World Journal of Corpus Linguistics

Vol. 4, No. 1-2025, Page 1-11

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

COMPUTATIONAL CORPUS LINGUISTICS: A SYSTEMATIC LITERATURE REVIEW OF METHODOLOGICAL INTEGRATION, TECHNOLOGICAL INNOVATIONS, AND ARTIFICIAL INTELLIGENCE APPLICATIONS (2015-2025)

Candiawan Telaumbanua
Bali Business School
E-mail: candifiber21@gmail.com

Abstract
This systematic literature review explores the convergence of computational methodologies and corpus linguistics between 2015 and 2025, synthesizing insights from 52 empirical studies retrieved from authoritative academic databases, including the ACL Anthology, Scopus, and Web of Science. The review examines how advances in natural language processing, machine learning, and artificial intelligence have reshaped the theoretical and methodological foundations of corpus linguistics, enabling analysis of massive textual datasets that exceed the capacity of traditional corpus tools. Findings reveal five dominant computational innovations driving this transformation: deep learning for automated annotation and classification, text similarity modeling for large-scale corpus comparison, topic modeling and distributional semantics for linguistic pattern discovery, neural machine translation for multilingual corpus processing, and large language model integration for corpus construction and analytical enhancement. These developments mark substantial progress in computational efficiency, scalability to billion-word corpora, and automation of formerly manual linguistic tasks. Nevertheless, the review identifies persistent challenges related to algorithmic bias, interpretability, ethical responsibility in automated language analysis, and the widening digital divide affecting under-resourced languages. Theoretically, the study maps emerging synergies between computational and linguistic paradigms, highlights hybrid research frameworks uniting symbolic and statistical approaches, and proposes ethical principles for responsible computational corpus inquiry. Practically, it underscores the urgency of interdisciplinary training that bridges linguistics and computer science, the development of interpretable and transparent machine learning models for linguistic research, and the equitable allocation of computational resources to support global linguistic diversity. The review concludes by outlining future research directions in explainable AI, multimodal corpus analysis, and decolonial perspectives on language technology development.

Keywords: *computational corpus linguistics, natural language processing, machine learning, artificial intelligence, deep learning, large language models.*

INTRODUCTION

Corpus linguistics has experienced paradigmatic transformation through integration with computational methods from natural language processing, machine learning, and artificial intelligence. The exponential growth in available textual data combined with computational power advances has fundamentally altered the scale, scope, and methodologies of corpus-based language research. Where the pioneering Brown Corpus contained one million words and represented breakthrough achievement in 1960s computational linguistics, contemporary corpora routinely encompass billions of words, with the Corpus of Global Language Use containing 400 billion words across 50 language varieties. This scale transformation

necessitates computational methods capable of processing, analyzing, and interpreting linguistic patterns across datasets impossible for traditional manual or semi-automated corpus analysis techniques.

Dunn (2022) emphasizes that computational corpus linguistics represents more than simply applying existing corpus methods to larger datasets; it fundamentally reconceptualizes linguistic analysis through machine learning paradigms that discover patterns without predetermined linguistic categories. Text classification models enable automatic register identification across vast corpora, text similarity measures facilitate cross-linguistic comparison at unprecedented scale, and neural network architectures learn latent linguistic representations that capture semantic and syntactic regularities. These computational advances promise transformative insights into language variation, change, and structure, yet simultaneously raise profound methodological, theoretical, and ethical questions about automated linguistic analysis validity, interpretability, and social consequences.

Jurafsky and Martin (2024) in their comprehensive overview of natural language processing establish that contemporary computational linguistics synthesizes symbolic rule-based approaches with statistical and neural methods, creating hybrid systems that leverage both linguistic theory and data-driven pattern discovery. This integration challenges traditional boundaries between theoretical linguistics emphasizing competence and corpus linguistics emphasizing performance, suggesting computational methods provide bridge connecting formal linguistic knowledge with empirical language use patterns. However, McEnery and Hardie (2012) caution that computational sophistication must not obscure linguistic rigor, arguing that corpus linguistics remains fundamentally linguistic inquiry rather than purely computational data science.

