Exploring the Future of Corpus Linguistics: Innovations in AI and Social Impact

Ersilia Incelli*

Sapienza University of Rome, Italy

Abstract: This paper explores the evolving landscape of corpus linguistics, focusing on the impact of artificial intelligence (AI) and its social implications. Over the past two decades, the study of language through corpus linguistics has evolved significantly, prompting ongoing reflection on the field's transformation. These reflections naturally give rise to pressing questions related to how corpus linguistics will evolve in a world defined by rapid technological progress and changing societal priorities. To validate the suppositions and reflections addressed in this contribution, the study explores a corpus that comprises scholarly papers from scientific journals, and a collection of AI-related articles taken from the media. This dual corpus enables a comparative analysis of how AI-driven corpus linguistics is represented, in order to explore how the integration of artificial intelligence is transforming corpus linguistics, and hence the methodological, theoretical, and socio-political implications of this shift. The methodological framework combines quantitative corpus analysis with qualitative discourse analysis. Collocation and keyword frequency retrieval is applied to identify prevalent themes. As expected academic literature emphasizes methodological advancements and data-driven rigor, while media discourse highlights ethical concerns and societal implications. These findings support the overview and contribute to understanding how AI is shaping both the practice and perception of corpus linguistics in contemporary society.

Keywords: AI-driven corpus linguistics, AI-generated text, AI integration, collocations, language processing, social implications.

1. INTRODUCTION

Having devoted over two decades to the study of language through the lens of corpus linguistics, I often find myself reflecting on the transformations this field has undergone. These reflections inevitably lead to a forward-looking question: What lies ahead for corpus linguistics in a world increasingly shaped by rapid technological advances and shifting societal needs? As researchers, it is our nature to anticipate and adapt, driven by curiosity and a commitment to advancing our discipline. With these thoughts in mind this current contribution formulates a pressing research question: how is the integration of artificial intelligence influencing methodological approaches and transforming corpus linguistics, and what are its broader social and linguistic implications? By addressing the above research question, it may be possible to determine where corpus analysis is heading in this specific context.

Thus, the study seeks to investigate how AI is shaping corpus linguistics methodologies and what implications this has for the future of language analysis. The paper examines how AI-driven approaches align with or diverge from traditional corpus linguistics methodologies. Additionally, insights from AI-driven linguistic studies are incorporated to address concerns regarding AI bias, representation, social impact and ethical considerations in corpus analysis. Indeed, sociolinguistic perspectives provide a broader context for understanding how AI-driven corpus linguistics affects discourse and influences linguistic change. In terms of methodology, the paper outlines an empirical approach to analyzing AI's role in corpus linguistics. This includes: surveying corpus linguistics researchers to understand their current use of AI-driven tools and their perspectives on methodological shifts; looking at a newspaper and online media articles and texts which focus on the social implications. These methodological steps provide evidence supporting the reflections made here on AI’s role in corpus linguistics rather than relying on abstract observations in order to move beyond a descriptive discussion.

To explore and develop the observations which make up the various sections, a corpus was created consisting of newspaper and online media articles dealing with AI-driven related issues especially with reference to language, and a second sub-corpus of scientific research articles was collected on the subject of AI integrated language analysis involving technical processes like natural language processing and machine learning. The two sub-corpora together totalled 110,856 words. Although it is a small corpus, it can be considered nevertheless representative of idiosyncratic features which pointed to the future of AI-driven corpus linguistics. The corpus generated reoccurring language features and common collocations recurrent through both sub-corpora which formed the basis for the reflections on AI-driven corpus linguistics.

*Address correspondence to this author at the Sapienza University of Rome, Italy; E-mail: ersilia.incelli@uniroma1.it

Nonetheless, it is necessary to point out that this study does not primarily focus on corpus analysis or corpus linguistic retrieval techniques. Such an approach is not the primary objective of the reflections reported here, although the considerations make use of the small collected corpus so as to identify the most frequently occurring and semantically significant lexical items in AI-driven corpus linguistics documents. The analysis of these key terms led to the formulation of forward-looking reflections on the conceptual and technical dimensions of AI's role in corpus linguistics. In other words, rather than conducting an extensive corpus study, this contribution highlights selected lexical units and core AI and NLP terms that emerged from the corpus, illustrating the strong emphasis on both the technical mechanisms and broader conceptual implications of AI within the field**.** For this purpose, the study is not intended to be a mere juxtaposition of observations, but rather an analysis informed by some empirical findings. While it does not adopt a traditional corpus-based linguistic approach, it draws on a targeted small-scale corpus to identify key lexical patterns and terminological trends within AI-driven corpus linguistics discourse. These lexical patterns serve as a basis for broader reflections on the conceptual and methodological shifts occurring in the field. Furthermore, in addressing the core question of where corpus linguistics is heading, this study emphasizes that AI does not eliminate the corpus itself, but reshapes how it is analyzed and interpreted. While AI-driven tools offer new efficiencies and capabilities, corpus linguistics remains fundamentally reliant on language data. This discussion is substantiated with some empirical findings and existing research, ensuring that conclusions are not based merely on speculation, but on an informed and data-driven analysis.

