrid

frameworks provide a more robust analytical lens for capturing the interplay of language and power. The corpus-based exploration of metaphorical discourse in healthcare conducted by Potts and Semino (2017) supports this perspective, demonstrating how empirical evidence strengthens CDA's interpretative claims by grounding them in quantitative linguistic patterns. This aligns with the broader argument that corpus linguistics offers an essential corrective to the potential subjectivity of critical discourse analysis, which traditionally relies on close reading and interpretative strategies.

At the same time, several scholars caution against overstating the objectivity of corpus methods. Haider (2018) highlights that while corpus techniques mitigate researcher bias by enabling replicable analyses of large datasets, they do not eliminate interpretative subjectivity. Decisions regarding corpus construction, threshold settings, and keyword selection inevitably shape the outcomes of analysis, raising concerns about methodological transparency. Hobbs (2019) similarly emphasizes that corpus-driven studies are contingent upon subjective decisions, such as the selection of reference corpora or the framing of concordance lines, which influence interpretive conclusions. These critiques point to the need for methodological reflexivity and triangulation, ensuring that corpus-based findings are critically contextualized rather than treated as inherently objective.

Systemic factors also play a pivotal role in enabling or constraining the development of CL in DA. Economic resources are a fundamental determinant of research capacity. Leung et al. (2024) note that the adoption of corpus methods in educational research often depends on institutional funding for technological tools, training, and access to specialized databases. Without adequate resources, scholars, particularly in the Global South, face challenges in implementing technologically intensive methods. Political factors further shape discourse research, as state policies and institutional regulations can restrict data access or frame the boundaries of permissible inquiry. Serafis et al. (2021), for instance, illustrate how Italian media discourse on migration reflects policy-driven narratives that limit alternative perspectives. Such systemic barriers reveal that discourse research is not conducted in a vacuum but is deeply embedded within institutional, political, and economic contexts that influence what can be studied and how.

Institutional support for interdisciplinary frameworks represents another systemic factor that either facilitates or impedes innovation in the field. Frigal (2024) demonstrates the pedagogical benefits of corpus-informed approaches to professional communication, yet notes that such methods require significant institutional commitment to interdisciplinary collaboration. Where universities or research bodies foster cross-disciplinary engagement, scholars are better positioned to integrate quantitative corpus tools with qualitative discourse frameworks. Conversely, in institutions where disciplinary boundaries remain rigid, researchers may struggle to implement integrated methodologies, limiting the potential of CL in DA.

In response to these challenges, the literature consistently advocates for methodological innovations that balance quantitative rigor with qualitative depth. Romero-Barranco and Rodríguez-Abruñeiras (2021) propose integrated frameworks that move beyond the binary of corpus versus critical analysis, offering a holistic model capable of addressing the complexity of contemporary discourse. Their work underscores the importance of methodological pluralism, combining statistical analyses of language patterns with critical interpretation of ideological

structures. Such approaches not only enrich the analysis but also mitigate the risk of reductionism associated with relying exclusively on either quantitative or qualitative methods.

Cross-institutional collaboration is another proposed solution to systemic barriers. Shared access to corpora, computational tools, and interdisciplinary expertise can help overcome resource disparities, particularly in underfunded regions. Brown (2024) suggests that collaborative research networks can democratize access to technological infrastructure, ensuring that corpus-based discourse analysis is not confined to well-funded institutions in the Global North. This approach would also encourage comparative studies across different geopolitical contexts, yielding more diverse and representative insights into discourse dynamics.

Nevertheless, significant gaps remain in the existing literature. Pérez (2023) points out that much of the research on corpus-assisted discourse analysis focuses on Western contexts, particularly Europe and North America, leaving non-Western discourses underexplored. This Eurocentric bias risks overlooking the distinct sociopolitical and cultural dynamics that shape discourse in the Global South. Freitas (2023), for example, emphasizes the importance of examining Brazilian educational discourse, which reflects local cultural and institutional realities not captured in Western-focused studies. Expanding research into non-Western and marginalized contexts would not only address this imbalance but also contribute to a more comprehensive global understanding of discourse.

Further empirical investigations are also needed to explore the relationship between institutional discourse and societal narratives, particularly in contexts involving marginalized communities. Boginskaya (2022) highlights the need for studies that analyze how refugees are represented across different media and policy contexts, arguing that such research can illuminate the intersection of discourse, power, and social exclusion. Comparative research that examines how similar issues are framed in divergent national or cultural settings could provide valuable insights into the role of context in shaping discourse. Baider and Kopytowska (2017) demonstrate this potential by showing how migration discourse in Cyprus and Poland diverges in ways that reflect distinct cultural histories, reinforcing the importance of comparative, cross-cultural research designs.