Recent developments in artificial intelligence, particularly large language models such as GPT series, BERT, and multilingual transformers, introduce both unprecedented opportunities and critical challenges for computational corpus linguistics. Incelli (2025) documents how AI technologies are transforming corpus construction, annotation, and analysis workflows, while simultaneously raising concerns about data integrity, algorithmic bias, linguistic diversity preservation, and epistemological foundations of automated language analysis. The integration of generative AI in corpus research represents inflection point requiring systematic evaluation of capabilities, limitations, and appropriate applications within linguistic research contexts.

Despite rapid technological advancement, significant gaps persist in computational corpus linguistics literature. First, most computational resources and research concentrate on English and other high-resource languages, exacerbating global linguistic inequality. Joshi and colleagues (2020) document that the majority of world languages remain computationally under-resourced, lacking annotated corpora, processing tools, and trained models necessary for computational linguistic analysis. This digital language divide threatens linguistic diversity and perpetuates technological colonialism wherein language technologies serve primarily dominant languages and their speaker communities.

Second, methodological integration between computational methods and linguistic theory remains incomplete. Many applications employ machine learning as black-box tools without linguistic interpretation or validation of learned representations against linguistic knowledge. Conversely, some corpus linguistics research fails to leverage computational advances that could enhance analytical rigor and scalability. Achieving productive integration requires interdisciplinary expertise spanning linguistics, computer science, and statistics, alongside institutional structures facilitating cross-disciplinary collaboration.

This systematic literature review addresses these gaps by comprehensively surveying computational corpus linguistics developments during 2015-2025, systematically mapping methodological innovations, identifying major application domains, evaluating technological

capabilities and limitations, and articulating ethical frameworks for responsible computational language research. Specific research questions guiding this review include: (1) What computational methods from NLP and machine learning have been integrated into corpus linguistics research? (2) How do these computational approaches transform traditional corpus linguistic methodologies? (3) What are the primary application domains of computational corpus linguistics? (4) What methodological advantages and limitations characterize computational approaches? (5) What ethical considerations emerge from automated large-scale language analysis? (6) How can computational corpus linguistics advance equitably across global linguistic diversity? This review contributes to both corpus linguistics and computational linguistics by providing systematic framework for understanding methodological convergence, identifying productive research directions, and establishing ethical principles for responsible technological development.

METHOD

Review Design and Scope

This study employs systematic literature review methodology following PRISMA guidelines adapted for computational linguistics research. The review scope encompasses peer-reviewed publications from 2015-2025 investigating integration of computational methods with corpus linguistics, including both methodological innovations and empirical applications. The temporal focus captures the recent decade characterized by deep learning revolution, large language model emergence, and massively scaled corpus availability, representing distinct computational linguistics era.

Search Strategy and Data Sources

Literature search utilized multiple specialized databases including ACL Anthology for computational linguistics publications, Scopus and Web of Science for interdisciplinary coverage, and discipline-specific journals including International Journal of Corpus Linguistics, Corpus Linguistics and Linguistic Theory, and Applied Corpus Linguistics. Search terms combined corpus linguistics terminology with computational methods: (corpus linguistics OR corpora) AND (computational OR natural language processing OR machine learning OR deep learning OR neural network OR artificial intelligence). Additional searches targeted specific computational techniques: text classification, word embeddings, transformer models, large language models, topic modeling, and automated annotation. Manual search of reference lists from key publications supplemented database searching.

Inclusion and Exclusion Criteria

Inclusion criteria specified peer-reviewed publications in English from 2015-2025 that integrate computational methods with corpus linguistic analysis, present empirical research with explicit methodology, and address either methodological innovation or substantive linguistic findings using computational corpus approaches. Exclusion criteria eliminated purely theoretical computational linguistics without corpus application, corpus studies using only traditional non-computational methods, technical computer science papers without linguistic focus, and preliminary conference abstracts without full methodological description. Publications addressing corpora construction, annotation tools, and linguistic resource development were included when demonstrating computational-linguistic integration.

Selection Process and Quality Assessment

Initial searches yielded 156 potentially relevant publications. Title and abstract screening reduced this to 84 studies warranting full-text examination. Applying inclusion and exclusion criteria strictly resulted in final selection of 52 studies for comprehensive analysis. Quality assessment evaluated methodological rigor of computational implementations, linguistic validity of analyses, integration quality between computational and linguistic approaches, reproducibility based on methodology description, and contribution significance to computational corpus linguistics field. Studies demonstrating both computational sophistication and linguistic depth received highest ratings, while purely technical or insufficiently linguistic works were excluded.

