ISSN 1799-2591 (print); ISSN 2053-0692 (online)
Theory and Practice in Language Studies, Vol. 16, No. 2, pp. 586-596, February 2026
DOI: https://doi.org/10.17507/tpls.1602.24

# AI-Assisted Corpus Linguistics: Integrating NLP Models Into Corpus Analysis

Abdullah Saad Al-Qarni*
Royal Commission for Jubail and Yanbu, Jubail English Language and Preparatory Year Institute, Jubail, Saudi Arabia

*Abstract*—**Integrating natural language processing (NLP) and artificial intelligence (AI) models into corpus linguistics has opened new avenues for linguistic analysis, yet their suitability for rigorous academic research remains debated due to issues like opacity and interpretability. This systematic review explores how NLP models transform traditional corpus linguistics methodologies, focusing on their applications, benefits, and challenges. Employing a PRISMA-guided approach, the study reviewed peer-reviewed literature from 2013 to 2025 across databases like Scopus and ACL Anthology, using keywords such as "AI in corpus linguistics" and "NLP corpus analysis". Inclusion criteria targeted studies applying NLP models (e.g., BERT, GPT) to linguistic tasks, resulting in 12 selected studies after screening 922 records. A quality assessment using the CASP checklist ensured robustness, followed by thematic synthesis of findings. Results highlight that NLP models enhance corpus analysis by automating tasks like keyword extraction and pragmatic annotation, while offering scalability and semantic depth. Applications span discourse analysis, diachronic studies, and sociolinguistic variation, supported by tools like CorpusChat and Hugging Face Transformers. However, challenges include model biases, lack of transparency, and domain mismatch. The study explores that AI-driven NLP models significantly advance corpus linguistics but require addressing ethical, privacy, and reproducibility concerns to ensure academic rigor. Future research should focus on developing domain-specific models and enhancing interpretability to fully harness AI's potential in linguistic studies.**

*Index Terms*—**natural processing language, artificial intelligence, corpus analysis, corpus linguistic**

## I. Introduction

The use of AI and NLP in the field of corpus linguistics has sparked discussion, since many linguists wonder whether large language models (LLMs) like BERT and ChatGPT are worthy of thorough linguistic analysis (Raiaan et al., 2024; Suijkerbuijk et al., 2025). Some researchers claim that LLMs are not helpful for serious academic work given the unclear information they use and how their algorithms work, while others are unsure if generative AI has value in finding patterns in language (Cámara et al., 2023). However, it is still worth exploring how NLP models powered by AI can improve corpus linguistics by saving time, emphasizing meaning and supporting the use of various languages (Gatla, 2024; Mohamed et al., 2024; Zhao, 2025). This can be observed in data-driven learning (DDL), where studies have used corpus-based methods to improve teaching and learning for many years, along with the new field of corpus-based language pedagogy (CBLP) (Fauzanz et al., 2022; Liu & Ma, 2025; Lusta et al., 2023). Although AI is changing these fields, little research has been done on how it could affect traditional corpus linguistics practices.

Similar discussions at the 2024 TaLC conference raised questions about the place of traditional corpus software now that AI is available in applications like ChatGPT (Cheung & Crosthwaite, 2025; Granger, 2024). Many academics now wonder if it makes sense to keep promoting DDL and CBL strategies, after people began using AI-driven languages during education and understanding. While there is no conclusive agreement yet, proposing to add AI to corpus linguistics to make it more accessible and functional was thought to help it preserve its position. The situation is urgent because educational and research fields are now adopting AI tools more quickly, leaving behind traditional text analysis tools like AntConc and Sketch Engine (Cheung & Crosthwaite, 2025).

Researchers are now studying the ways AI/NLP models and corpus linguistics can work together. Crosthwaite and Baisa (2023) believe that combining generative AI and DDL results in a "natural partnership" since there are multiple benefits, including: fewer barriers compared to hard corpus query tools (Yu et al., 2024); simpler AI interfaces than most corpus resources (Curry et al., 2024); the strong computational ability of LLMs to analyze very large corpora well beyond what older AntConc can handle (Chen & Chang, 2024); personalized language results suited for different domains like interest or (Piperno et al., 2025) state that these benefits suit the requirements of modern corpus linguistics, especially to deal with vast and diverse data using automation.

Nevertheless, traditional corpus linguistics still has some unique positives. Since corpora are transparent, like the British National Corpus (BNC2014) was, researchers can trust the data as highlighted by Sinclair's (2004) principle (Lin & Adolphs, 2023). Unlike LLMs, corpora give us texts that come from reliable human sources and can be verified which is especially important for studies in linguistics (Alaqlobi et al., 2024). Besides, findings based on corpora are reproducible, but this is not possible for LLMs because they are not always predictable (Li et al., 2024). In educational contexts, where

* Corresponding Author. Email: abdullah.saad91@outlook.com

respecting privacy and copyright rules matters, using traditional corpora is seen as more ethical than doing so with AI-powered tools (bin Subait et al., 2025; Ihekweazu et al., 2024; Mohamed et al., 2024). In addition, using DDL means actively searching for information and thinking about the results, unlike relying passively on AI to provide answers.