The study continues as follows: section 2 presents a brief review of the beginnings in corpus linguistics, followed by a description of lexical units and the emergence of terms in the literature which reflect the advent and integration of AI technologies, chartering what this could mean for corpus linguistics research analysis. Through corpus retrieval of key collocations, section 4 discusses the social impact and relevance of AI-driven studies in corpus linguistics, pointing to how corpus linguistics can evolve in both the AI landscape and its social context, identifying challenges, opportunities, and the ethical considerations that accompany these advances.

2. HISTORICAL PERSPECTIVES

AI's integration into corpus linguistics is not an entirely new phenomenon, and this paper incorporates a historical review of AI's influence on corpus studies over the past few years. By tracing the evolution of AI in this field, the study will ensure that its claims are contextualized within ongoing developments rather than presenting AI's role as a novel or speculative future trend.

When corpus linguistics first emerged as a field, it represented an innovative, data-driven approach to studying language. Defined as the analysis of language through large text collections (corpora), it provided researchers with a foundation for examining patterns and structures beyond theoretical models by relying on empirically gathered data, as opposed to theoretical models (Sinclair, 1991; Biber, 1993; Stubbs, 2001). The scope of analysis was often constrained by the significant effort required to process and examine large datasets. Early efforts employed manually compiled datasets and basic tools; nonetheless, the potential of corpus linguistics to uncover insights about language use was apparent even at an early date. Key figures like John Sinclair (1991) and Michael Stubbs (2001) emphasized the capacity of corpora to illuminate lexical semantics and patterns of meaning. Their pioneering work laid the groundwork for advancements that would expand the field's scope dramatically. In fact, over time, corpus linguistics transitioned from small-scale, manually compiled corpora and manual analyses to leveraging vast, machine-readable datasets. This transformation was driven by technological innovations, making corpus linguistics increasingly accessible and versatile. Its evolution also began to reflect the growing recognition of its interdisciplinary potential.

Corpus linguistics has traditionally relied on human expertise in the design and compilation of corpora. Early examples, such as the Brown Corpus (Francis and Kuçera, 1964), were created by gathering texts from various genres to represent the language as it was used in a specific time period. As computational power increased and software tools became more sophisticated, corpus linguistics underwent a major shift. The availability of larger corpora and the rise of computational methods allowed for the automated extraction of linguistic patterns, such as word frequency distributions, collocations, and syntactic structures (Biber *et al.*, 1998). As we move further into the 21st century, the future of corpus linguistics is likely to be shaped by two key forces: the evolution of artificial intelligence (AI) and the increasing emphasis on the social dimensions of language (Gu, 2023; Bleorţu, 2024; McEnery and Brooke, 2024). This contribution reflects on these two significant factors and their impact on ongoing evolutions in the field.

In recent years, the integration of artificial intelligence has introduced a wave of innovation that is transforming how we approach corpus linguistics and education. Techniques like natural language processing and machine learning are no longer fringe tools, they are central to methodology for many scholars (Gruetzemacher, 2022). These advances inspire critical reflection on the direction the field is taking. And with these reflections come questions such as, are these tools being leveraged to their fullest potential, or is there a risk of becoming overly reliant on technology at the expense of deeper linguistic insight? (McEnery and Hardie, 2012).

3. KEY COLLOCATIONS DEMONSTRATING AI'S ROLE IN SHAPING CORPUS LINGUISTICS: AN AI-POWERED FUTURE

An important component of the following considerations in this section involves reviewing existing literature and ongoing research projects that integrate AI into corpus analysis, providing an empirical basis for assessing current trends and future directions. The question of AI's impact on the trajectory of corpus linguistics is indeed fundamental. Rather than suggesting that AI will erase the social dimension of linguistic analysis, this study argues that AI-driven corpus methods are reshaping how linguistic data is processed, analyzed, and interpreted. AI's role does not eliminate qualitative or human-centred approaches, but rather extends the range of analytical possibilities (Jurafsky and Martin, 2024). At the same time, there is a risk that an overreliance on computational modelling could underestimate critical sociolinguistic perspectives, and this tension must be acknowledged.

The following list of some retrieved key collocations and their contextual surroundings provide empirical support for subsequent considerations, demonstrating how AI is actively shaping the field of corpus linguistics. The collocations are followed by sample sentences from the scientific literature illustrating the collocations in context. A keyword analysis (using Sketchengine, Kilkarif *et al*., 2006) identifies the most frequently occurring and semantically significant words.

Core linguistic terms: *Corpus, annotation, pragmatics, syntax, semantics, discourse, tagging, transcription;*

AI and Natural Language Processing (NLP) terms: *Machine learning, deep learning, transformer models, neural networks, tokenization, embeddings, AI-generated text, AI systems, AI models; Application-oriented terms**:** annotation, language processing, text generation, sentiment analysis, translation, computational linguistics, automation, efficiency;*

Ethical terms: *Bias, fairness, transparency, data ethics, misinformation*.

These keywords indicate a strong focus on both the technical and conceptual aspects of AI's role in corpus linguistics, but also on the social and ethical considerations (these latter were more prevalent in newspaper discourse). The following are also frequent lexical semantic patterns, often juxtapositioned with 'corpus linguistics', followed by a verb structure, i.e. CL + verb; for example, *driven by AI*, *powered by machine learning*, *applied in NLP*, *enhanced by deep learning.*

Retrieved key collocations and frequent lexical units included: *Artificial Intelligence, corpus linguistics, AI-driven corpus analysis, machine learning, natural language processing techniques, AI-powered language models, computational linguistics, AI integration, AI-enhanced text analysis, automated linguistic annotation, AI-based collocation extraction, deep learning in corpus studies.*

The following examples (1) – (5) illustrate some of the above collocations and verb structures retrieved first as concordances, and then as expanded text in the corpus. Key collocations are in italics.