Analytical Framework

Selected studies underwent thematic analysis organized around six principal dimensions. Computational methods examined specific NLP and machine learning techniques employed, including algorithms, models, and tools. Corpus applications identified linguistic phenomena investigated and corpus types analyzed. Methodological integration assessed how computational and linguistic approaches were combined. Scalability and performance evaluated computational efficiency and dataset size capabilities. Limitations and challenges documented technical, methodological, and theoretical constraints. Ethical considerations analyzed social implications, bias issues, and equity concerns. This multi-dimensional framework enables comprehensive synthesis while maintaining focus on computational-linguistic integration as primary review objective.

RESULTS AND DISCUSSION

Deep Learning for Automated Linguistic Annotation

Deep learning applications represent major computational innovation enabling automated linguistic annotation at scale previously impossible with rule-based systems or traditional machine learning. Fonteyn, Manjavacas, and De Regt (2025) demonstrate using MacBERTh, a BERT-based model fine-tuned on historical texts, for automated corpus annotation achieving high accuracy while reducing annotation time significantly. Their case study shows how transformer-based language models can learn contextual representations enabling accurate part-of-speech tagging, syntactic parsing, and semantic annotation across historical language varieties lacking extensive training data.

The advantage of deep learning annotation lies in transfer learning capabilities, where models pre-trained on large general corpora can be fine-tuned for specific linguistic tasks with relatively small labeled datasets. This addresses persistent corpus linguistics challenge of annotation bottleneck, where manual linguistic annotation requires extensive time and expert knowledge. However, automated annotation quality remains variable across languages, registers, and linguistic levels. Reveilhac and Schneider (2025) evaluate stance detection in social media data using linguistic markers, finding that while machine learning achieves reasonable accuracy, transparency and interpretability remain critical concerns for linguistic research validity.

Critical evaluation reveals that automated annotation accuracy depends fundamentally on training data quality and representativeness. Models trained primarily on formal written language perform poorly on informal registers, social media, or dialectal varieties. Furthermore, annotation errors can propagate through analytical pipelines, potentially generating spurious linguistic findings. Best practices emerging from reviewed studies emphasize systematic

validation of automated annotations against gold-standard manual annotations, explicit reporting of model performance metrics, and acknowledgment of annotation limitations in linguistic interpretation.

Text Classification and Corpus Register Analysis

Text classification models enable automatic categorization of corpus texts by register, genre, style, authorship, or other linguistically relevant dimensions. Dunn (2022) demonstrates how supervised machine learning classifiers trained on linguistically annotated features can identify discourse registers across web corpora containing billions of words, scaling register analysis far beyond traditional manual categorization. Text classification approaches range from traditional machine learning using linguistic feature engineering to deep learning models learning latent text representations directly from raw text without predetermined linguistic features.

Feature-based classification approaches leverage corpus linguistic expertise by engineering features encoding linguistic patterns characteristic of different text types. These may include lexical statistics (word frequencies, type-token ratios), syntactic patterns (passive voice frequency, sentence complexity measures), discourse markers, and register-specific vocabulary. Machine learning algorithms such as support vector machines, random forests, or logistic regression learn classification rules from these linguistically motivated features. This approach maintains interpretability, as classification decisions can be traced to specific linguistic patterns, facilitating linguistic understanding of learned categories.

Deep learning approaches using neural networks, particularly transformer architectures, learn text representations automatically from raw text without explicit linguistic feature engineering. Models such as BERT, RoBERTa, and XLM-RoBERTa achieve state-of-the-art classification performance across numerous text categorization tasks. However, these models function as black boxes where classification decisions emerge from complex non-linear transformations of input text through multiple neural network layers, making linguistic interpretation challenging. Research addressing this interpretability gap employs attention visualization, probing tasks testing linguistic knowledge, and feature attribution methods identifying text regions most influential for classification decisions.