Figure 1 depicts practical NLP applications through numerous domains. This figure highpoints how NLP is cohesive into real-world scenarios, including machine translation for bridging language barriers, sentiment analysis for social media monitoring, voice recognition systems for virtual assistants, and chatbots for customer service.


Figure 1. NLP Applications (Wibawa & Kurniawan, 2024)

This conflict brings up a key concern: is it possible to join the thoroughness and transparency of corpus linguistics with the user-friendliness and advanced features of AI/NLP to address the challenges of each approach? This paper answers the question by carefully reviewing how NLP models are used in analyzing corpora, using the latest research as a basis. The following research questions help to organize the review:

*RQ1:* How have NLP models transformed traditional corpus linguistics methodologies?

*RQ2:* What are the primary applications of AI in corpus analysis across linguistic subfields?

*RQ3:* What tools and platforms support AI-assisted corpus linguistics, and how are they different from traditional models?

*RQ4:* What are the benefits and challenges of integrating NLP models into corpus analysis?

- *RQ4 (a):* What are the benefits of integrating NLP models into corpus analysis?
- *RQ4 (b):* What are the challenges of integrating NLP models into corpus analysis?

This paper is structured as follows: Section 2 of this paper sets out the methods, relying on a PRISMA-based systematic review. Section 3 summarizes the findings based on topics like NLP methods, applications, tools, advantages and challenges. Section 4 presents main ideas, notes the existing literature shortcomings, outlines the findings' implications and finishes by offering suggestions for further study. Section 5 concludes the paper.

## II. Methodology

The study follows a systematic approach and the guidelines of PRISMA to ensure transparency, rigor and reproducibility. It includes developing a search strategy, deciding on the inclusion and exclusion criteria, screening the studies, reviewing them using the CASP list, collecting and merging the data and examining the information. Among the available research, the review considered 12 studies that are most relevant to connecting NLP to corpus linguistics.

### *A. Search Strategy*

It included searching for peer-reviewed studies that apply NLP models named BERT, GPT and Word2Vec to AI-assisted corpus linguistics. Important terms used were 'AI in corpus linguistics,' 'NLP corpus analysis,' 'BERT corpus studies,' 'LLM in linguistics,' and 'transformer-based corpus analysis.' Scopus, Web of Science, IEEE Xplore, ScienceDirect and ACL Anthology were used for their wide range of linguistic and computational research. The time frame was between 2013-2025, documenting the progress of NLP models, with a particular emphasis on models created after BERT (after

2018). Applying Boolean operators (AND OR) and wildcards (e.g., 'NLP * corpus') extended the search so that only relevant studies were considered.

### *B. Inclusion/Exclusion Criteria*

The inclusion and exclusion criteria were designed to focus on studies that bridge NLP and corpus linguistics.

TABLE 1
INCLUSION AND EXCLUSION CRITERIA

| Inclusion Criteria | Exclusion Criteria |
|---|---|
| Published in peer-reviewed journals or conference proceedings | Pure NLP or linguistics papers without intersection with corpus analysis |
| Explicitly apply or analyze NLP models (e.g., BERT, GPT, Word2Vec) in corpus linguistics | Non-peer-reviewed sources (e.g., preprints, blogs) |
| Focus on linguistic tasks such as discourse analysis, sentiment analysis, or diachronic studies | Studies lacking empirical application of NLP models |
| Written in English | Non-English publications |

### *C. Screening Process*

The screening process was conducted in three stages to ensure systematic selection of relevant studies. Records identified from, including Web of science (n = 250), Scopus (n = 280), IEEE Xplore (n = 130), ACL Anthology (n = 140), and ScienceDirect (n = 122). Moreover, before screening process, these records removed including duplicate records removed (n = 567), and records removed for other reasons (n = 0). In addition, 355 records were screened. 286 records were removed, which did not qualify according to the criteria. 69 articles were retrieved and 2 were excluded from retrieve. Moreover, 67 articles were assessed for eligibility. Finally, 12 articles meeting criteria and assessment process were included in this review as shown in Figure 2.