1. *Corpus linguistics* traditionally relies on software to analyse extensive *language datasets*, uncovering patterns and insights. However, the rise of *generative AI* offers opportunities to *enhance linguistic research* by *automating* tasks, improving adaptability, and fostering innovation. (Kalaš, 2025:1)

2. In modern linguistic research, the *application of Artificial Intelligence* has led the field and provided powerful tools and prospects for linguists. (Jiang and Chen, 2024:58).

3. *LSTM (Machine Learning)* is used for extracting character features, joint vector representation and constructing *text generation models* and *generating natural language text*. (Danni *et al*., 2023).

4. In this study, we explore the potential of *LLMs in assisting corpus-based linguistic studies through automatic annotation of texts* with specific categories of linguistic information (Danni *et al*., 2023).

5. We also release *large-scale datasets* containing sentences where these *collocations* occur, which can be used for training *MWE representations*, or as a resource for corpus linguistics and lexicography (Fisas *et al*., 2020).

From the key collocations we can perceive how AI is advancing with immense potential. As artificial intelligence has become more advanced, integrating AI into corpus linguistics, significantly enhances the scope and depth of language analysis, opening up new avenues for research, application, and social impact. We continue here to summarize key areas examined in the corpus and in the literature.

The potential for this integration lies in several key areas: natural language processing (NLP) (Groenewald *et al*., 2024), machine learning (ML) (Alpaydin, 2020), deep learning (DL) (Goodfellow, *et al*., 2016), and automated language generation (Radford *et al*., 2019). Developments in these areas promise to open new frontiers for corpus linguistics, making it more powerful, efficient, and accessible.

Techniques that were once cutting-edge, like simple keyword-in-context searches, have been superseded by sophisticated algorithms capable of processing and analyzing language data at unprecedented scales (Jurafsky and Martin, 2024). These tools are no longer merely aids, they are integral to advanced tools which have elevated corpus linguistics from hypothesis-driven research to dynamic, automated discovery. NLP, for example, facilitates detailed linguistic analyses, including syntactic parsing, part-of-speech tagging, and named entity recognition. NLP techniques, such as part-of-speech tagging (Manning *et al*., 2014), syntactic parsing (Jurafsky and Martin, 2024), and named entity recognition (Finkel *et al*., 2005), can be applied to corpus data to automate the analysis of linguistic features. Meanwhile, unsupervised learning techniques, like clustering and topic modelling, uncover latent patterns in corpora without requiring predefined hypothe-ses, opening new avenues of inquiry (Blei *et al*., 2003).

More recent advancements, such as transformer-based architectures like GPT and BERT, exemplify AI's impact on language analysis. These models, trained on massive datasets, capture not only word-level meanings, but also complex syntactic and contextual relationships (Vaswani *et al*., 2017). By processing data at unprecedented scales, AI-powered tools detect subtle linguistic shifts, offering insights into diachronic language change and contemporary usage trends. Researchers can now conduct far more nuanced and efficient analyses of language data than was previously possible, in that using deep neural networks, AI systems can learn highly complex linguistic patterns across various levels of analysis, including phonology, syntax, and semantics. For instance, AI can be used to model the semantic relationships between words and phrases in a corpus, opening new avenues for understanding word meaning in context (Vaswani *et al*., 2017). For example, the well-known analogy of words like 'king' and 'queen' or 'Paris' and 'France,' recognizing that 'king' is to 'queen' as 'man' is to 'woman,' and that 'Paris' is the capital of 'France.' This enables AI to understand analogies, synonyms, and word associations in context, and makes the abstract concept of semantic relationships more tangible by showing specific word pairs and how AI processes them. These models could potentially aid in the analysis of ambiguous or context-dependent language, which remains a challenge for traditional corpus linguistics.

These types of tools are especially useful for large-scale diachronic studies that track language change over time (Davies, 2010). For example, tracking the frequency of word combinations in English texts over centuries, represents one of the ways AI has been integrated into corpus linguistics (e.g. Google's n-gram tool). By processing vast quantities of text data, AI can detect subtle shifts in word meaning, syntactic structures, and usage patterns that may be imperceptible to human researchers (Michel *et al*., 2011).

Machine learning, a subset of AI, involves the use of algorithms that learn from data to make predictions or decisions without being explicitly programmed. In corpus linguistics, machine learning techniques could be applied to improve pattern detection and predict linguistic trends. Supervised learning methods, such as classification algorithms, can be used to identify particular linguistic categories or structures within a corpus, such as sentiment, intention, or speech acts. For example, until recently sentiment was difficult to capture through automation. Now, machine learning algorithms can go beyond hypothesis testing and actively generate hypotheses based on patterns observed within the data. For example, let's say a company wants to analyze customer feedback from online reviews to determine whether customers feel positively, negatively, or neutrally about their products; through supervised Learning Applications like a classification algorithm, such as a Support Vector Machine (SVM) or a deep learning model like BERT, is trained on labeled reviews (e.g., *positive, negative, neutral*). Once trained, the model can classify new, unseen reviews automatically. Another example is a deep learning model trained on social media sentiment data which can discover that certain emojis (such as 😀) strongly correlate with excitement or humour, leading to new insights that researchers had not considered. Unsupervised learning techniques such as clustering and topic modelling could allow AI to discover latent topics and categories within a corpus, leading to new areas of research that might not have been taken into account otherwise (Blei *et al*., 2003).