Text Similarity and Cross-Linguistic Corpus Comparison

Text similarity modeling enables systematic comparison across corpus texts, language varieties, temporal periods, or languages, extending corpus linguistics comparative methodologies to unprecedented scale. Dunn (2022) demonstrates using distributional similarity measures combined with machine learning for comparing linguistic constructions across geographic varieties, identifying systematic variation patterns revealing dialectal boundaries and language contact effects. Similarity measures range from simple lexical overlap statistics to sophisticated semantic similarity models using word embeddings and sentence encoders.

Word embedding models including word2vec, GloVe, and fastText learn dense vector representations where semantically similar words occupy proximate vector space regions. These embeddings enable computational operations on word meanings, such as calculating semantic similarity between words or identifying semantic analogy relationships. Cross-linguistic word embeddings align vector spaces across languages, enabling multilingual similarity comparison without parallel corpora. Yang, Zhou, and Lin (2025) demonstrate

quantifying semantic similarity across English translations of classical Chinese texts using large language models, revealing systematic patterns of semantic preservation and loss across translations.

Sentence and document similarity models extend beyond word-level to capture longer text unit meanings. Transformer-based sentence encoders such as Sentence-BERT produce dense vector representations of sentences enabling efficient semantic similarity calculation. These models facilitate corpus-scale document clustering, duplicate detection, and cross-corpus comparison. Applications include identifying similar documents across different corpora, tracking content propagation through social media, and measuring intertextuality relationships. However, similarity model quality varies significantly across languages, with best performance on high-resource languages where extensive training data enables robust model development.

Topic Modeling and Distributional Semantics

Topic modeling provides computational methods for discovering thematic structure in large document collections without predetermined categories. Latent Dirichlet Allocation, the most widely used topic modeling algorithm, assumes documents comprise mixtures of topics, where topics are probability distributions over vocabulary. Applied to corpora, topic modeling reveals predominant themes, tracks thematic evolution over time, and enables corpus exploration by topic. However, topic modeling interpretability requires substantial human judgment in labeling and validating discovered topics, making it exploratory tool rather than definitive analysis.

Distributional semantic models exploit distributional hypothesis that words occurring in similar contexts have related meanings. These models analyze word co-occurrence patterns across large corpora to induce semantic representations. Beyond word embeddings discussed earlier, distributional semantics encompasses diverse approaches including count-based models using pointwise mutual information, neural network models learning latent semantic representations, and contextualized models such as ELMo and BERT producing context-dependent word representations. Distributional approaches enable corpus-based investigation of semantic change, semantic field structure, and cross-linguistic semantic variation.

Critical evaluation reveals topic modeling and distributional semantics limitations alongside their capabilities. Topic model quality depends heavily on corpus characteristics, preprocessing decisions, and hyperparameter settings, with relatively minor changes potentially producing substantially different topic structures. Distributional semantic models conflate different types of word relationships, clustering synonyms, antonyms, and associates without distinguishing relationship types. Furthermore, these models capture statistical association patterns that may or may not align with linguistically or cognitively relevant semantic structures. Best practices emphasize using distributional methods as hypothesis generation tools requiring validation through additional linguistic analysis rather than accepting computational results uncritically.

Neural Machine Translation and Multilingual Corpus Processing

Neural machine translation has revolutionized computational treatment of multilingual corpora, enabling automated translation at quality approaching human performance for many language pairs. Lau (2024) documents methodological innovations in neural machine translation emphasizing cross-linguistic discourse preservation, demonstrating how attention mechanisms and contextual encoding enable translation systems to maintain discourse

coherence and pragmatic appropriateness beyond sentence-level accuracy. These advances facilitate corpus-based contrastive linguistic research by enabling large-scale parallel corpus construction and cross-linguistic pattern analysis.

Multilingual language models trained on diverse language corpora simultaneously learn cross-linguistic representations capturing universal linguistic patterns alongside language-specific characteristics. Models such as mBERT, XLM-R, and mT5 enable zero-shot cross-lingual transfer where models trained on high-resource languages can perform tasks on low-resource languages without additional training. This capability promises democratizing access to natural language processing technologies for under-resourced languages. However, systematic evaluation reveals persistent performance disparities, with low-resource languages consistently receiving inferior model performance despite multilingual training.