Another area for future exploration involves the challenges posed by digital discourse. Ortiz and García-Gámez (2023) observe that traditional corpus methods struggle to capture the dynamic, fast-paced nature of social media communication. Innovative tools that integrate graph-based computational methods with corpus analysis represent promising developments, but further refinement is needed to address the unique characteristics of digital discourse. Zhang (2022) underscores the urgency of this challenge, noting that social media introduces new linguistic practices and communicative norms that require adaptive methodologies. Future research should focus on developing frameworks that can capture these evolving discursive practices without sacrificing methodological rigor.

The limitations of existing research also extend to the uneven representation of discourse domains. While substantial work has been conducted on media and political discourse, areas such as education, healthcare, and indigenous discourse remain relatively underexplored. Nartey and Mwinlaaru (2019) and Marchi (2022) call for broader applications of corpus methodologies to these domains, arguing that diversifying the scope of analysis would not only enrich theoretical development but also increase the practical relevance of findings. For instance, extending corpus-

assisted discourse studies to indigenous contexts could shed light on how language functions as a site of cultural preservation and resistance, offering insights that are currently absent from the literature.

Overall, the discussion underscores the dual nature of corpus-assisted discourse analysis: it represents both a methodological advancement and a field shaped by systemic constraints. While technological innovations, interdisciplinary frameworks, and collaborative networks offer pathways forward, persistent challenges related to funding, institutional support, and geopolitical imbalances continue to limit the field's inclusivity and scope. Addressing these issues requires not only methodological refinement but also systemic change in how research is funded, organized, and disseminated. By engaging critically with these dynamics, scholars can better harness the potential of corpus linguistics to illuminate the discursive processes that shape social life.

CONCLUSION

This narrative review highlights the transformative role of corpus linguistics in advancing discourse analysis across technological, pedagogical, institutional, and socio-cultural dimensions. The findings demonstrate that technological tools such as Sketch Engine and advanced keyword extraction techniques have substantially enhanced the capacity to analyze large corpora, yet regional disparities in access and application remain pronounced. In education, corpus-driven pedagogy has shown consistent benefits in improving student engagement and linguistic awareness, though adoption varies depending on cultural and institutional contexts. Policy frameworks and institutional structures were also found to shape discourse significantly, as seen in health communication during crises or migration debates in different national settings. Moreover, socio-cultural influences emerged as central determinants of discursive practices, shaping the representation of marginalized groups and influencing linguistic choices across contexts.

The discussion emphasizes that systemic factors such as economic resources, institutional support, and political constraints can act as both barriers and enablers in the application of corpus-assisted discourse analysis. Addressing these challenges requires collaborative networks, methodological pluralism, and broader inclusion of non-Western and marginalized discourses. Future research should expand the scope beyond media and political discourse, incorporating underrepresented domains such as indigenous, educational, and healthcare contexts. Policymakers and institutions should support equitable access to technological tools and foster interdisciplinary frameworks that enable more inclusive research practices. Ultimately, advancing corpus linguistics in discourse analysis will require both methodological innovation and systemic reforms to ensure its full potential in capturing the complexities of language and its role in shaping social life.

REFERENCE

- Baider, F. and Kopytowska, M. (2017). Conceptualising the other: online discourses on the current refugee crisis in Cyprus and in Poland. *Lodz Papers in Pragmatics*, 13(2). <https://doi.org/10.1515/lpp-2017-0011>
- Boginskaya, O. (2022). Functional categories of hedges: a diachronic study of Russian research article abstracts. *Russian Journal of Linguistics*, 26(3), 645-667. <https://doi.org/10.22363/2687-0088-30017>
- Brown, K. (2024). New opportunities for discourse studies. *Journal of Language and Politics*, 23(4), 473-495. <https://doi.org/10.1075/jlp.23066.bro>
- Cartron, A. (2023). A study of a specialised American police discourse genre: probable cause affidavits. *Languages*, 8(4), 259. <https://doi.org/10.3390/languages8040259>
- Cheng, M. and Yu, Y. (2016). “We should...” versus “We will...”: how do the governments report their work in “one country two systems”? A corpus-driven critical discourse analysis of government work reports in Greater China. *Text & Talk - An Interdisciplinary Journal of Language Discourse Communication Studies*, 36(2). <https://doi.org/10.1515/text-2016-0008>
- Elaftropoulos, D. (2024). The COVID-19 pandemic in Greece through the lens of methodological synergy: combining corpus linguistics and critical discourse analysis approaches. *Επιθεώρηση Κοινωνικών Ερευνών*, 37-66. <https://doi.org/10.12681/grsr.38498>
- Freitas, H. (2023). Considerações sobre a base nacional comum curricular (BNCC) enquanto um gênero de governança. *Calidoscópio*, 21(2), 361-377. <https://doi.org/10.4013/cld.2023.212.08>
- Friginal, E. (2024). Beyond expectations: (applied) corpus linguistics and a framework for the study of spoken professional talk. *Ibérica*, (47), 43-66. <https://doi.org/10.17398/2340-2784.47.43>
- Haider, A. (2018). Using corpus linguistic techniques in (critical) discourse studies reduces but does not remove bias: evidence from an Arabic corpus about refugees. *Poznan Studies in Contemporary Linguistics*, 55(1), 89-133. <https://doi.org/10.1515/psicl-2019-0004>
- Hobbs, V. (2019). The discourse of divorce in conservative Christian sermons. *Critical Discourse Studies*, 17(2), 193-210. <https://doi.org/10.1080/17405904.2019.1665079>
- Ismail, H., Baharun, H., Abdullah, H., & Majid, S. (2020). Power and voices of authority in the media narrative of Malaysian natives: combining corpus linguistics and discourse analysis approaches. *3L The Southeast Asian Journal of English Language Studies*, 26(2), 50-66. <https://doi.org/10.17576/3l-2020-2602-04>
- Leung, C., Chow, M., & Ge, H. (2024). Compiling data for investigating language use in classroom discourse: a corpus-driven method. *MethodsX*, 13, 103048. <https://doi.org/10.1016/j.mex.2024.103048>

