

Corpus Linguistics in Modern Linguistic Research: A Theoretical Review of Evolution, Methodologies, and Future Horizons

Cong Khanh Ly  ¹

¹ M.Ed. in TESOL, English lecturer, Faculty of Foreign Languages, Ho Chi Minh City University of Industry and Trade (HUIT), Vietnam

✉ Corresponding author:
khanhlc@huit.edu.vn



This work is licensed under CC BY-NC-ND 4.0.
To view a copy of this license, visit
<https://creativecommons.org/licenses/by-nc-nd/4.0>

ABSTRACT: This research paper provides a systematic theoretical review of corpus linguistics within modern linguistic research. The primary objective is to clarify the shifting conceptual definitions of a corpus, trace its historical development from manual text collections to the digital era, and evaluate its core methodological frameworks. Using a qualitative desk-research design, this study filters and analyzes core academic texts and foundational literature published since the 1960s. The results show that the definition of a corpus has evolved from a loose text collection into a highly structured, digital dataset designed for specific research questions. Furthermore, historical data highlights major technological milestones, such as the creation of early computer corpora and the rise of mega-corpora, which shifted linguistics into an empirical science. The analysis also reveals that while the field remains theoretically divided between corpus-based and corpus-driven methods, modern applied linguistics increasingly blends both approaches to solve practical problems. The significance of these findings lies in their application to language education and artificial intelligence. Understanding corpus design helps practitioners build accurate language teaching materials while providing the baseline data needed to train natural language processing tools, ensuring that future digital innovation remains anchored in natural human communication.

KEYWORDS: corpus, corpus-based, corpus-driven, corpus linguistics, applied linguistics, digital data.

How to cite: Ly, C. K. (2026). Corpus Linguistics in Modern Linguistic Research: A Theoretical Review of Evolution, Methodologies, and Future Horizons. *World Conference on Science, Innovation and Human Progress*. Futurity Research Publishing. <https://doi.org/10.5281/zenodo.21261626>

Introduction

For a long time, linguistic research relied heavily on the personal intuition and theoretical reflection of scholars. Researchers often create their own examples to analyze language structures. However, the rise of computing technology in the late twentieth century shifted the discipline toward empirical, evidence-based approaches. At the center of this change is corpus linguistics, which investigates language through large, digital collections of real-world texts known as corpora. Today, corpus linguistics is recognized as a powerful methodological framework rather than an isolated theory. It allows linguists to discover what is actually frequent and natural in daily communication (McEnery & Brezina, 2022). Despite its widespread use, the field currently faces several challenges. First, as digital technology advances, the precise definition and boundaries of a corpus are becoming less clear (Gauthier, 2025). Many researchers struggle to collect data that truly reflects real language use. Second, scholars continue to debate whether a corpus should only be used to validate existing theories, or if the data itself should drive the creation of new linguistic descriptions (Tognini-Bonelli, 2001). Finally, there is a gap between rapid technological developments and basic research practices. Many language practitioners use advanced software and digital vocabulary materials, but they often lack a clear understanding of how the underlying corpus data was designed (Crosthwaite, 2024). This challenge is further intensified in the era of artificial intelligence, where language teachers increasingly rely on automated platforms and Large Language Models (LLMs) to generate and evaluate ELT coursebooks without grounding these data-driven analyses in established linguistic theories (Ly, 2026a).

This lack of methodological clarity can lead to errors in data analysis. To address these issues, researchers need a systematic overview of the field. This research paper provides a concise review of corpus linguistics by examining its historical development, clarifying its core methodologies, and outlining its future directions in modern applied linguistics.

Research Aim and Research Questions

The primary objective of this research paper is to provide a systematic and critical review of corpus linguistics within modern linguistic research. Specifically, this study aims to clarify the shifting definitions of a corpus, trace its historical development from early text collections to the digital era, and evaluate its core methodological frameworks. By doing so, the paper seeks to address current challenges in data interpretation and outline the future directions of corpus applications in an increasingly technology-driven academic landscape.

To achieve this main goal, the study synthesizes the evolving conceptual boundaries of a corpus in existing literature, identifies the key technological shifts that drove the historical evolution of the discipline, and analyzes the practical contributions and methodological divisions in modern corpus applications.

To guide the investigation, this study addresses the following three research questions:

1. **RQ1:** How is the concept of a corpus defined, characterized, and classified in foundational and current linguistic literature?
2. **RQ2:** What are the key historical milestones and technological changes that have shaped the evolution of corpus linguistics?

3. **RQ3:** What are the core methodological approaches in corpus studies, and how do they inform the future directions of modern applied linguistics?

