World Journal of Corpus Linguistics

Vol. 4, No. 1-2025, Page 12-21

Available online at https://pusatpublikasi.com/index.php/corpus/

CORPUS LINGUISTICS IN DISCOURSE ANALYSIS: A SYSTEMATIC LITERATURE REVIEW OF METHODOLOGICAL INNOVATIONS AND EMPIRICAL APPLICATIONS (2015-2025)

Gabriel Mau Mali
Warmadewa University
E-mail: gerrymali74@gmail.com

Abstract
This systematic literature review investigates methodological innovations and empirical applications of corpus linguistics in discourse analysis from 2015 to 2025. Drawing on 45 empirical studies retrieved from major academic databases; Scopus, Web of Science, and Google Scholar, this research identifies emerging methodological patterns, technological advancements, and ongoing theoretical challenges within Corpus-Assisted Discourse Studies (CADS). The findings demonstrate that integrating corpus linguistics with critical discourse analysis has produced substantial methodological synergy, enabling systematic, evidence-based interpretation of linguistic patterns across large-scale textual corpora. The review delineates five principal domains of application: media and political discourse, social group representation, health and environmental communication, multimodal discourse analysis, and the integration of artificial intelligence technologies. Despite these advances, methodological constraints persist, including issues of researcher bias, corpus representativeness, and limited resources for non-English language data. The study's theoretical contribution lies in providing a comprehensive mapping of CADS as a transdisciplinary framework that fuses quantitative corpus methodologies with qualitative discourse interpretation. Practically, the review underscores the need for greater methodological transparency, development of corpus tools for under-resourced languages, and ethically informed adoption of AI-driven methods in discourse research. Ultimately, this review offers a systematic conceptual foundation for scholars employing corpus-based approaches in discourse studies and highlights future research trajectories involving multimodal analysis, diachronic corpora, and the expansion of CADS in Global South contexts.

Keywords: *corpus linguistics; discourse analysis; corpus-assisted discourse studies; critical discourse analysis; systematic literature review; methodological innovation*.

INTRODUCTION

Discourse analysis as a linguistic research method has undergone significant transformation with the emergence of corpus-based approaches. The integration of corpus linguistics with discourse analysis, known as Corpus-Assisted Discourse Studies (CADS), offers an empirical approach for identifying systematic linguistic patterns in large-scale textual datasets. This approach enables researchers to move beyond single-text analysis or small samples characteristic of traditional discourse analysis toward evidence-based generalizations that can be qualitatively interpreted within broader social and political contexts.

The theoretical foundations of CADS rest on recognition that language constitutes social practice through which power relations, ideologies, and social structures are constructed and maintained. Fairclough (2015) establishes that discourse functions as dialectical element of social practice, simultaneously shaped by and shaping social structures. When combined with corpus linguistic methods enabling systematic identification of repetitive linguistic patterns

across large text collections, discourse analysis gains empirical grounding that strengthens its analytical claims (Baker, 2006; Partington, Duguid, and Taylor, 2013).

Baker and colleagues (2008) demonstrate that combining critical discourse analysis with corpus linguistics creates useful methodological synergy, where quantitative corpus methods provide empirical evidence of linguistic patterns while discourse analysis offers deep interpretation of ideological implications and power relations embedded in language use. This synergy addresses limitations inherent in purely qualitative discourse analysis, particularly subjectivity in text selection and inability to make generalizable claims about discourse patterns (Gillings, Mautner, and Baker, 2023).

Computational technology development and increasing availability of digital corpora have driven CADS proliferation in the past decade. Taboada (2025) observes that corpus linguistic methods have become increasingly sophisticated, with topic modeling, sentiment analysis, and other computational techniques expanding analytical possibilities. However, Incelli (2025) warns that artificial intelligence integration in corpus linguistics raises methodological and ethical questions requiring critical attention, particularly concerning data integrity, algorithmic bias, and deterministic versus non-deterministic technologies.

Despite methodological advancement, significant gaps remain in CADS literature. First, most studies concentrate on English-language corpora, creating imbalance in linguistic diversity representation (Biber, Conrad, and Reppen, 1998). Second, integration between corpus methods and qualitative discourse interpretation often remains unidirectional, with corpus analysis generating findings subsequently interpreted qualitatively without genuine methodological integration (Baker, 2020). Third, rapid AI technology development requires systematic evaluation of generative AI potential and limitations in corpus discourse research.