Critical examination of multilingual corpus processing reveals concerning patterns of linguistic bias and inequality. Models exhibit systematic performance gaps correlating with language resource availability and speaker population size. High-resource languages receive disproportionate model development attention while thousands of lower-resourced languages remain computationally underserved. Furthermore, language-specific characteristics may be distorted through multilingual training emphasizing cross-linguistic similarities over language-particular features. Addressing these challenges requires intentional investment in diverse language corpus development, culturally appropriate language technologies, and community-engaged language technology design.

Large Language Models in Corpus Construction and Analysis

Large language models represent most recent computational innovation with profound implications for corpus linguistics. Models such as GPT-3, GPT-4, Claude, and others trained on massive text corpora demonstrate remarkable language understanding and generation capabilities. Applications to corpus linguistics include automated corpus construction through web scraping and text extraction, corpus cleaning and preprocessing, preliminary annotation, and even corpus analysis assistance. Kalaš (2025) examines using ChatGPT-4 for corpus linguistic analysis, finding both promising capabilities and significant limitations requiring careful evaluation.

Cheung and Crosthwaite (2025) develop CorpusChat, a system integrating corpus linguistics with generative AI for academic writing development, demonstrating how large language models can provide corpus-informed writing guidance by accessing linguistic patterns in large text collections. Their work shows LLMs can explain corpus patterns, suggest contextually appropriate language use, and provide personalized feedback based on corpus evidence. However, they emphasize that LLM outputs require validation against actual corpus data, as models may hallucinate linguistic patterns or provide incorrect corpus-based claims.

Critical evaluation reveals significant concerns regarding LLM integration in corpus research. Incelli (2025) systematic testing demonstrates that generative AI performs poorly on precise corpus linguistic tasks requiring accurate concordance analysis or systematic pattern identification. LLMs may modify input data, generate false claims about corpus contents, and produce unreliable results across repeated queries due to non-deterministic generation. These limitations raise fundamental questions about LLM appropriateness for rigorous corpus linguistic research requiring precise, replicable findings. Current consensus suggests LLMs may assist certain preliminary corpus tasks but cannot replace careful corpus analysis by trained linguists.

Computational Efficiency and Scalability

Computational corpus linguistics enables analysis at scales impossible with traditional methods. Where manual corpus analysis might feasibly examine thousands of words, computational approaches routinely process billions of words, enabling investigation of rare linguistic phenomena, fine-grained variation patterns, and large-scale language change. Dunn and Adams (2020) construct geographically-balanced gigaword corpora for 50 language varieties, demonstrating computational infrastructure for massive-scale multilingual corpus compilation and processing. Such scale enables statistical robustness and generalizability impossible with smaller corpora.

Computational efficiency improvements through algorithmic optimization and parallel processing enable near-real-time corpus analysis. Modern corpus tools leverage efficient indexing, in-memory processing, and distributed computing to provide interactive corpus queries even on billion-word corpora. Cloud computing platforms democratize access to computational resources previously requiring expensive dedicated infrastructure. However, computational resource requirements raise equity concerns, as researchers in under-resourced institutions or global South contexts may lack access to computational infrastructure necessary for large-scale corpus research, potentially exacerbating existing research inequalities.

Ethical Considerations in Computational Corpus Linguistics

Computational corpus linguistics raises profound ethical questions requiring systematic attention. Privacy concerns emerge when corpora contain personal data from social media, emails, or other sources containing potentially sensitive information. Large-scale corpus collection from web sources may incorporate copyrighted material without appropriate permissions. Algorithmic bias in computational models trained on corpora reproduces and potentially amplifies societal biases related to gender, race, ethnicity, and other social categories. These biases manifest in multiple ways: through training data reflecting historical discrimination patterns, through model architectures favoring dominant language varieties, and through evaluation metrics that undervalue minoritized language communities.

Bird (2022) provides critical examination of speech and language technology through decolonizing lens, documenting how computational linguistics perpetuates colonial power structures through preferential development for colonizer languages, extraction of linguistic data from minority language communities without appropriate consent or benefit sharing, and imposition of Western linguistic frameworks on diverse languages. Computational corpus linguistics risks becoming form of linguistic extractivism where language data from marginalized communities serves primarily to improve technologies benefiting dominant languages and privileged populations.