By applying these techniques to large, unstructured corpora, researchers may uncover previously unrecognized language features, such as new syntactic constructions, dialects, or evolving patterns of social interaction. Methods such as clustering and topic modelling, could help discover previously unknown linguistic patterns and topics of interest, using algorithms based on word co-occurrences, allowing researchers to track shifting discourse trends, political movements, or changes in public sentiment over time. Nevertheless, as McEnery and Hardie (2012) caution, the power of these tools must be matched by a commitment to linguistic insight, ensuring that technological advancements do not overshadow the interpretive essence of the discipline.

One key challenge for the future of corpus linguistics will be ensuring linguistic diversity in AI models. Current AI models have been primarily trained on large-scale corpora of standard English and other dominant languages. However, the world is home to thousands of languages and dialects, many of which are underrepresented in current AI training datasets (Joshi *et al*., 2020). To ensure inclusivity, AI models should be trained on corpora that reflect a wider range of linguistic variation, including regional dialects, non-standard registers, and indigenous languages. For corpus linguistics, this means that future corpora will need to be more diverse, including multilingual corpora and corpora that represent underrepresented speech communities. Such inclusivity will ensure that AI tools are more effective and accessible to speakers of all languages and dialects, thus preserving linguistic diversity.

For corpus linguistics, AI-generated text can be valuable in creating synthetic corpora for specific domains or languages where data is scarce. Moreover, AI models can assist linguists by generating hypotheses or suggesting new avenues of inquiry based on patterns discovered in corpora. This integration could lead to more dynamic, real-time language studies and more comprehensive research methodologies. In fact, developments in AI-driven corpus linguistics in the ability to analyze corpora in real-time, will provide dynamic and immediate feedback to researchers and users. Current corpus analysis tools, while powerful, often require a significant amount of time to process data, especially for large corpora. However, advancements in AI-driven computational techniques will allow for the near-instantaneous processing of language data, enabling linguists to track live changes in language use, evolving slang, and even political or cultural shifts in real time. This real-time feedback loop could prove transformative in areas such as social media monitoring, political discourse analysis, and sentiment analysis. Furthermore, it could enhance the precision of language models used for tasks such as automated translation, question answering, and text generation, making them more responsive to ongoing language trends and shifts.

Additionally, the expansion of AI-driven corpus linguistics cannot be understood in isolation from broader geopolitical and economic forces. As AI technology is largely dominated by major global superpowers, issues of data control, accessibility, and epistemological authority become central to discussions of corpus linguistics' future. This study acknowledges that AI's role in corpus analysis is not just a technical evolution, but also a site of power dynamics, where institutions, corporations, and governments shape linguistic resources and knowledge production. In fact, the following section 4 incorporates a discussion of these political and ethical dimensions to avoid reducing AI's impact to mere technological progress.

4. AI-DRIVEN CORPUS LINGUISTICS FOR SOCIAL CHANGE: THE DETECTION OF SOCIAL CHALLENGES THROUGH COLLOCATION ANALYSIS

This section discusses the function of AI-driven corpus linguistics in detecting social issues through collocations, as well as how AI-driven corpus linguistics can identify key collocations that emphasise social issues such as inequality, environmental concerns, and ethical dilemmas. The analytical approach behind the reflections which follow in the text, is the result of an analysis based on the data retrieved from the corpus. The juxtaposition of observations is not intended as an isolated reflection, but rather as an attempt to illustrate the multifaceted applications of corpus linguistics in addressing social change, particularly in the AI era. The reflections arise from collocations which were retrieved pointing to social and ethical issues that frequently emerged in the two sub-corpora, providing semantic nuances which were then qualitatively interpreted. For example, recurrent lexical items referring to 'ethical' concerns: *algorithmic bias, data fairness, explainable AI, ethical NLP, transparency, data ethics, misinformation*; and frequent lexical items regarding 'social' implications: *potential social applications, increased social awareness, AI can enhance knowledge and understanding.*

Within a corpus of AI-driven linguistic texts, collocations often emerge that point to pressing global issues, such as inequality, environmental degradation, ethical dilemmas in technology, and political power structures. Frequent recurrent collocations for example are: *AI can enhance awareness/knowledge/ understanding, algorithmic bias, AI surveillance.* These collocations had other reoccurring collocates in the vicinity, for example, the lexical unit *AI can raise awareness* frequently appears in discussions surrounding social responsibility, education, and digital literacy. The collocates of *raise awareness* include words such as *bias, misinformation, ethical considerations, accessibility, AI models* and *digital divide*, indicating that discussions around AI are often framed in terms of its potential to address social inequities. Such lexical pairings highlight concerns about the ethical dimensions of AI and its role in shaping public consciousness.