This systematic literature review addresses these gaps by comprehensively mapping CADS methodological evolution, identifying major application domains, evaluating AI technology integration, and formulating recommendations for future research. Specific research questions guiding this review include: (1) What methodological innovations have emerged in corpus-based discourse analysis during 2015-2025? (2) What are the primary application domains of CADS in contemporary research? (3) What are methodological strengths and limitations of corpus approaches in discourse studies? (4) How has AI technology integration influenced CADS development? (5) What future research directions can advance the field? This review contributes to corpus linguistics and discourse studies by providing systematic framework for understanding field evolution and identifying promising research trajectories.

METHOD

Review Design

This study employs a systematic literature review methodology following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency, replicability, and comprehensiveness in literature selection and analysis. The systematic approach enables identification of methodological patterns, theoretical trends, and research gaps in corpus-based discourse analysis field through structured procedures for study identification, screening, eligibility assessment, and synthesis.

Search Strategy and Data Sources

Literature search was conducted across three major academic databases: Scopus, Web of Science, and Google Scholar, supplemented by specialized linguistics databases and manual searching of reference lists from key publications. The temporal scope covers publications from 2015 to 2025, capturing the most recent decade of methodological and technological developments in CADS. Search terms combined corpus linguistics terminology with discourse

analysis concepts using Boolean operators: (corpus OR corpora) AND (discourse analysis OR critical discourse analysis OR CADS OR corpus-assisted discourse studies) AND (method OR methodology OR approach). Additional searches targeted specific journals known for CADS publications including Discourse and Society, Corpora, International Journal of Corpus Linguistics, and Applied Linguistics.

Inclusion and Exclusion Criteria

Inclusion criteria specified peer-reviewed journal articles and book chapters published in English between 2015-2025 that employed corpus linguistic methods for discourse analysis, presented empirical research with explicit methodological description, and addressed theoretical or methodological aspects of corpus-discourse integration. Exclusion criteria eliminated conference abstracts without full papers, purely theoretical discussions without empirical application, studies using corpus methods for purposes other than discourse analysis, such as purely grammatical or lexicographic studies, and publications not accessible through institutional access or open-access repositories.

Selection Process and Quality Assessment

Initial database searches yielded 127 potentially relevant publications. Title and abstract screening reduced this to 68 studies warranting full-text examination. Applying inclusion and exclusion criteria strictly resulted in final selection of 45 studies for comprehensive analysis. Quality assessment evaluated each study across multiple dimensions: methodological rigor in corpus construction and analysis procedures, clarity of research design description enabling replicability, integration quality between quantitative corpus findings and qualitative discourse interpretation, theoretical framework adequacy for situating findings, and contribution significance to CADS field advancement.

Analytical Framework

Selected studies underwent thematic analysis organized around five principal dimensions. Methodological approaches examined corpus construction procedures, analytical tools employed, and integration strategies between quantitative and qualitative methods. Application domains identified discourse types and contexts investigated. Theoretical frameworks analyzed discourse theories and linguistic models underpinning research. Technical innovations documented novel computational methods, software tools, and analytical techniques. Limitations and challenges synthesized methodological constraints, validity concerns, and practical difficulties reported. This multi-dimensional analytical framework enabled comprehensive synthesis of CADS research landscape while maintaining focus on methodological evolution as primary review objective.

RESULTS AND DISCUSSION

Methodological Innovations in Corpus-Assisted Discourse Studies

Analysis of selected studies reveals several significant methodological innovations characterizing contemporary CADS research. Most fundamentally, there has been evolution from purely keyword and collocation-based approaches toward more sophisticated multi-method frameworks. Baker (2023) in the updated edition of his seminal work emphasizes triangulation importance, combining CADS-style collocation analysis with qualitative close reading of text samples. This methodological pluralism addresses concerns raised by earlier critics regarding corpus linguistics limitations in capturing contextual nuances and pragmatic meanings.

