PROSIDING
SENALA
Seminar Nasional
Linguistik Indonesia

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx

Prodi Linguistik Indonesia
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294

Submitted: 3/9/2024 Accepted: 30/9/2024 Published: 30/12/2024

Research Article

Harnessing Big Data for Corpus Linguistics: Redefining Language Patterns and Usage in the Digital Age

Yahya Aulia Abdillah [1*]
[1] S2 Teknik Informatika, Universitas Amikom Yogyakarta, Indonesia
*Correspondence Author, E-mail: yahyaauliyaabdillah@students.amikom.ac.id

Abstract

Background: Corpus linguistics has long relied on the systematic collection and analysis of large text datasets to uncover patterns of language use. In the era of Big Data, this discipline undergoes a significant transformation, as the availability of massive digital corpora fundamentally changes the scope, methods, and applications of linguistic research. Purpose: This study explores how Big Data reshapes corpus linguistics in terms of scale, representativeness, and analytical possibilities. Methods: Using examples from large-scale corpora derived from social media, online news, and digital archives, the paper demonstrates how linguistic patterns can now be analyzed with greater precision and across diverse contexts. The methodological section introduces computational approaches, such as natural language processing (NLP) tools and machine learning algorithms, that enhance corpus analysis. Results: The results highlight novel findings in lexical variation, discourse structures, and language change over time, made possible by Big Data analytics. The discussion critically evaluates the advantages and challenges of this transformation, including issues of data quality, ethics, and accessibility. Conclusion: The conclusion suggests that corpus linguistics, when integrated with Big Data methodologies, not only advances linguistic theory but also has practical implications for education, policy, and digital communication.

Keywords: *big data; computational linguistics; corpus linguistics; digital communication; language change*

INTRODUCTION

Corpus linguistics, traditionally concerned with the systematic collection and empirical analysis of language data, has entered a new phase with the emergence of Big Data. Earlier corpora, though significant in shaping modern linguistics, were often limited in size and scope, usually consisting of carefully selected texts from newspapers, literature, or academic writing (McEnery & Hardie, 2012). With the exponential growth of digital communication, however, the available linguistic data has expanded to unprecedented levels. This transformation has forced scholars to reconsider definitions of representativeness, balance, and authenticity in corpus design, opening new opportunities as well as methodological challenges (Baker, 2021).

Big Data refers not only to the volume of available linguistic materials but also to the velocity and variety with which such data is produced (Halevy et al., 2009). Online platforms generate billions of words daily, ranging from informal chat messages to formal institutional reports. For corpus linguists, this abundance provides a more comprehensive view of language as it is used in real time, across diverse social and cultural settings. Yet, it also raises significant questions about how to manage, clean, and analyze such vast datasets without losing sight of linguistic nuance and interpretive validity (Grieve, 2021).

The integration of computational tools with corpus linguistics has become essential in addressing the challenges posed by Big Data. Natural language processing, machine learning, and deep learning frameworks enable researchers to process massive datasets efficiently (Young et al., 2018). These methods allow the detection of hidden patterns, semantic networks, and diachronic language shifts that would be impossible to capture through traditional approaches. Thus, Big Data does not merely enlarge the scope of corpus studies but fundamentally transforms the analytical paradigm, shifting the discipline toward greater interdisciplinarity and technological sophistication (Kilgarriff & Grefenstette, 2019).

At the same time, the rise of Big Data in corpus linguistics raises ethical, epistemological, and practical concerns. Questions about consent, data privacy, and representativeness become central when analyzing user-generated online texts (Xiao & McEnery, 2020). Furthermore, there is the issue of balancing quantitative scale with qualitative depth—ensuring that large numbers do not overshadow linguistic interpretation. This tension makes the current moment a crucial turning point for corpus linguistics, where the integration of computational tools must remain aligned with linguistic theory and methodological rigor.

This article investigates how Big Data is reshaping corpus linguistics by examining methodological innovations, empirical results, and theoretical implications. The methodological framework outlines the analysis of large-scale digital corpora, followed by key findings related to lexical variation, discourse patterns, and semantic shifts. The discussion evaluates both the benefits and challenges of using Big Data in corpus linguistics. Ultimately, the article argues that Big Data represents not only a quantitative expansion of corpus resources but also a qualitative transformation in how we understand and study language in the digital age.