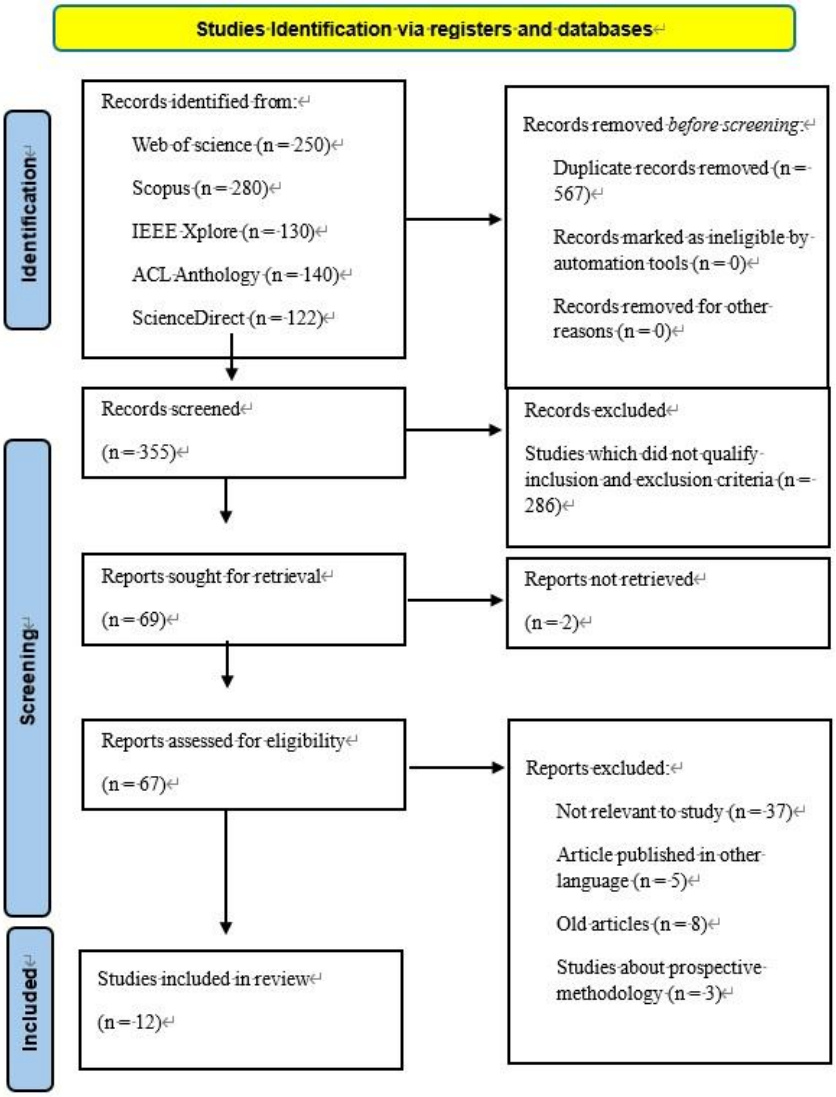


Figure 2. PRISMA Chart for Included Studies

*D. Quality Assessment*

A quality assessment was directed employing the Critical Appraisal Skills Programme (CASP) checklist to ensure the robustness of the selected studies for qualitative and mixed-methods studies. The CASP checklist evaluated studies based on criteria, i.e., clarity of research aims, methodology relevance, data analysis consistency, and ethical considerations. Each study was scored as "High," "Moderate," or "Low" quality based on the fulfillment of CASP criteria. Reasons for the assigned scores were documented to maintain transparency. The results are summarized in the quality assessment Table 2 below.

TABLE 2
INCLUDED STUDIES' QUALITY ASSESSMENT

| Author | Year | Title | CASP Output | Reason |
|---|---|---|---|---|
| (Cheung & Crosthwaite, 2025) | 2025 | CorpusChat: integrating corpus linguistics and generative AI for academic writing development | High | Clear aims, robust mixed-methods approach with surveys and focus groups, ethical considerations addressed, though limited by small sample size |
| (Alaqlobi et al., 2024) | 2024 | Artificial intelligence in applied (linguistics): a content analysis and future prospects | Moderate | Systematic review with statistical analysis (ANOVA, Chi-square) is rigorous, but limited sample diversity reduces generalizability |
| (Alpdemir & Alpdemir, 2024) | 2024 | AI-Assisted Text Composition for Automated Content Authoring Using Transformer-Based Language Models | High | Rigorous hybrid framework with clear methodology, detailed reporting of GPT-4 performance, and comparison to human coders |
| (Yu et al., 2024) | 2024 | Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology | High | Detailed comparison of GPT-3.5, GPT-4, and human coders using local grammar framework, strong empirical focus, though high annotation costs noted |
| (Xu & Huang, 2025) | 2025 | Corpus-based Translation and Interpreting Studies in the Age of AI: Innovations and Challenges | Moderate | The conceptual and analytical approach is clear, but limited discussion of model biases and domain mismatch reduces robustness |
| (Uchida, 2024) | 2024 | Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations | Moderate | Clear evaluation of ChatGPT 3.5 against COCA corpus, but lacks detailed methodological transparency and struggles with genre identification |
| (Crosthwaite & Baisa, 2023) | 2023 | Generative AI and the end of corpus-assisted data-driven learning? Not so fast! | High | Comprehensive comparative analysis of corpora and generative AI, addresses ethical issues and transparency, robust theoretical grounding |
| (Curry et al., 2024) | 2024 | Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT | Moderate | Replication case studies are well-designed, but non-reproducible outputs and limited suitability for nuanced tasks lower quality |

| | | | | |
|---|---|---|---|---|
| (Flowerdew, 2024) | 2024 | Data-driven learning: From Collins Cobuild Dictionary to ChatGPT | High | Clear pedagogical focus with case studies, robust evaluation of ChatGPT in DDL, though limited to general corpora |
| (Chen & Chang, 2024) | 2024 | An entropy-based corpus method for improving keyword extraction: An example of sustainability corpus | High | Rigorous quantitative entropy-based method, clear reporting of keyword extraction improvements, though computationally intensive |
| (bin Subait et al., 2025) | 2025 | Artificial Intelligence-based Natural Language Processing for sarcasm detection and classification on Arabic Corpus | Moderate | Strong results with AINLP-ALBTCN achieving high accuracy, but limited to Arabic corpora, reducing applicability |
| (Mo & Crosthwaite, 2025) | 2025 | Exploring the advantages of generative AI large language models for stance and engagement in academic writing | High | Robust methodology comparing LLMs and human essays using Hyland's taxonomy, addresses biases and plagiarism detection challenges |