Gillings, Mautner, and Baker (2023) provide comprehensive methodological framework for CADS emphasizing bidirectional workflow between corpus analysis and discourse interpretation. Rather than unidirectional process where corpus findings are subsequently interpreted, they advocate iterative cycle where qualitative hypotheses inform corpus queries, corpus results generate new qualitative questions, and both approaches mutually inform each other throughout research process. This represents significant advancement from earlier CADS models where quantitative and qualitative phases remained more distinct. Studies employing such integrative approaches demonstrate richer analytical insights and more nuanced interpretations of discourse patterns.

Technical methodological innovations include adoption of advanced concordance analysis techniques going beyond simple keyword-in-context displays. Researchers increasingly employ dispersion plots to track term distribution across corpus sections, n-gram analysis to identify fixed multi-word expressions, and semantic prosody analysis to uncover evaluative associations of seemingly neutral terms. Topic modeling using Latent Dirichlet Allocation has emerged as valuable exploratory tool, though Taboada (2025) cautions that topic modeling should be understood as means to end rather than analysis itself, requiring substantial interpretive work to connect computational topics with meaningful discourse categories.

Major Application Domains

Media and political discourse constitute the most prevalent application domain in reviewed literature. Studies examine how newspapers, television news, and social media construct representations of political actors, events, and issues. Baker and colleagues (2008) pioneering work on refugee and asylum seeker representation in UK press established template widely adopted in subsequent research. Recent studies extend this approach to examine discourse around migration, terrorism, populism, and polarization across multiple national contexts. Common findings across these studies include identification of systematic linguistic strategies for in-group legitimation and out-group delegitimation, use of metaphorical framing to naturalize particular ideological positions, and deployment of euphemism and dysphemism to manage evaluative meanings.

Social group representation research employs CADS to analyze how different demographic groups are constructed in various discourse types. Studies examine representations of gender, race, ethnicity, age, disability, and other identity categories across media, institutional, and everyday discourse. These investigations typically combine frequency analysis of demographic terms with examination of their collocational profiles to identify systematic associations. For instance, studies of gender representation analyze differential collocates of male and female terms to reveal implicit stereotyping, while research on ethnic minority representation documents patterns of criminalization or marginalization through linguistic association.

Health and environmental discourse has emerged as significant application area, particularly following COVID-19 pandemic. Corpus studies examine how public health crises are discursively constructed in media and policy documents, how medical and scientific terminology diffuses into public discourse, and how competing framings of environmental issues contend in policy debates. Research on COVID-19 discourse documents rapid lexical innovation including neologism formation and semantic change in existing terms. Studies of climate change discourse analyze metaphorical framing variations across different political and cultural contexts, revealing systematic differences in how climate issues are conceptualized and

solutions proposed.

Multimodal discourse analysis represents expanding frontier for CADS. While traditional corpus linguistics focuses exclusively on linguistic features, recent research integrates analysis of visual elements, layout, typography, and other semiotic resources. Studies combine corpus linguistic analysis of textual components with qualitative analysis of accompanying images, developing frameworks for systematic description of text-image relationships and their contribution to overall discourse construction. This multimodal turn responds to recognition that contemporary discourse, particularly in digital environments, rarely consists of pure text but rather multimodal assemblages requiring integrated analytical approaches.

Artificial intelligence integration represents newest application domain, with researchers exploring both AI as object of study and as methodological tool. Studies analyze discourse about AI technologies examining how artificial intelligence is represented in media, policy, and public discussion. Simultaneously, researchers investigate AI tools potential for corpus analysis itself. However, Incelli (2025) comprehensive evaluation finds significant limitations in current generative AI capabilities for CADS research, particularly regarding data integrity, false inferences, and inability to perform reliable concordance and function-to-form analysis. This suggests that while AI offers promising directions, human expertise remains essential for valid corpus discourse research.

Theoretical Frameworks and Discourse Models

Reviewed studies predominantly draw on Critical Discourse Analysis frameworks, particularly Fairclough three-dimensional model analyzing text, discursive practice, and sociocultural practice dimensions. Fairclough (2015) emphasis on discourse as dialectical element of social practice provides theoretical justification for examining systematic linguistic patterns as evidence of underlying ideological structures and power relations. Studies applying Fairclough framework typically move from micro-level textual description through meso-level analysis of production and consumption practices to macro-level explanation of sociocultural context and consequences.