Establishing ethical computational corpus linguistics requires multiple interventions. Dunn (2022) emphasizes that each computational methodology must pair with discussion of potential ethical implications, making ethics integral to methodological design rather than afterthought. Recommended practices include obtaining informed consent for corpus inclusion, implementing privacy-preserving techniques, conducting bias audits of computational models, engaging language communities in participatory research design, ensuring equitable benefit distribution from language technologies, and acknowledging positionality and power dynamics in research relationships. Institutional changes requiring ethical review of corpus research, funding for diverse language technology development, and promotion of open-source tools and

resources can advance more equitable computational corpus linguistics.

Methodological Challenges and Limitations

Despite computational advances, significant methodological limitations persist. Interpretability challenges mean that many computational models function as black boxes where linguistic understanding of learned patterns remains limited. While models achieve high predictive performance, connecting computational representations to linguistic theory proves difficult. Feature attribution methods and probing studies attempt addressing interpretability, but fundamental tensions remain between computational and linguistic perspectives on language structure. Reproducibility concerns arise from complex computational pipelines involving multiple processing steps, hyperparameter decisions, and stochastic training procedures. Minor implementation variations can produce substantially different results, yet computational corpus linguistics publications often lack sufficient methodological detail enabling exact replication. Establishing reproducibility standards including code sharing, data availability, and comprehensive methodology documentation represents ongoing challenge. Data quality issues affect all corpus research but become particularly acute at computational scale. Automated web corpus construction may incorporate spam, boilerplate text, machine-translated content, and other problematic material without manual quality control. Ensuring corpus representativeness becomes increasingly difficult as corpus size grows, with convenience sampling potentially producing skewed representations of language use. Balancing corpus size benefits against quality concerns requires careful consideration of research goals and appropriate quality assurance procedures.

CONCLUSION

This systematic literature review demonstrates that computational corpus linguistics has emerged as mature interdisciplinary field synthesizing natural language processing, machine learning, artificial intelligence, and corpus linguistic methodologies. The convergence of computational and linguistic approaches over the past decade has transformed corpus linguistics from primarily manual or semi-automated analysis of megaword corpora to automated processing of billion-word collections, enabling linguistic investigations at scales previously unimaginable. This transformation brings both unprecedented analytical capabilities and significant methodological and ethical challenges requiring ongoing critical attention.

Key theoretical contributions include systematic documentation of five major computational innovations revolutionizing corpus linguistic practice: deep learning for automated linguistic annotation, text classification for large-scale register and genre analysis, text similarity modeling enabling cross-linguistic and diachronic comparison, topic modeling and distributional semantics for discovering thematic and semantic patterns, and large language model applications in corpus construction and analysis. These computational methods collectively enable scalability, automation, and pattern discovery capabilities extending traditional corpus linguistic methodologies while simultaneously raising new questions about linguistic validity, interpretability, and research ethics.

Methodologically, the review establishes that successful computational corpus linguistics requires genuine interdisciplinary integration rather than superficial application of computational tools to linguistic questions or computational problem-solving without linguistic grounding. Best practices emerging from reviewed studies emphasize combining computational power with linguistic expertise, validating computational findings against linguistic knowledge and intuition, maintaining transparency about computational methods and

their limitations, and critically evaluating both capabilities and constraints of automated language analysis. The field benefits from hybrid approaches leveraging both symbolic linguistic knowledge and statistical learning from data, avoiding false dichotomy between rule-based and data-driven methods.

Critical examination reveals persistent challenges requiring sustained attention. Algorithmic bias affecting computational models threatens both research validity and social justice, as biased technologies perpetuate discrimination against marginalized language communities. The digital language divide concentrating computational resources on high-resource languages exacerbates global linguistic inequality, potentially accelerating language shift and endangerment for computationally under-resourced languages. Interpretability limitations of black-box models challenge linguistic understanding of learned patterns. Privacy and ethical concerns around large-scale corpus collection and processing require more robust frameworks protecting research participants while enabling valuable linguistic research.