Literature Review

1. Definitions and Classifications of a Corpus

To understand the scientific value of corpus linguistics, researchers must first establish what a corpus actually is. Early studies defined it simply as any collection of texts chosen for language analysis (Francis & Kučera, 1982). As computer tools advanced, Sinclair (2004b) emphasized that a true corpus must consist of naturally occurring texts, collected according to strict rules. Modern scholars state that a corpus must be digital, organized, representative, and limited in size (McEnery & Hardie, 2012).

Previous literature divides corpora into specific types based on research goals. General corpora, like the British National Corpus, represent a language as a whole (Biber et al., 2021). In contrast, specialized corpora focus on narrow areas, such as academic (Chang, 2014) or medical language (Florescu & Ohniwa, 2025). While these divisions are well-known, the rapid rise of internet communication challenges these traditional categories, creating a clear need to review how these definitions adapt to new digital formats. Moreover, from a pedagogical perspective, the structural representative of specialized corpora serves as a baseline for evaluating materials, ensuring that selected resources achieve an appropriate skills-systems balance and align with learners' needs (Ly, 2025).

2. Historical Development and Technological Shifts

The history of corpus linguistics is closely tied to computer progress. Before the 1960s, empirical language study relied on manual data collection, in which researchers used paper index cards to count word frequencies (Kennedy, 1998). This early era required a lot of hard work, which greatly limited the size and scope of datasets. The field experienced a major change with the creation of the Brown Corpus in the 1960s, which contained one million words of written American English (Meyer, 2002). This project proved that computers could successfully store and analyze large amounts of language data.

During the 1990s, the discipline grew quickly with the development of large corpora containing hundreds of millions of words (Brezina, 2018). Scholars like Biber (1993) showed that these big datasets allowed for accurate statistical analyses of grammatical patterns and text variations. Understanding this history is important because it shows how current computer limits directly affect data reliability.

3. Methodological Divisions: Corpus-Based versus Corpus-Driven

A major point of discussion in previous research is the methodological division between corpus-based and corpus-driven approaches. Tognini-Bonelli (2001) clearly explained this distinction, which remains central to modern debates. A corpus-based approach uses corpus data mainly to check, test, or change existing linguistic theories (Xiao, 2008). In this model, the theory remains the main driver, and the corpus serves as a tool to validate it.

On the other hand, a corpus-driven approach claims that the data itself should be the only source of linguistic description (Biber, 2015). Researchers allow the patterns within the corpus to create new theoretical ideas without pre-existing assumptions. Recent reviews show that while this distinction is

clear in theory, practical applications often mix the two approaches (McEnery & Hardie, 2012). Many modern studies combine elements of both methods without a clear reason. This lack of clarity can weaken the validity of the research findings, which is why a clear review of these core methods is essential. Recent conceptual investigations reiterate this concern, arguing that advanced computational and AI-supported tools must function purely as analytical supports rather than autonomous evaluators to maintain theoretical integrity (Ly, 2026b).

Research Methodology

This study uses a qualitative desk research design to conduct a systematic review of the corpus linguistics literature. As a theoretical review, the methodology focuses on a clear process of selecting, filtering, and analyzing published academic texts to ensure reliability (Snyder, 2019).

The research process followed three main steps. First, an initial search was conducted across major databases, including Scopus, Google Scholar, Web of Science, and ResearchGate. The search terms included "*corpus linguistics definition*", "*history of corpus linguistics*", and "*corpus-based versus corpus-driven*". The search was limited to books, book chapters, and peer-reviewed journal articles.

Second, strict selection rules filtered the texts. To be included, publications had to focus directly on the theoretical development or methodology of corpus linguistics and be published by recognized academic sources. General opinion pieces, brief abstracts, and non-academic web sources were excluded, leaving a final selection of core texts from the 1980s to recent papers.

Finally, the selected literature was analyzed using thematic analysis based on three categories: conceptual definitions, historical milestones, and methodological frameworks. Comparing these categories allows this study to identify the main theoretical shifts in the discipline.

Research Results and Discussion

The analysis of the selected literature provides clear answers to the three research questions regarding the definitions, history, and methods of corpus linguistics.

1. Conceptual Definitions and Classifications

Regarding RQ1, the results show that the definition of a corpus has evolved from a loose collection of texts into a highly structured dataset. Foundational literature initially viewed a corpus as a general collection of texts used for language analysis (Francis & Kučera, 1982; Sinclair, 2004a). In contrast, current literature establishes strict criteria: modern corpora must be digital, systematically organized, and representative of a specific language variety (McEnery & Hardie, 2012; Wynne, 2005). To achieve representativeness, researchers now balance their data selection across different text genres rather than just gathering a large volume of words. They must account for variables like spoken versus written modes, formal versus informal registers, and demographic factors (Ne'matova & Hakima, 2025; Szmrecsanyi & Engel, 2023).