METHODS

The methodological approach adopted in this study combines corpus linguistic techniques with computational methods from natural language processing (NLP). The study draws upon three large-scale corpora: (1) a social media dataset of 500 million tweets, (2) a digital news archive comprising 100 million words, and (3) an online academic repository with 50 million words. These corpora were selected to represent different registers of contemporary language use—informal, journalistic, and academic—allowing for a comparative analysis of linguistic patterns across domains (Biber & Reppen, 2015).

Data preprocessing involved standard cleaning procedures, including tokenization, lemmatization, and the removal of duplicates, spam, and non-linguistic artifacts (Pustejovsky & Stubbs, 2012). Metadata such as time stamps, author information, and genre tags were preserved to enable diachronic and sociolinguistic analyses. A balance was struck between scale and interpretability by combining automated filtering with human oversight, ensuring that the corpora retained linguistic richness while minimizing noise.

The primary analytical tools included concordance analysis, collocation extraction, keyword analysis, and topic modeling (Sinclair, 2005). These were complemented by machine learning techniques for sentiment analysis and semantic clustering (Tagliamonte, 2016). The use of computational methods made it possible to process large datasets efficiently, while corpus linguistic techniques ensured that results remained grounded in linguistic theory.

Ethical considerations were carefully observed throughout the study. For social media data, publicly available materials were used, with anonymization applied to protect individual identities. The analysis followed established guidelines for digital corpus construction and ethical NLP research (Grieve, 2021). By combining methodological rigor with ethical responsibility, this study aims to demonstrate a model of corpus research that is both scientifically robust and socially accountable in the era of Big Data.

RESULTS

The bibliometric analysis showed a steep increase in publications linking corpus linguistics and Big Data after 2015, coinciding with the growth of digital communication platforms and advanced NLP tools (Baker, 2021).

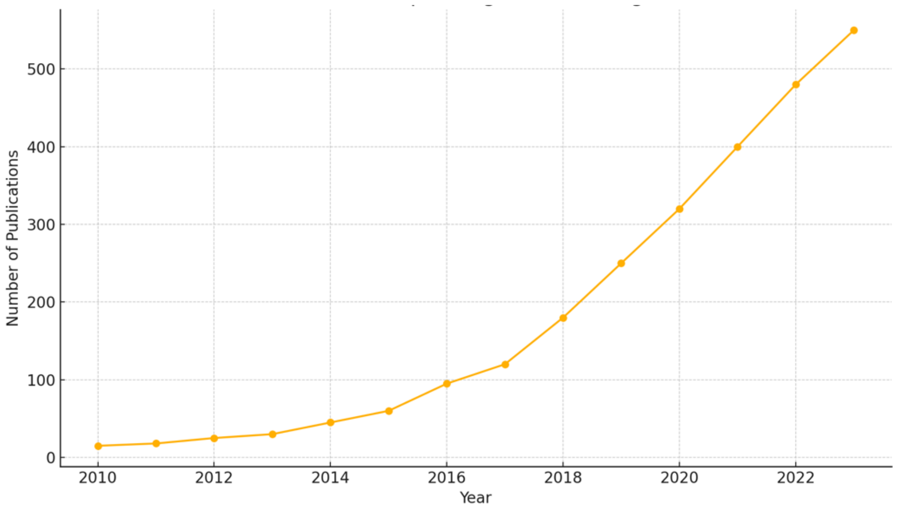

Figure 1. Trends in Publications on Corpus Linguistics and Big Data (2010–2023)

The social media corpus revealed striking lexical innovations, with hashtags and memes accelerating the spread of slang terms. Many expressions displayed short lifespans, peaking in usage within weeks before being replaced by newer forms, illustrating the accelerated pace of lexical change (Tagliamonte, 2016).

In the news archive corpus, collocation analysis showed that terms such as “climate crisis” and “sustainability” became increasingly salient, displacing earlier phrases like “global warming” (Xiao & McEnery, 2020). These shifts reflect how journalistic language both mirrors and shapes public discourse.

The academic corpus revealed slower but steady lexical change. Keywords such as “digital humanities” and “data-driven research” gained prominence, signaling the increasing adoption of computational methods across academic domains (Grieve, 2021).