### *E. Data Extraction and Synthesis*

Data extraction was systematically performed on the 12 selected studies to capture key information relevant to the research questions. Extracted data included metadata (authors, year, publication venue), NLP models used (e.g., BERT, GPT, spaCy), corpus details (type and size), linguistic tasks (such as sentiment analysis, or Named Entity Recognition), and key findings (e.g., performance metrics, qualitative insights). The extraction process followed a standardized template to ensure consistency. Synthesis was conducted using a thematic approach, organizing findings into categories: NLP techniques, applications, tools, benefits, challenges, and emerging trends. Narrative synthesis complemented this by providing qualitative insights into how AI enhances corpus linguistics, drawing on the studies' discussions of methodological advancements and limitations.

TABLE 3
SELECTED STUDIES

| Author | Year | Title | Methods | Findings | Limitations |
|---|---|---|---|---|---|
| (Cheung & Crosthwaite, 2025) | 2025 | CorpusChat: integrating corpus linguistics and generative AI for academic writing development | Developed and piloted CorpusChat, an AI-driven platform, with two chatbots built using the BAWE and CHAT corpora, tested on undergraduate arts students in Hong Kong through surveys and focus group interview. | Broad spectrum of uses and views, highlighting both AI potential in pedagogy and research, and concerns about ethics and content quality. | Limited to academic writing; small sample size |
| (Alaqlobi et al., 2024) | 2024 | Artificial intelligence in applied (linguistics): a content analysis and future prospects. | A systematic review and content analysis of 73 scholarly articles. Used statistical tools such as Chi-square and one-way ANOVA. | Produced high-quality, human-like content from limited inputs, offering a practical solution for resource-constrained text generation tasks. | Lack of domain-specific models |
| (Alpdemir & Alpdemir, 2024) | 2024 | AI-Assisted Text Composition for Automated Content Authoring Using Transformer-Based Language Models | A hybrid framework combining Controllable Text Generation (CTG) via Large Language Models, fine-tuned models, and sentence transformers to generate Turkish-language articles in the style of specific real authors. | GPT-4 outperformed GPT-3.5 and achieved accuracy close to that of a human coder. | Limited corpus diversity |

| | | | | | |
|---|---|---|---|---|---|
| (Yu et al., 2024) | 2024 | Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology | Involved comparing GPT-3.5, GPT-4, and a human coder in annotating apology components in English texts using the local grammar framework, to evaluate the potential of LLMs for pragma-discursive corpus annotation | LLMs improve annotation speed and accuracy | High annotation costs |
| (Xu & Huang, 2025) | 2025 | Corpus-based Translation and Interpreting Studies in the Age of AI: Innovations and Challenges | A conceptual and analytical approach to examine how AI—particularly generative AI and machine learning—can enhance corpus-based translation and interpreting studies (CTIS) by automating tasks like data collection, corpus compilation, annotation, and feature extraction. | AI's transformative potential for CTIS, while emphasizing the need for critical human oversight and collaborative, balanced approaches to ensure ethical and effective use. | Domain mismatch issues |
| (Uchida, 2024) | 2024 | Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations | Involved evaluating ChatGPT 3.5's performance on various corpus linguistic tasks—such as generating frequency lists, collocations, grammatical patterns, and genre identification—by comparing its outputs with results from the COCA corpus. The study assessed the level of congruence between the LLM's outputs and COCA data for the top 20 items in each category. | Strong alignment in frequency lists (75.0%), moderate success in collocations (42.8%) and grammatical patterns (53.0%), but poor performance in genre identification. | While early LLMs may lack academic rigor. |
| (Crosthwaite & Baisa, 2023) | 2023 | Generative AI and the end of corpus-assisted data-driven learning? Not so fast! | Comparative analytical approach to examine the strengths and limitations of both corpora and generative AI (GenAI) in language learning and analysis | Highlighted the reliability, transparency, and depth of corpus-based methods, particularly in analyzing multi-word units and known data sources, while acknowledging GenAI's ability to address some limitations of traditional DDL. | Over-reliance on pre-trained models |
| (Curry et al., 2024) | 2024 | Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT | Employed three replication case studies to evaluate ChatGPT's applicability for automated qualitative analysis in corpus-based discourse studies, focusing on semantic categorization, concordance analysis, and function-to-form analysis. Each task tested ChatGPT's performance on typical discourse-analytic procedures to assess its effectiveness and limitations | ChatGPT showed moderate success in semantic categorization but performed poorly in concordance and function-to-form analysis, raising concerns about its reliability, replicability, and suitability for nuanced, empirical linguistic research. | Non-reproducible outputs |
| (Flowerdew, 2024) | 2024 | Data-driven learning: From Collins Cobuild Dictionary to ChatGPT | The study used a few case studies, including the Collins Cobuild Dictionary, an academic writing workshop, and the replication of workshop tasks using ChatGPT, to explore developments in data-driven learning (DDL). | Evolution of DDL and suggests that large language models like ChatGPT present both transformative opportunities and | Limited to general corpora |