Below are contextual examples of the collocations on the themes of: ethics and AI, AI and social challenges, and AI and the environment. The following extracts in examples (6) – (10) taken from the corpus, illustrate the semantic nuances which emerge, leading to debate and discussion on the role of AI in corpus linguistics for social change. The examples below illustrate the multifaceted relationship between AI and ethical considerations, social challenges, and environmental impacts, highlighting the importance of responsible AI development and deployment. Similarly, corpus analysis of AI-related texts frequently reveals collocations that point directly to issues of inequality and systemic bias. The term *algorithmic bias*, for instance, appears with high frequency alongside words such as *racial disparities, gender discrimination, socio-economic inequality,* and *marginalized communities*. This pattern suggests that discussions of AI do not occur in a political vacuum but are closely tied to ongoing debates about fairness and justice in automated decision-making. The prevalence of such collocations reinforces the argument that AI-driven systems must be designed with fairness in mind to mitigate rather than exacerbate social inequalities. This is reflected in examples (6) and (7). Key collocates are in italics.

6. The integration of *AI* in beauty industry applications has *raised awareness* about *biases* in *AI models*, particularly concerning skin tone analysis. Companies like Haut.AI and Renude are actively working to refine these models to provide more accurate and inclusive recommendations. This effort highlights the growing *awareness* and proactive measures being taken to address *AI biases* in consumer products (Vogue Business.com.2025)

7. In October 2019, researchers discovered that an *algorithm* used in U.S. hospitals to predict patient care needs *exhibited racial bias*. The study revealed that the *algorithm favored* white patients over black patients, leading to disparities in the allocation of medical resources. This example underscores how *algorithmic bias* can perpetuate existing *inequalities* in healthcare systems (Obermeyer *et al*., 2019).

Environmental concerns also emerged in the collected corpus of AI-related docuemnts, particularly through collocations surrounding the themes of *AI and sustainability/environment*. The data showed frequent associations with words like *carbon footprint, climate modelling, conservation efforts,* and *energy efficiency*, demonstrating that AI is being discussed as both a potential solution to and contributor to ecological challenges. The presence of terms like *ethical AI, responsible innovation,* and *green technology* further reflects an awareness of the need for sustainable development within the technological sphere, as shown in example (8).

8. The Vatican has expressed concerns over *AI's environmental impact*, highlighting that the significant energy consumption of *AI technologies* contributes to environmental

degradation. This underscores the need for *sustainable AI development practices* to mitigate adverse ecological effects. (Robertson, 2025)

The retrieval of key collocations has also revealed the intersections of AI with broader issues of surveillance and privacy. The collocation *AI and surveillance* frequently co-occurs with words such as *privacy concerns, state control, ethical implications,* and *mass data collection*, underscoring the societal anxiety surrounding the expansion of AI-powered monitoring technologies. These linguistic patterns provide insight into the ways public discourse frames AI as both an opportunity and a threat, depending on the context in which it is deployed. See example (9).

9. The resignation of Hoan Ton-That, CEO of Clearview AI, a company known for its extensive facial recognition database, has brought renewed attention to the *ethical implications of AI-driven surveillance*. Clearview *AI's practices* of scraping billions of images from the internet without consent have sparked debates over *privacy rights* and the potential for misuse of *surveillance technologies*. (Forbes. com. 2025)

The following example (10) draws attention to the ethical implications and social challenges of AI -driven models.

10. In 2021, the United Nations adopted the Recommendation on the *Ethics of Artificial Intelligence*, emphasizing principles such as *fairness, transparency, and accountability* to ensure *AI systems* benefit society while respecting human rights (Unesco. org. 2024)

As shown from the examples above, ultimately, AI-driven corpus linguistics can do more than simply track linguistic trends; it offers a window into the societal issues that shape and are shaped by language. The recurring collocations found in AI-related texts reveal the pressing concerns of our time, from inequality and discrimination to environmental sustainability and digital ethics. By systematically analyzing these patterns, researchers can uncover hidden biases, trace the evolution of public discourse, and contribute to more informed policymaking. Far from being a neutral tool, corpus linguistics, particularly when enhanced by AI, becomes a powerful mechanism for social insight, helping to identify and address some of the most urgent challenges facing contemporary society.

The examples above illustrate the real-world applications and challenges associated with AI technologies, emphasizing the importance of addressing biases and ethical considerations in AI development and deployment. Through the retrieval of the data and collocations the following reflections can be made, which are confirmed in the literature.

A notable dimension of change is the growing awareness of advancements in corpus linguistics which may have a significant impact on society as a whole. As corpus linguistics continues to evolve alongside AI technologies, its potential for addressing a range of societal issues becomes increasingly evident. The ability to analyze and interpret large language corpora in the context of social issues, such as identity, health, race, gender, has far-reaching implications for how society understands and addresses issues related to language, communication, and social change. Thus, the social potential of corpus linguistics cannot be overlooked, which must be integrated alongside the potential of AI, in order for it to be a positive challenge rather than a threat to the social order (The Guardian, 11 Dec., 2024; 14 Dec., 2024; 28 Dec., 2024).