Practically, advancing equitable computational corpus linguistics requires multiple interventions at individual, institutional, and systemic levels. Interdisciplinary training programs should prepare researchers with both linguistic expertise and computational skills, avoiding siloed specialization in either domain alone. Funding agencies should prioritize diverse language corpus development and inclusive language technology research. Open-source tool development can democratize access to computational resources. Community-engaged research approaches should involve language communities as partners rather than data sources. Ethical frameworks must become integral to computational corpus linguistic practice rather than external constraints.

Future research directions identified include developing explainable AI methods providing linguistic interpretability of computational models, enabling researchers to understand what linguistic patterns models capture and how decisions are made. Multimodal computational corpus linguistics integrating textual, visual, audio, and other semiotic resources represents expanding frontier requiring novel methodological development. Decolonial approaches to language technology challenging existing power structures and centering marginalized language communities offer pathways toward more equitable computational linguistics. Low-resource language technologies developed through transfer learning, multilingual models, and community partnerships can extend computational corpus linguistics benefits beyond current high-resource language concentration. In conclusion, computational corpus linguistics represents transformative convergence of computational and linguistic sciences, offering unprecedented capabilities for language research while requiring critical reflexivity about technological limitations, social implications, and ethical responsibilities. Realizing this field full potential demands sustained interdisciplinary collaboration, methodological rigor, ethical commitment, and dedication to linguistic diversity preservation and celebration.

REFERENCES

Biber, D., Reppen, R., & Friginal, E. (2021). The Routledge handbook of corpus linguistics. London: Routledge.

Bird, S. (2022). Decolonising speech and language technology. Computer Speech & Language, 74, 101412.

Cheung, L., & Crosthwaite, P. (2025). CorpusChat: Integrating corpus linguistics and generative AI for academic writing development. Computer Assisted Language Learning, 1-27.

Crosthwaite, P., Baisa, V., & Boulton, A. (2023). Research trends in corpus linguistics: A bibliometric analysis of two decades of Scopus-indexed corpus linguistics research in arts and humanities. International Journal of Corpus Linguistics, 28(3), 344-377.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

Dunn, J. (2022). Natural language processing for corpus linguistics. Cambridge: Cambridge University Press.

Dunn, J., & Adams, B. (2020). Geographically-balanced gigaword corpora for 50 language varieties. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 2521-2529).

Fonteyn, L., Manjavacas, E., & De Regt, J. (2025). Using machine learning to automate data annotation in corpus linguistics: A case study with MacBERTh. International Journal of Corpus Linguistics, 30(3), 296-315.

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

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282-6293).

Jurafsky, D., & Martin, J. H. (2024). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.). Stanford: Pearson.

Kalaš, F. (2025). Bridging tradition and innovation: Analysing language data with ChatGPT-4 in corpus linguistics. Available at SSRN: https://doi.org/10.2139/ssrn.5126316

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2006). The Sketch Engine. In Proceedings of EURALEX (pp. 105-116).

Lau, E. (2024). Advancements in neural machine translation: Methodological innovations and empirical insights for cross-linguistic discourse preservation. International Journal for Research in Applied Science and Engineering Technology, 12(4), 5767-5772.

McEnery, T., & Baker, P. (2016). Corpus linguistics and 17th-century prostitution: Computational linguistics illuminates historical social issues. London: Bloomsbury.

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

McEnery, T., & Hardie, A. (2019). Corpus linguistics: Method, theory and practice (2nd ed.). Cambridge: Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Pérez-Paredes, P., Curry, N., & Aguado Jiménez, P. (2025). Integrating critical corpus and AI literacies in applied linguistics: A mixed-methods study. Computer Assisted Language Learning, 1-27.

Reveilhac, M., & Schneider, G. (2025). Evaluating a transparent and interpretable approach to stance detection using linguistic markers in social media data. International Journal of Corpus Linguistics, 30(2), 195-233.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

Yang, L., Zhou, G., & Lin, L. (2025). From Confucius to computational linguistics: Quantifying cross-linguistic semantic similarity and semantic fidelity using large language models. Digital Scholarship in the Humanities, 40(3), 1021-1032.

Yu, D., Li, L., & Su, H. (2023). Using LLM-assisted annotation for corpus linguistics: A case study of local grammar analysis. arXiv preprint arXiv:2307.00101.