| | | | | | |
|---|---|---|---|---|---|
| | | | | significant challenges for its future. | |
| (Chen & Chang, 2024) | 2024 | An entropy-based corpus method for improving keyword extraction: An example of sustainability corpus. | An entropy-based corpus method that optimizes keyword extraction by automatically excluding function and generic words, assigns objective weights to keyword parameters (log-likelihood, frequency, and range) | Entropy method improves keyword extraction | Computationally intensive |
| (bin Subait et al., 2025) | 2025 | Artificial Intelligence-based Natural Language Processing for sarcasm detection and classification on Arabic Corpus. | Introduced the AINLP-ALBTCN technique for Arabic sarcasm detection, combining data preprocessing, word2vec embeddings, Bidirectional Temporal Convolutional Networks (BTCN) for classification, and Sand Cat Swarm Optimization (SCSO) for hyperparameter tuning | Method achieved a superior sarcasm classification accuracy of 95.59%, outperforming existing models. | Limited to Arabic corpora |
| (Mo & Crosthwaite, 2025) | 2025 | Exploring the affordances of generative AI large language models for stance and engagement in academic writing. | Compared 30 academic essays from three LLMs with human-written essays on the same topics, annotating texts for stance and engagement using Hyland's (2005) taxonomy to analyze linguistic properties and strategies | LLMs exhibited a narrower and more repetitive use of stance and engagement features compared to humans, with disciplinary patterns mostly aligned except in philosophy. Highlighted challenges in detecting LLM-based plagiarism and inconsistencies in academic writing instruction. | Bias in pre-trained models |

## III. Findings

This section synthesizes findings from the 12 selected studies, organized thematically to address NLP techniques, applications, tools, benefits, challenges, and emerging trends in AI-assisted corpus linguistics. Critical analysis of each study's methodology, findings, limitations, and contributions is integrated into the relevant subsections to provide a comprehensive evaluation.

### *A. Traditional Corpus Linguistics Methodologies*

*RQ1:* How have NLP models transformed traditional corpus linguistics methodologies?

Using NLP has greatly improved the analysis of corpora by helping to go beyond standard practices such as Part-of-Speech (POS) tagging and parsing. Alaqlobi et al. (2024) explored and analyzed 73 articles by using ANOVA and Chi-square methods and they reported that BERT and GPT models reduce time and effort in linguistic analysis by creating very good, human-like content from limited information. Nonetheless, because they rely on models that work for many tasks, they are not always well-suited for specific areas, requiring specialized changes. In 2024, Chen and Chang suggested using an information measure, working on parameter weights like log-likelihood and frequency, to get high accuracy for keywords related to sustainability. Despite its accuracy, the complex calculation needed in this method makes it difficult to implement, which indicates it needs to be improved. Moreover, Alpdemir and Alpdemir (2024) proposed a hybrid framework combining Controllable Text Generation and fine-tuned transformer models, demonstrating GPT-4's near-human accuracy in generating Turkish-language articles. The limited corpus diversity in their study raises concerns about generalizability, but the approach underscores transformers' scalability. Likewise, Uchida (2024) evaluated ChatGPT 3.5 against the COCA corpus, finding strong alignment in frequency lists (75%) but poor performance in genre identification due to opaque outputs, emphasizing the need for greater transparency in LLMs. These studies collectively highlight the power of embedding models (e.g., Word2Vec, BERT) and LLMs in enhancing semantic analysis and automation, though methodological transparency and computational efficiency remain critical challenges.

### *B. Applications in Corpus Analysis*

*RQ2:* What are the primary applications of AI in corpus analysis across linguistic subfields?

AI-assisted corpus linguistics leverages NLP models to address a wide range of linguistic phenomena, demonstrating versatility across diverse applications. In discourse analysis, Curry et al. (2024) conducted replication case studies to evaluate ChatGPT's performance in semantic categorization, concordance, and function-to-form analysis, achieving moderate success in identifying discourse markers but struggling with nuanced tasks due to non-reproducible outputs, which limits its reliability for empirical discourse studies and underscores the need for enhanced reproducibility mechanisms to ensure robust results. For diachronic studies, Xu and Huang (2025) adopted a conceptual approach to explore AI's role in corpus-based translation and interpreting studies (CTIS), finding that AI streamlines data collection and feature extraction, significantly enhancing the analysis of linguistic evolution over time; however, their study's failure to address model biases and domain mismatch issues weakens its robustness, highlighting the need for empirical validation and human oversight to ensure accuracy in specialized domains. In sociolinguistic variation, bin Subait et al. (2025) introduced the AINLP-ALBTCN technique, which combines word2vec embeddings with Bidirectional Temporal Convolutional Networks, achieving an impressive 95.59% accuracy in sarcasm detection within Arabic corpora, though its focus on a single language limits broader applicability, yet the innovative methodology showcases AI's potential for sociolinguistic tasks. In sentiment and stance analysis, Mo and Crosthwaite (2025) compared LLM-generated and human-written academic essays using Hyland's (2005) taxonomy, revealing that LLMs exhibit repetitive stance and engagement features, with biases in pre-trained models posing challenges; their robust methodology strengthens the study's credibility, but addressing these biases is critical for reliable academic writing analysis. In pragmatic annotation, Yu et al. (2024) found that GPT-3.5 and GPT-4 are faster and more accurate than humans, though high annotation costs limit scalability, so affordable strategies are still necessary. They demonstrate that NLP models have the power to revolutionize corpus linguistics, though they still require efforts to resolve problems related to reproducibility, bias and costs.