In the past, the primary goal often centred on academic pursuits, such as understanding language structure, variation, and use. Today, there is an increasing emphasis on applying research findings to real-world issues (McEnery and Hardie, 2012), from enhancing educational materials (Biber *et al*., 2002; Römer, 2009) to addressing biases in AI systems (Blodgett *et al*., 2020). In other words, corpus linguistics is no longer confined to purely academic inquiries, as in its beginnings. Instead, it increasingly intersects with the fields of social sciences and media studies where its findings can shape public understanding and policy (Baker *et al*., 2013; McEnery and Brookes, 2024). Linguistic work no longer exists in isolation, or purely for language learning acquisition, but on the contrary, it has the potential to influence or even improve lives, with potentially profound social implications. This interdisciplinary expansion not only enriches the field, but also positions corpus linguistics as a key player in addressing global challenges, such as, the environment (Alexander, 2009), health, (Semino, 2022; Brookes and Collins, 2023), immigration, (Baker *et al*., 2013; Taylor, 2021), gender inequality, (Jaworska and larrivée, 2011), social inequality (Incelli, 2021; Gomez-Jimenez and Toolan, 2022), political discourse, (Partington, 2012), language diversity and minority languages (Knight *et al*., 2020), to name just a few of the areas of relevant social impact.

AI systems, particularly those trained on large, historical corpora, often inherit and perpetuate societal biases, including those related to race, gender, and socio-economic status. This can have profound consequences, especially when AI is deployed in critical applications such as criminal justice, and public policy (Bolukbasi *et al*., 2016). Corpus linguistics has the potential to identify and challenge such biases by analyzing corpora for patterns of discriminatory language use. For example, researchers could examine corpora of news articles, political speeches, or social media posts to track the ways in which certain groups are represented or misrepresented (Baker *et al*., 2013). By analyzing vast corpora of texts, AI systems can detect subtle biases, stereotypes, and discriminatory language patterns that might otherwise go unnoticed. This can be applied to fields as far and wide as journalism, law, advertising, and politics, where language shapes public perceptions and policies. By analyzing these corpora with AI-driven tools, linguists can uncover harmful linguistic practices, such as racial profiling, misogyny, or homophobia, and work toward addressing these issues in broader society, recognize and counter biased language could be an important step in promoting fairness and equality (Jaworska, 2020). In this way, corpus linguistics can become a tool for social change, empowering activists, policymakers, and communities to better understand and challenge harmful language use (Zhao *et al*., 2018).

Perhaps one of the most significant developments is the 'democratization' of corpus linguistics (McEnery and Hardie, 2012). In the beginning, access to corpora and tools was limited to a select few, often requiring significant institutional backing. Today, platforms like the *Corpus of Contemporary American English* (COCA) (Davies, 2010) and the growing availability of open-access corpora have made it possible for a much broader audience to engage with corpus linguistics. Brezina (2018) highlights the importance of accessibility in statistics and corpus tools, ensuring that even non-specialists can benefit from linguistic insights. This shift has profound implications for the future of the field, broadening its relevance and reach.

In the field of education and enhancing language learning and teaching, AI-powered corpus linguistics can revolutionize language education by providing more personalized and data-driven learning experiences. By analyzing vast corpora of spoken and written language, AI systems can identify the most relevant linguistic features that learners struggle with, enabling the development of tailored teaching materials. Additionally, automated feedback systems can provide learners with real-time corrections on their language use, making language learning more efficient and engaging (Brown *et al*., 2019). For example, there are now major software language learning programmes which have this instant feedback tool.

AI can also help educators track linguistic trends and patterns in student writing or speaking, providing insights into common errors or misconceptions For example, corpus-based tools can assist in designing curricula that reflect real-world language use, ensuring that students are exposed to language as it is naturally employed in various contexts, genres, and registers (Thompson, 2018; Lau, 2024). Additionally, within the field of language learning, corpus linguistics combined with AI can also play a significant role in promoting multilingualism and cross-cultural communication. As globalization continues, the need for effective communication across languages and cultures is more critical than ever. AI-powered tools that leverage multilingual corpora can help break down language barriers, enabling people from different linguistic backgrounds to communicate seamlessly.

Machine translation, a field that has already benefited from AI-driven improvements (e.g., Google Translate's transition to neural machine translation), can be further enhanced by incorporating more diverse and contextually aware linguistic data from larger, more comprehensive corpora (Lau, 2024). Similarly, AI-powered tools can help non-native speakers learn new languages by providing real-time translation, pronunciation assistance, and language learning aids.

AI and corpus linguistics can also contribute to the preservation of endangered languages. Many languages around the world are at risk of extinction, and AI-driven corpus analysis provides a powerful tool for preserving and revitalizing these languages. By compiling digital corpora of rare or endangered languages, linguists can ensure that these languages are documented and can be studied by future generations (Knight *et al*., 2020). Advanced technologies can assist in developing language learning resources and materials for these endangered languages. For example, AI-powered applications could create language models that help learners of endangered languages practice speaking, writing, and listening. This could be especially important for communities that wish to revitalize their native languages or for indigenous groups who are seeking to preserve their cultural heritage (see Morris, *et al*., 2024) for the Welsh language.

With these changes come new responsibilities. The sheer scale of data we now work with raises ethical questions (Brezina, 2018). Who owns this data? How should it be used? How can we ensure that our analyses respect the diversity of voices represented in these corpora? These are questions we grapple with as we look to the future, knowing that the answers are likely to shape not just our field, but also how society understands and interacts with language. These are not just technical issues; they are ethical ones, as Boyd and Crawford (2012) argue in their critical examination of big data. Handling these questions is as much a part of research as designing algorithms or analyzing concordance lines. One of the most pressing concerns is the issue of privacy and data security. As AI systems rely on large datasets, there is a need for strict ethical guidelines regarding the collection and use of data, particularly in the case of personal or sensitive information. Biases in language models can perpetuate harmful stereotypes or lead to the exclusion of certain linguistic communities, which can have a detrimental effect on society as a whole.