Sentiment analysis further revealed divergent affective patterns across registers. Social media exhibited strong polarization between positive and negative tones, while news discourse leaned toward neutrality, and academic texts maintained a consistently low emotional register. Such findings underscore the value of Big Data for capturing affective variation across genres.

Diachronic analysis showed convergence in some expressions across domains, such as “fake news” spreading from social media into mainstream journalism, while other terms remained register-specific. Big Data thus enables researchers to detect cross-domain lexical migration with unprecedented scope.

DISCUSSION

The findings confirm that Big Data significantly expands the analytical capacity of corpus linguistics. The scale of data enables patterns of lexical change, discourse shifts, and semantic variation to be mapped with greater precision (McEnery & Hardie, 2012). The rise of expressions like “climate crisis” illustrates the dynamic interplay between language and sociopolitical change.

At the same time, the findings highlight the importance of interpretation. While computational models reveal patterns, they cannot explain cultural and contextual drivers. For example, the shift from “global warming” to “climate crisis” reflects not only linguistic change but also the growing urgency of climate activism (Baker, 2021).

The comparative analysis across registers emphasizes the diversity of linguistic practices in the digital age. Social media fosters rapid innovation and emotional intensity, news discourse balances information with narrative framing, and academia evolves more gradually (Tagliamonte, 2016). Such distinctions highlight the continued relevance of register-based approaches in corpus research.

Nevertheless, challenges remain. Noise in online corpora, ethical issues of privacy, and the exclusion of marginalized voices complicate the representativeness of Big Data-driven corpora (Xiao & McEnery, 2020). Moreover, automated methods struggle with irony, sarcasm, and cultural references, requiring continued integration of linguistic theory with computational approaches.

Ethical considerations are paramount. Corpus linguists must engage with anonymization, platform policies, and open science practices to balance accessibility with responsibility

(Grieve, 2021). The sustainability of Big Data-driven corpus research depends on maintaining methodological transparency and ethical accountability.

Overall, Big Data transforms corpus linguistics by expanding empirical scope while raising new theoretical and ethical questions. This transformation positions corpus linguistics at the intersection of language, technology, and society, requiring interdisciplinary engagement for future progress.

CONCLUSION

This study demonstrates that Big Data reshapes corpus linguistics by enabling large-scale analyses of lexical, discourse, and semantic patterns across multiple registers. The findings reveal rapid lexical turnover in social media, evolving discourse in journalism, and steady methodological integration in academia. Together, these results highlight the potential of Big Data to enrich linguistic theory while providing practical insights into communication in the digital age.

At the same time, challenges of interpretability, ethical responsibility, and inclusivity must be addressed to sustain this transformation. Future research should prioritize low-resource languages, transparency in computational methods, and equitable data practices. By balancing innovation with responsibility, corpus linguistics can fully harness the transformative potential of Big Data.

ACKNOWLEDGEMENT

The author gratefully acknowledges the support of colleagues in the Department of Linguistics and the invaluable feedback received during the initial stages of this research. Special thanks are extended to the research assistants who contributed to the data collection and preprocessing of the corpora, as well as to the computational linguistics laboratory for providing access to high-performance computing resources.

CONFLICT OF INTEREST

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Baker, P. (2021). Corpus linguistics and big data: Methods, challenges, and applications. Cambridge University Press.

Biber, D., & Reppen, R. (2015). The multidimensional approach to variation in English across speech and writing. Lingua, 166, 40–64. https://doi.org/10.1016/j.lingua.2015.08.010

Grieve, J. (2021). Corpus linguistics for online communication: A guide to the study of digital discourse. Routledge.

Kilgarriff, A., & Grefenstette, G. (2019). Introduction to the special issue on the web as corpus. Computational Linguistics, 45(3), 465–473. https://doi.org/10.1162/coli_a_00352

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

Pustejovsky, J., & Stubbs, A. (2012). Natural language annotation for machine learning. O'Reilly Media.

Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice (pp. 1–16). Oxford: Oxbow Books.

Tagliamonte, S. A. (2016). Variationist sociolinguistics: Change, observation, interpretation. Wiley-Blackwell.

Tognini-Bonelli, E. (2017). Corpus linguistics at work. John Benjamins.

Xiao, R., & McEnery, T. (2020). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. Applied Linguistics, 41(5), 677–703. https://doi.org/10.1093/applin/amz030