Systemic Functional Linguistics, particularly Halliday framework for analyzing ideational, interpersonal, and textual metafunctions, provides another influential theoretical foundation. This framework aligns well with corpus approaches because it offers systematic method for categorizing linguistic features according to their functional contributions to meaning-making. Studies employing SFL typically conduct detailed grammatical analysis of corpus texts, examining features such as transitivity patterns revealing agency and causality representations, modality expressing epistemic and deontic stance, and theme-rheme structures organizing information flow.

Cognitive linguistic frameworks, particularly Conceptual Metaphor Theory, inform studies examining metaphorical patterns in corpora. These investigations identify systematic metaphor usage through corpus-based analysis of source-target domain mappings, demonstrating how conceptual structures shape discourse about abstract topics. Research on political discourse, for instance, documents warfare, journey, and building metaphors prevalence for conceptualizing political processes, while environmental discourse studies reveal nature-as-machine versus nature-as-living-organism metaphorical framings divergent implications.

Methodological Strengths and Advantages

CADS primary methodological strength lies in scalability enabling analysis of text volumes impractical for traditional close reading approaches. This scalability permits identification of patterns that might be invisible in small samples but become evident when examining thousands or millions of word tokens. The empirical grounding provided by frequency data, statistical significance testing, and systematic concordance analysis strengthens analytical claims beyond impressionistic observations possible from limited textual analysis. Biber and Conrad (2019) emphasize that corpus-based findings regarding linguistic patterns possess generalizability impossible with small-scale qualitative analysis alone.

Reduction of researcher bias constitutes another significant advantage. While no analysis achieves complete objectivity, corpus methods systematicity reduces subjective text selection where researchers might unconsciously favor examples supporting preconceived hypotheses. Comprehensive corpus examination forces researchers to confront disconfirming evidence and account for variation across texts. Furthermore, corpus approach transparency, where data and procedures can be explicitly documented and potentially replicated by other researchers, enhances research accountability and enables verification of findings.

Discovery of non-obvious patterns represents particularly valuable CADS contribution. As Gillings, Mautner, and Baker (2023) note, corpus analysis can reveal meanings not readily apparent to unaided reading. Collocational analysis might uncover systematic associations that readers do not consciously register but that cumulatively construct particular representations. Discourse prosody analysis can identify evaluative colorings of apparently neutral terms that only become visible through examination of typical contexts. These discoveries often provide empirical support for critical discourse analysts intuitions while also generating unexpected findings challenging existing interpretations.

Methodological Limitations and Challenges

Despite methodological advantages, CADS faces several significant limitations requiring acknowledgment. Decontextualization risk remains persistent challenge, as corpus methods typically analyze isolated linguistic features or short concordance lines removed from their broader textual and situational contexts. While this decontextualization enables pattern identification across texts, it can obscure pragmatic meanings dependent on specific communicative situations. Critics argue that focusing on frequent patterns may lead researchers to overlook significant but infrequent discourse strategies or to miss interpretive subtleties visible only through sustained engagement with complete texts in their original contexts.

Corpus representativeness and sampling issues present ongoing methodological challenges. Biber, Conrad, and Reppen (1998) establish that corpus composition fundamentally shapes research findings, yet many CADS studies employ convenience samples of readily available texts rather than carefully stratified representative corpora. Studies of media discourse, for instance, often analyze major national newspapers while underrepresenting local, regional, or alternative media sources. This sampling bias can produce skewed understanding of discourse patterns if analyzed corpus does not adequately represent discourse domain under investigation.

The quantitative-qualitative integration challenge persists despite methodological innovations. Many studies struggle to achieve genuine dialogue between corpus analysis and discourse interpretation, instead presenting corpus findings followed by separate qualitative discussion without clear articulation of how the two inform each other. This integration

difficulty partly reflects different epistemological assumptions underlying quantitative corpus linguistics and interpretive discourse analysis. Developing frameworks for productive conversation between these traditions remains ongoing methodological challenge requiring continued innovation.