### *C. Tools and Platforms*

*RQ3:* What tools and platforms support AI-assisted corpus linguistics, and how are they different from traditional models?

Traditionally, linguistic patterns were investigated through tools like AntConc and Sketch Engine which depended on handcrafted queries and statistics. Nowadays, AI platforms are emerging that lessen the need for tedious queries and make these tasks easier for everyone. Cheung and Crosthwaite (2025) introduced CorpusChat, a platform that uses BAWE and CHAT corpora and was studied with Hong Kong undergraduate arts students through surveys and interviews. Unlike AntConc that relies on users being skilled in its programming, CorpusChat showcases the use of AI, highlighting its prospects for improving teaching, yet its narrow reach and focus on academics require it to be tried in other areas to confirm if it can work as reliably as traditional tools. Flowerdew (2024) compared ChatGPT to dictionary-based tools such as Collins Cobuild and Sketch Engine and found that ChatGPT excels by providing clear explanations for downloading corpora and correcting tasks, features lacking in other tools due to their reliance on old information. While ChatGPT counts on common information sources, other tools use specialized and verifiable databases, resulting in more advanced educational features. According to Alpdemir and Alpdemir (2024), using Python libraries such as spaCy and Hugging Face Transformers for custom pipelines helped achieve accurate text composition, exceeding what AntConc can do alone, though the lack of corpus variety in their study highlights the importance of having more widely representative datasets for similar insights that traditional tools provide. The results of these studies reveal that AI platforms are easy to use and scalable, but they lack transparency and precision that similar traditional approaches have which can be addressed through general use and rigorous testing.

### *D. Benefits of NLP Integration*

*RQ4 (a): What are the benefits of integrating NLP models into corpus analysis?*

Introducing NLP systems into corpus linguistics enables performance improvements in language analysis. Researchers Chen and Chang (2024) and Yu et al. (2024) have shown that using automation such as entropy-based keyword extraction and LLMs helps reduce the manual work needed for complex tasks. These models have an additional benefit, as Alpdemir and Alpdemir (2024) demonstrated that transformer-based ones such as GPT-4 are efficient for generating Turkish text in a large amount of data. According to Alaqlobi et al. (2024), NLP models offer superior contextual understanding, revealing fine linguistic details that earlier approaches could not detect.

### *E. Challenges and Limitations*

*RQ4 (b): What are the challenges of integrating NLP models into corpus analysis?*

There are several challenges in corpus linguistics that influence how well NLP models can be applied and trusted in academic and practical situations. According to Mo and Crosthwaite (2025), having biased models is a major obstacle, since the unreliability of their findings calls for strong and effective strategies to address these biases in academic research. Uchida (2024) and Curry et al. (2024) also point out that models like ChatGPT are not interpretable, as their outputs cannot be replicated or verified by others which is a major concern for the academic community. Yu et al. (2024) highlight the challenge presented by limited well-built annotated corpora for NLP, pointing out that accessible, well-curated annotations are essential for advanced field development since they help in effectively training the models. Domain mismatch is also an issue and according to Abdelaal (2023), Lin et al. (2025), Sobti et al. (2024), and Veres (2022), general

models tend to perform poorly in areas such as translation and applied linguistics, needing to be specially adjusted for better outcomes.

## IV. Discussion

By embracing AI and NLP, corpus linguistics have changed the way language data is handled and applied in numerous settings. This discussion analyzes the key findings, any gaps in literature and what these discoveries mean for both future research and practical implementation and it also looks at the potential wider impact of AI-assisted corpus linguistics on the field. The tools and methods in corpus linguistics have been enhanced by using NLP, as machine processing has reduced manual work, offered in-depth analysis and made these tools easier for everyone to use, apart from those that mainly work with statistical instruments (Gatla, 2024; Zhao, 2025). Similarly, GPT-4 and other LLMs make annotation faster and more accurate, reducing the time spent by humans, but issues with high cost may make it tough to use them in settings without much funding. In addition, the research demonstrates that AI is flexible and well-developed, though it reveals that sometimes it is more economical to use manual approaches because costly advanced AI tools are not always the best choice.