5. CONCLUSIONS

The future of corpus linguistics lies at the intersection of linguistic theory, computational techniques, and AI-driven innovations. As the field continues to integrate with artificial intelligence, the possibilities for enhancing linguistic research, improving educational outcomes, and promoting social equity are vast. Through the power of machine learning, natural language processing, and automated content generation, corpus linguistics can not only deepen our understanding of language but also serve as a tool for addressing some of society's most pressing issues (McEnery and Brookes, 2024).

As we move forward, it will be essential for linguists and policymakers to collaborate in shaping the ethical and responsible use of AI in corpus linguistics, as well as in language learning and in education in general, ensuring that these technologies are harnessed in ways that benefit all sectors of society. It is for the reasons mentioned above, that the future of corpus linguistics promises to be a dynamic intersection of AI technology and social impact. With the continued advancement of NLP, machine learning, and deep learning, corpus linguistics has the potential to evolve into an even more powerful tool for understanding language, identifying social trends, and promoting justice and inclusion. By addressing the challenges involved in ethical implications of responsibility and inclusivity, corpus linguistics can continue to be a force for positive social change, preserving linguistic diversity, and contributing to more equitable, multilingual, and fair societies.

I started this contribution on a personal note and I finish looking ahead, seeing a field poised at a crossroads, rich with possibilities. The next chapter of corpus linguistics will undoubtedly bring challenges, but also extraordinary opportunities to deepen our understanding of language and its role in society. As researchers, we are not just passive observers of these changes; we are active participants in shaping the future of the discipline. As I conclude, I find myself filled with a mix of curiosity and optimism. The challenges are significant, but so are the opportunities.

REFERENCES

Alexander, Richard. 2009. Framing Discourse on the Environment: A Critical Discourse Approach. Routledge, New York. https://doi.org/10.4324/9780203890615

Baker, Paul, Costas Gabrielatos, and Tony McEnery. 2013. Discourse Analysis and Media Studies: Using Corpora. Routledge. https://doi.org/10.1017/CBO9780511920103

Biber, Douglas. 1993. "Representativeness in Corpus Design". Literary and Linguistic Computing, 8(4): 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press. https://doi.org/10.1017/CBO9780511804489

Blei, David. Andrew Ng, and John Lafferty. 2003. "Latent Dirichl*et* *al*location". Journal of Machine Learning Research, 3: 993–1022.

Bleorţu, Cristina. 2024. The Use of Artificial Intelligence (AI) in Linguistics. Available at: https://www.researchgate.net/publication/386568996_2024_The_Use_of_Artificial_Intelligence_AI_in_Linguistics.

Blodgett Su, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. "Language (Technology) is Power: A Critical Survey of "Bias" in NLP", Computation and Language. https://doi.org/10.18653/v1/2020.acl-main.485

Bolukbasi Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". Advances in Neural Information Processing Systems, 29.

Boyd Danah, Kate Crawford. 2012. "Critical Questions for Big Data". Information, Communication & Society, 15(5): 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brezina, Vaclav. 2018. Statistics in Corpus Linguistics: A Practical Guide. Cambridge University Press. https://doi.org/10.1017/9781316410899

Brookes, Gavin, and Luke Collins. 2023. Corpus Linguistics for Health Communication, A Guide for Research. Taylor & Francis Ltd, Routledge Books. https://doi.org/10.4324/9781003099659

Brown, A., Taylor, R., Wilson, K. 2019. "Automated feedback in language learning: A practical approach". Language Learning Journal, 45(3): 245–261.

Danni Yu, Li Luyang, and Su Hang. 2023. Using LLM-assisted Annotation for Corpus Linguistics: A Case Study of Local Grammar Analysis. arXiv - CS - Artificial Intelligence.

Davies, Mark. 2008. "The Corpus of Contemporary American English (COCA): 520 Million Words, 1990–Present." Available online at https://www.english-corpora.org/coca/.

Davies, Mark. 2010. The Corpus of Historical American English (COHA): 400 Million Words, 1810-2009.

Dawn Knight, Steve Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić, Enlli Môn Thomas. 2020. “The national corpus of contemporary Welsh: project report | Y corpws cenedlaethol Cymraeg cyfoes: adroddiad y prosiect”. Project Report. CorCenCC.2018.

Finkel, Jenny. R., Trond Grenager, Christopher D. Manning. 2005. “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”. ACL. https://doi.org/10.3115/1219840.1219885

Fisas, Beatriz, Luis Espinosa Anke, Joan Codina-Filbá, and Leo Wanner. 2020. CollFrEn: Rich Bilingual English–French Collocation Resource. This is the repository for the MWE-LEX 2020.

Gomez-Jimenez, Eva M., and Michael Toolan, M. 2022. The Discursive Construction of Economic Inequality, CADS Approaches to the British Media. Bloomsbury publishing.