- Lin, A. (2014). Critical discourse analysis in applied linguistics: a methodological review. *Annual Review of Applied Linguistics*, 34, 213-232. <https://doi.org/10.1017/s0267190514000087>
- Левінін, М. (2024). Language teaching challenges through the lens of corpus linguistics. *TEFLIN Journal - A Publication on the Teaching and Learning of English*, 35(1), 40-65. <https://doi.org/10.15639/teflinjournal.v35i1/40-65>
- Marchi, A. (2022). Corpus linguistics in the study of news media. In *The Routledge Handbook of English Language and Digital Humanities* (pp. 576-588). <https://doi.org/10.4324/9780367076399-40>
- Ortiz, A. and García-Gámez, M. (2023). Strategies for the analysis of large social media corpora: sampling and keyword extraction methods. *Corpus Pragmatics*, 7(3), 241-265. <https://doi.org/10.1007/s41701-023-00143-0>
- Pérez, M. (2023). The representation of migration in parliamentary settings: critical cross-linguistics corpus-assisted discourse analyses. *Humanities and Social Sciences Communications*, 10(1). <https://doi.org/10.1057/s41599-023-02496-y>
- Potts, A. and Semino, E. (2017). Healthcare professionals' online use of violence metaphors for care at the end of life in the US: a corpus-based comparison with the UK. *Corpora*, 12(1), 55-84. <https://doi.org/10.3366/cor.2017.0109>
- Rasulo, M. (2021). Are gold hoop earrings and a dab of red lipstick enough to get even Democrats on the offensive? *Journal of Language Aggression and Conflict*, 9(1), 155-183. <https://doi.org/10.1075/jlac.00057.ras>
- Richter, K., Gaskaree, B., & Mirzai, M. (2022). A functional analysis of lexical bundles in the discussion sections of applied linguistics research articles: a cross-paradigm study. *Russian Journal of Linguistics*, 26(3), 625-644. <https://doi.org/10.22363/2687-0088-27752>
- Romero-Barranco, J. and Rodríguez–Abruñeiras, P. (2021). Current trends in corpus linguistics and textual variation. *Research in Corpus Linguistics*, 9(2), i-xiii. <https://doi.org/10.32714/ricl.09.02.01>
- Roshdy, R. (2024). Exploring the interface between corpus and statistical modelling: a multi-methodological analysis of lexical variation in Islamic legal discourse. *JoSTrans*, 42(42), 126-154. <https://doi.org/10.26034/cm.jostrans.2024.5986>
- Serafis, D., Raimondo, C., Assimakopoulos, S., Greco, S., & Rocci, A. (2021). Argumentative dynamics in representations of migrants and refugees: evidence from the Italian press during the ‘refugee crisis’. *Discourse & Communication*, 15(5), 559-581. <https://doi.org/10.1177/17504813211017706>

- Wang, J. and Jin, G. (2023). Critical discourse studies eleven years on in China (from 2011 to 2021): a critical review. *Discourse Studies*, 25(3), 361-382. <https://doi.org/10.1177/14614456231158519>
- Zhang, J. (2022). The “Chinese Dream” as a discursive symbol. *Chinese Semiotic Studies*, 18(4), 605-631. <https://doi.org/10.1515/css-2022-2082>