### *A. Gaps Identified*

Despite these improvements, there are still major areas lacking in literature which weakens AI-supported corpus linguistics. Since ChatGPT 3.5 lacked clarity in its production and did not meet academic standards in identifying genre, this was a repeated issue of transparency. Because many studies do not give detailed information on how models are configured, it becomes difficult to check their findings and trust the results of AI. Additionally, ChatGPT does not allow researchers to replicate its output in tasks such as concordance and analyzing how word forms relate to grammatical functions. This matter is important as difficulties from black-box models can influence stance calculations and the accuracy of insights offered by AI in academic papers. Relating to corpus translation, depending only on pre-trained models ends up resulting in performance gaps, since they cannot handle specialized domains well and thus alternative models should be designed. It seems that, despite the tools that AI offers, its integration in corpus linguistics usually lacks the accuracy and flexibility needed for all kinds of linguistic studies and this problem could be addressed by making the procedures clear for others to follow.

### *B. Implications of Findings*

AI makes it possible for linguists to use corpora easily, thus revealing linguistic patterns more swiftly. Computational linguists have advantages using their own NLP systems for composing texts, but the narrow variety in their corpuses points out the need for more representative data to make them capable of dealing with more situations. CorpusChat uses AI to help teach writing in school, but a modest number of student participants and its use in academics mean the tool's benefits need to be tested more broadly. Digital humanities make use of NLP to process vast volumes of text, but still face challenges in diachronic studies, so it's up to humans to prevent some biases and errors. They point to the role AI plays in combining classic and computer-based approaches, but they also lead to questions about ethics, privacy and copyright. The data files LLMs learn from which are often just from the internet increase the risk of bias, unchecked facts and privacy issues if private data appears in them. Since proprietary texts are under copyright, it is challenging for embracing tools, for instance CorpusChat, to be introduced in schools. Clear policy actions are needed: strong rules and guidance on ethics should be formed to overcome these challenges in AI use for research and education.

### *C. Future Directions*

The findings also point to future research directions that could address these challenges and further advance AI-assisted corpus linguistics. To overcome domain mismatch in disciplines such as translation or sociolinguistics, it is necessary to build NLP models designed for each field, since general models usually miss the special details of those areas. Explainable AI which allows one to easily understand models, helps solve the black-box problems and encourages belief in modeled predictions. Developing open access corpora with comments would deal with the lack of quality annotations, encouraging more complete model training and a wider variety of research work. Moreover, considering human-in-the-loop techniques could improve interactive corpus analysis by using AI to analyze huge amounts of data and having people refine and avoid errors. Additionally, these directions should protect privacy and ethics by using clear data sources and obeying copyright guidelines in educational settings, as these are especially important for teaching.

## V. Conclusion

The review highlights that AI and NLP models greatly improve corpus linguistics, making automation, scaling and exploring semantics easier in different applications. CorpusChat and Hugging Face Transformers are examples of how corpus analysis can be made both simpler and quicker. It also points out that AI can handle many challenges in language analysis. But problems such as biases in ML models, difficulty explaining their results, mismatch with certain applications and privacy and copyright issues must be solved to sustain the field. With more focus on models designed for single domains, easy-to-interpret results, shared data and roles for humans, research in the future can narrow these gaps, promoting an inclusive, robust and ethical approach that advances learning in language and humanities.

## REFERENCES


[1] Abdelaal, N. (2023). The role of corpora in enhancing translation accuracy and fluency feasibility of using corpora as a tool in translation practice. *Australian Journal of Applied Linguistics*, *6*(3), 205-218.

[2] Alaqlobi, O., Alduais, A., Qasem, F., & Alasmari, M. (2024). Artificial intelligence in applied (linguistics): a content analysis and future prospects. *Cogent Arts & Humanities*, *11*(1), 2382422.

[3] Alpdemir, Y., & Alpdemir, M. N. (2024). *AI-Assisted Text Composition for Automated Content Authoring Using Transformer-Based Language Models.* 2024 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET), 1-6.

[4] bin Subait, W., Asiri, M. M., Alzaidi, M. S. A., Alanazi, M. H., Alshammeri, M., Yafoz, A., Alsini, R., & Khadidos, A. O. (2025). Artificial Intelligence-based Natural Language Processing for sarcasm detection and classification on Arabic Corpus. *Alexandria Engineering Journal*, *125*, 320-331.

[5] Cámara, J., Troya, J., Burgueño, L., & Vallecillo, A. (2023). On the assessment of generative AI in modeling tasks: an experience report with ChatGPT and UML. *Software and Systems Modeling*, *22*(3), 781-793.

[6] Chen, L.-C., & Chang, K.-H. (2024). An entropy-based corpus method for improving keyword extraction: An example of sustainability corpus. *Engineering Applications of Artificial Intelligence*, *133*, 108049.

[7] Cheung, L., & Crosthwaite, P. (2025). CorpusChat: integrating corpus linguistics and generative AI for academic writing development. *Computer Assisted Language Learning*, *133*(Part B), 1-27.

[8] Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, *3*(3), 100066.

[9] Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), 100082.

[10] Fauzanz, A., Basthomi, Y., & Ivone, F. M. (2022). Effects of using online corpus and online dictionary as data-driven learning on students' grammar mastery. *LEARN Journal: Language Education and Acquisition Research Network*, *15*(2), 679-704.