Technical and resource constraints affect CADS research particularly for non-English languages. While sophisticated corpus analysis software exists for English, many languages lack comparable tools for morphological analysis, part-of-speech tagging, or semantic annotation. Building quality corpora requires significant time, expertise, and technical infrastructure not equally available to researchers across institutional and geographic contexts. These resource disparities contribute to English-language dominance in CADS literature, limiting insights into discourse patterns in world linguistic diversity.

Artificial Intelligence Integration: Opportunities and Concerns

Artificial intelligence integration in CADS represents rapidly evolving frontier with significant implications for future research directions. Incelli (2025) provides critical evaluation of generative AI, specifically ChatGPT, for corpus approaches to discourse studies through replication case studies. Findings reveal that while ChatGPT performs reasonably well at semantically categorizing decontextualized keywords, several serious limitations emerge. For concordance analysis, ChatGPT performs poorly, generating false inferences about concordance lines and sometimes modifying input data. Function-to-form analysis likewise shows poor performance, with ChatGPT failing to identify and analyze linguistic features systematically.

These limitations raise fundamental questions about AI affordances for supporting automated qualitative analysis within CADS. Incelli signals issues of repeatability and replicability, critical concerns for scientific research validity. The non-deterministic nature of large language models means identical queries may produce different results, undermining reliability essential for empirical research. Furthermore, AI modifications of input data pose serious ethical challenges regarding data integrity. For corpus discourse research requiring precise analysis of actual language use, such data alterations are unacceptable.

However, AI offers valuable potential for specific CADS tasks where its capabilities align with methodological requirements. Automated corpus construction through web scraping, initial data cleaning and preprocessing, and preliminary thematic categorization represent areas where AI assistance could enhance research efficiency while human oversight maintains quality control. Machine learning approaches to topic modeling, sentiment analysis, and pattern recognition continue advancing, offering increasingly sophisticated tools for exploratory corpus analysis when critically and carefully deployed.

The key conclusion regarding AI integration is that while these technologies offer promising assistance for certain CADS tasks, they cannot replace human expertise in corpus discourse analysis. The interpretive work central to discourse analysis, requiring deep contextual understanding, theoretical sophistication, and critical awareness of language-power relationships, remains fundamentally human endeavor. Future CADS research should explore productive human-AI collaboration models where computational power augments rather than replaces human analytical capabilities.

CONCLUSION

This systematic literature review demonstrates that corpus linguistics in discourse analysis has evolved into mature methodological approach characterized by sophisticated integration of quantitative and qualitative techniques. The emergence of Corpus-Assisted Discourse Studies as distinct subfield reflects successful synthesis of corpus linguistic empiricism with discourse analysis interpretive depth, creating methodological synergy addressing limitations of purely quantitative or purely qualitative approaches. Analysis of 45 empirical studies reveals consistent methodological innovations, expanding application domains, and ongoing refinement of theoretical frameworks guiding CADS research.

Key theoretical contributions of this review include comprehensive mapping of CADS methodological evolution showing progression from simple frequency-based approaches toward sophisticated multi-method frameworks emphasizing iterative integration between corpus analysis and discourse interpretation. The identification of five major application domains demonstrates field breadth spanning media and political discourse, social group representation, health and environmental communication, multimodal analysis, and AI discourse. These applications share common concern with revealing systematic linguistic patterns constructing social reality, legitimating power relations, and reproducing or challenging ideological structures.

Methodologically, the review establishes that CADS primary strength lies in scalability, empirical grounding, and capacity for discovering non-obvious discourse patterns through systematic analysis of large text collections. These advantages enable researchers to make generalizable claims about discourse patterns with greater confidence than possible from small-sample qualitative analysis. However, significant limitations persist, particularly regarding decontextualization risks, corpus representativeness challenges, quantitative-qualitative integration difficulties, and resource constraints especially for non-English languages.

Critical evaluation of AI integration reveals cautious stance warranted regarding generative AI current capabilities for corpus discourse research. While AI offers valuable assistance for specific tasks such as corpus construction and preliminary categorization, fundamental limitations in data integrity, repeatability, and analytical reliability prevent AI from replacing human expertise in corpus discourse analysis. Future research should explore productive human-AI collaboration models rather than AI automation of analytical processes.