Gu Feng, 2023. “Corpus-based critical discourse analysis on AI Policy: A comparison between North America and Developing Countries in East Asia”. Asian Journal of Social Science Studies 8(3):14. https://doi.org/10.20849/ajsss.v8i3.1371

Incelli, Ersilia. 2021. “But what’s so bad about inequality? Ideological positioning and argumentation in the representation of economic inequality in the British Press”. Pp. 77 – 100 in Argumentation, Ideology and Discourse in Evolving Specialized Communication, edited by J. Bowker, E. Incelli, C. Prosperi-Porta, Lingua e Linguaggi. Special Issue, 42.

Jaworska, Sylvia. 2020. *Corporate discourse*, Cambridge University Press. https://doi.org/10.1017/9781108348195.031

Jaworska, Sylvia and Pierre larrivée. 2011. “Women, power and the media: Assessing the bias”, Journal of Pragmatics, 43(10):2477-2479. https://doi.org/10.1016/j.pragma.2011.02.008

Jiang, Shaohua, and Zeng Chen. 2024. Applications and Prospects of Artificial Intelligence in Linguistic Research. 3C Tecnología. Glosas de innovación aplicada a la pyme 13(1): 57-76.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, Monojit Choudhury. 2020. “The State and Fate of Linguistic Diversity and Inclusion in the NLP World”. ACL Anthology, pp. 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Jurafsky, Daniel, and James. H. Martin, 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. (Updated Edition). Pearson.

Kalaš, Filip. 2025. Bridging Tradition and Innovation: Analysing Language Data with Chatgpt-4 in Corpus Linguistics. Available at SSRN. https://doi.org/10.2139/ssrn.5126316

Kilkarif, Adam, Pavel Rychly, Pavel Smerz, David, Tugwellet. 2006. The Sketch Engine, Lexicography MasterClass and ITRI, University of Brighton, U.K.

Lau, Ethan. 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. DOI:10.22214/ijraset.2024.61039

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, David McClosky. 2014. Stanford CoreNLP: A Java Suite for NLP Tools.

McEnery, Tony, and Andrew Hardie. 2012. Corpus Linguistics: Method, Theory, and Practice. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

McEnery, Tony, and Gavin Brookes. 2024. “Corpus Linguistics and the Social Sciences”. *Corpus* Linguistics and Linguistic Theory. 20 (3): 591-613. https://doi.org/10.1515/cllt-2024-0036

Michel Jean-Baptiste, Yuan Kui Shen, Aviva Aiden, Adrian Veres. 2011. “Quantitative analysis of culture using millions of digitized books”. Science, 331(6014): 176–182. https://doi.org/10.1126/science.1199644

Morris, Jonathan, Ignatius Ezeani, Katharine Young, Lynne Davies, Mahmoud El-Haj, Gareth Watkins, Dawn, Knight (Eds). 2024. Language and Technology in Wales: Volume II. Bangor University.

Nelson Francis, W. N., and Henry. Kuçera, 1964. The Brown Corpus of Standard American English. Brown University Press.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366: 447–453. https://doi.org/10.1126/science.aax2342

Partington, Alan. 2012. “Corpus Analysis of Political Language” in The Encyclopedia of Applied Linguistics, edited by C. A. Chapelle, Wiley. https://doi.org/10.1002/9781405198431.wbeal0250

Robertson, Michelle. 2025. The Vatican’s Stance on AI: Understanding Antiqua et Nova. OLA Communications Officer.

Gruetzemacher, Ross. 2022. “The Power of Natural Language Processing”. Harvard Business Review. April. 19.

Semino, Elena. 2022. “Health Communication”. Pp. 276-290 in Introducing Linguistics, edited by J. Culpeper, B, Malory, C. Nance, D., van Olmen, D., Atanasova, S. Kirkham, A. Casaponsa, London: Routledge.

Sinclair, John. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Stubbs, Michael. 2001. Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell.

Taylor, Charlotte. 2021. “Metaphors of migration over time”. Discourse and Society, 32, (4). https://doi.org/10.1177/0957926521992156

The Guardian, Dec. 14, 2024 The Guardian view on AI’s power, limits, and risks: it may require rethinking the technology. Accessed December 2024, https://www.theguardian.com/society/2024/dec/11/ai-tone-shifting-tech-could-flatten-communication-apple-intelligence.

The Guardian, Dec. 11, 2024. Losing our voice? Fears AI tone-shifting tech could flatten communication. Accessed December 2024, https://www.theguardian.com/society/ 2024/dec/11/ai-tone-shifting-tech-could-flatten-communication-apple-intelligence

The Guardian, 28 Dec. 2024. How will AI reshape 2025? Well, it could be the spreadsheet of the 21st century. Accessed December 2024, https://www.theguardian.com/ commentisfree/2024/dec/28/llms-large-language-models-gen-ai-agents-spreadsheets-corporations-work.

Vaswani Ashish, Noam Shazeer, Nikki Parmar, and Jakob Uszkoreit. 2017. “Attention is All You Need”. Advances in Neural Information Processing Systems, 30.

Zhao Jieyu, Wang Tianlu, Mark Yatskar, Chang Kai-Wei. 2019. “Gender Bias Gender Bias in Contextualized Word Embedding”. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/N19-1064

Received on 10-01-2025 Accepted on 06-02-2025 Published on 10-03-2025

https://doi.org/10.6000/2818-3401.2025.03.01