[11] Flowerdew, J. (2024). Data-driven learning: From Collins Cobuild Dictionary to ChatGPT. *Language Teaching*, *58*(3), 1-18.

[12] Gatla, T. R. (2024). A groundbreaking research in breaking language barriers: NLP and linguistics development. *International Journal of Advanced Research and Interdisciplinary Scientific Endeavours*, *1*(1), 1-7.

[13] Granger, S. (2024). From early to future learner corpus research. *International Journal of Learner Corpus Research*, *10*(2), 247-279.

[14] Ihekweazu, C., Zhou, B., & Adelowo, E. A. (2024). Ethics-Driven Education: Integrating AI Responsibly for Academic Excellence. *Information Systems Education Journal*, *22*(3), 36-46.

[15] Li, Y., Huo, Y., Jiang, Z., Zhong, R., He, P., Su, Y., Briand, L. C., & Lyu, M. R. (2024). Exploring the effectiveness of llms in automated logging statement generation: An empirical study. *IEEE Transactions on Software Engineering*, *50*(12), 3188-3207.

[16] Lin, P., & Adolphs, S. (2023). Corpus linguistics. In *The Routledge Handbook of Applied Linguistics* (pp. 296-308). Routledge.

[17] Lin, Y., Wang, R., & Chu, C. (2025). Addressing Domain Mismatch in Unsupervised Neural Machine Translation. *IEEE Transactions on Audio, Speech and Language Processing*, *33*, 472-482.

[18] Liu, J., & Ma, Q. (2025). Examining corpus-based language pedagogy (CBLP) practices in data-driven learning (DDL) for low-proficiency L2 English learners: A meta-analysis. *Educational Technology & Society*, *28*(2), 53.

[19] Lusta, A., Demirel, Ö., & Mohammadzadeh, B. (2023). Language corpus and data driven learning (DDL) in language classrooms: A systematic review. *Heliyon*, *9*(12), e22731.

[20] Mo, Z., & Crosthwaite, P. (2025). Exploring the affordances of generative AI large language models for stance and engagement in academic writing. *Journal of English for Academic Purposes*, *75*, 101499.

[21] Mohamed, Y. A., Khanan, A., Bashir, M., Mohamed, A. H. H., Adiel, M. A., & Elsadig, M. A. (2024). The impact of artificial intelligence on language translation: a review. *IEEE Access*, *12*, 25553-25579.

[22] Mohamed, Y. A., Mohamed, A. H., Kannan, A., Bashir, M., Adiel, M. A., & Elsadig, M. A. (2024). Navigating the Ethical Terrain of AI-Generated Text Tools: A Review. *IEEE Access*, *12*, 197061-197120.

[23] Piperno, R., Bacco, L., Dell'Orletta, F., Merone, M., & Pecchia, L. (2025). Cross-lingual distillation for domain knowledge transfer with sentence transformers. *Knowledge-Based Systems*, *311*, 113079.

[24] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, *12*, 26839-26874.

[25] Sobti, R., Guleria, K., & Kadyan, V. (2024). Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. *Multimedia Tools and Applications*, *83*(35), 81933-81995.

[26] Suijkerbuijk, M., Prins, Z., de Heer Kloots, M., Zuidema, W., & Frank, S. L. (2025). BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 1-39.

[27] Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics*, *4*(1), 100089.

[28] Veres, C. (2022). Large language models are not models of natural language: they are corpus models. *IEEE Access*, *10*, 61970-61979.

[29] Wibawa, A. P., & Kurniawan, F. (2024). Advancements in natural language processing: Implications, challenges, and future directions. *Telematics and Informatics Reports*, *16*, 100173.

[30] Xu, H., & Huang, Y. (2025). Corpus-based Translation and Interpreting Studies in the Age of AI: Innovations and Challenges. In *Translation Studies in the Age of Artificial Intelligence* (pp. 85-99). Routledge.

[31] Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, *29*(4), 534-561.

[32] Zhao, D. (2025). The impact of AI-enhanced natural language processing tools on writing proficiency: An analysis of language precision, content summarization, and creative writing facilitation. *Education and Information Technologies*, *30*(6), 8055-8086.

**Abdullah Saad Al-Qarni** works at Royal Commission for Jubail & Yanbu, Jubail English Language and Preparatory Year Institute (JELPYI) Saudi Arabia. Dr. Abdullah Al-Qarni is the Deputy of Planning and Development at the Jubail English Language and Preparatory Year Institute (JELPYI), Royal Commission for Jubail & Yanbu, Kingdom of Saudi Arabia. He holds a PhD in Applied Linguistics, where his doctoral research employed corpus linguistics methods to analyze learner corpora and to evaluate the effectiveness of Corpus Pattern Analysis (CPA) developed by Patrick Hanks. His current research interests include corpus linguistics, learner corpora, and the integration of artificial intelligence with corpus linguistics methodologies to advance language research and pedagogy. His work contributes to the development of innovative, data-driven, and technology-enhanced approaches to English language teaching and learning.