Practically, this review offers several recommendations for researchers and the field. First, greater methodological transparency is essential, with studies providing detailed documentation of corpus construction procedures, analytical tools employed, and integration strategies between quantitative and qualitative phases. Second, development of corpus analysis tools for under-resourced languages requires investment and international collaboration to address current English-language dominance. Third, training in CADS methodology should emphasize both computational skills and theoretical sophistication in discourse analysis, moving beyond either purely technical or purely interpretive approaches.

Future research directions identified include expansion of multimodal CADS combining textual analysis with systematic examination of visual and other semiotic resources in contemporary discourse. Diachronic corpus studies tracking discourse evolution over time offer valuable insights into language change and social transformation relationships. Comparative cross-linguistic and cross-cultural CADS research could illuminate how similar social phenomena are discursively constructed differently across languages and cultural contexts. Finally, critical attention to AI technologies ethical implications in discourse research, particularly concerning data privacy, algorithmic bias, and power asymmetries between Global North and South in AI development and deployment, represents important emerging concern.

In conclusion, corpus linguistics in discourse analysis represents vibrant and evolving

field characterized by methodological innovation, theoretical sophistication, and expanding empirical applications. While challenges remain, particularly regarding methodological integration and resource equity, CADS offers powerful approach for systematic investigation of how language constructs social reality, reproduces power relations, and enables resistance to dominant ideologies. Continued development of this approach, guided by critical reflexivity about its possibilities and limitations, promises valuable contributions to understanding language role in social life.

REFERENCES

Baker, P. (2006). Using corpora in discourse analysis. London: Continuum.

Baker, P. (2020). Corpus-assisted discourse analysis. In P. Baker (Ed.), Researching discourse (pp. 131-149). London: Routledge.

Baker, P. (2023). Using corpora in discourse analysis (2nd ed.). London: Bloomsbury.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society, 19(3), 273-306.

Biber, D., & Conrad, S. (2019). Register, genre, and style (2nd ed.). Cambridge: Cambridge University Press.

Biber, D., Connor, U., & Upton, T. A. (2007). Discourse on the move: Using corpus analysis to describe discourse structure. Amsterdam: John Benjamins.

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Conrad, S. M. (2002). Corpus linguistic approaches for discourse analysis. Annual Review of Applied Linguistics, 22, 75-95.

Fairclough, N. (2015). Language and power (3rd ed.). London: Routledge.

Flowerdew, J., & Richardson, J. E. (Eds.). (2018). The Routledge handbook of critical discourse studies. London: Routledge.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. Journal of English Linguistics, 36(1), 5-38.

Gillings, M., Mautner, G., & Baker, P. (2023). Corpus-assisted discourse studies. Cambridge: Cambridge University Press.

Gillings, M., Mautner, G., & Baker, P. (2024). Taking the road less travelled: How corpus-assisted discourse studies can enrich qualitative explorations of large textual datasets. British Journal of Management, 35(2), 883-903.

Incelli, E. (2025). Exploring the future of corpus linguistics: Innovations in AI and social impact. International Journal of Mass Communication, 3, 1-10.

Incelli, E. (2023). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. International Journal of Corpus Linguistics, 29(1), 129-155.

Jaworska, S., & Nanda, A. (2018). Doing well by talking good: A topic modelling-assisted discourse study of Corporate Social Responsibility. Applied Linguistics, 39(3), 373-399.

Mautner, G. (2019). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (Eds.), Methods of critical discourse studies (3rd ed., pp. 122-143). London: Sage.

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

Partington, A., Duguid, A., & Taylor, C. (2013). Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS). Amsterdam: John Benjamins.

Rao, Y., & Taboada, M. (2021). Gender bias in the news: A scalable topic modelling and visualization framework. Frontiers in Artificial Intelligence, 4, 664737.

Stubbs, M. (1996). Text and corpus analysis: Computer-assisted studies of language and culture.

Oxford: Blackwell.

Stubbs, M. (2001). Words and phrases: Corpus studies of lexical semantics. Oxford: Blackwell.

Taboada, M. (2025). Topic modelling is a means to an end: On topic modelling in corpus linguistics and discourse analysis. Discourse Studies, 27(1), 3-8.

Wodak, R., & Meyer, M. (Eds.). (2016). Methods of critical discourse studies (3rd ed.). London: Sage.