Furthermore, classifications have expanded. While early research relied mostly on general corpora to study whole languages, current studies frequently utilize specialized, parallel, and learner corpora (Hakima & Iroda, 2025). These specialized tools allow researchers to target the specific linguistic features of distinct professional domains. This shift indicates that a corpus is no longer just a

passive storage of words, but an active tool designed for specific research questions (Egbert et al., 2022)s.

2. Historical Milestones and Technological Shifts

For RQ2, the historical data highlights two major technological milestones. The first milestone was the transition from manual counting to computer storage in the 1960s, marked by the creation of the Brown Corpus (Francis & Kučera, 1982; Kholkovskaia, 2017). This shift allowed researchers to move past paper index cards and analyze one million words efficiently.

The second milestone occurred in the 1990s with the rise of mega-corpora like the British National Corpus, which expanded data sizes to hundreds of millions of words (Biber, 1993; Friginal et al., 2014). These technological changes did not just make data collection faster; they changed linguistic theory by proving that rare grammatical patterns can only be detected through massive datasets. Instead of relying on personal opinions, linguists can now use real frequency numbers to describe how people actually use language. This empirical shift forced researchers to re-examine traditional grammar rules, as corpus data often revealed that actual language use differs significantly from textbook examples (Li, 2024).

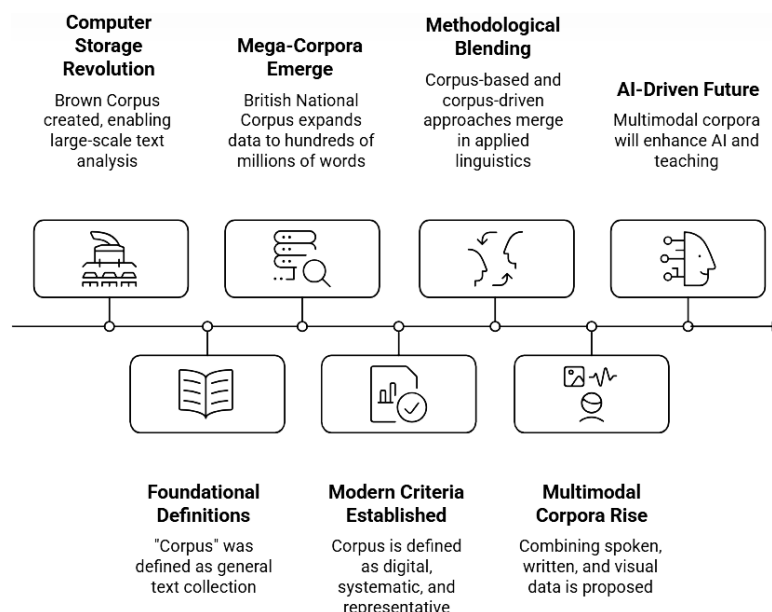
3. Methodological Approaches and Future Horizons

For RQ3, the results confirm that the field remains theoretically divided between corpus-based and corpus-driven methodologies (Lu et al., 2021; Tognini-Bonelli, 2001). The corpus-based approach uses data to test pre-existing theories, while the corpus-driven approach allows data to form new theories from scratch (Meyer, 2014). However, this strict division is becoming less useful in practice. Modern researchers usually combine both methods to analyze actual texts more effectively. A study might begin with an existing theoretical framework but remain open to new patterns that emerge from the software analysis.

These methodologies directly inform the future of the discipline. As language research enters the era of artificial intelligence, the focus is moving toward multimodal corpora that combine spoken, written, and visual data (Zhao & Nazir, 2022). Future studies will focus on how these integrated datasets can improve language teaching materials and natural language processing tools, ensuring that corpus linguistics remains central to digital innovation (Ali et al., 2026). In local contexts, such as higher education in Vietnam, integrating these technology-enhanced corpus tools can effectively foster learner autonomy and bridge the gap between theoretical insights and digital classroom applications (Tran & Ly, 2025).

Figure 1

Key milestones in corpus linguistics (Source: author's own development)



Conclusions

This theoretical review has synthesized the development of corpus linguistics across its definitions, history, and methods. The recapitulation shows that a corpus has evolved from a simple text collection into a structured, digital tool. Major technological milestones, including early computer storage and the rise of mega-corpora, have transformed linguistics into an empirical science, making traditional methodological divisions more flexible in modern practice.

These findings carry important implications for language education and digital research. For practitioners, corpus insights help design accurate teaching materials based on real-world usage rather than intuition. For the broader scientific community, corpus methodologies provide the baseline data needed to train modern artificial intelligence and natural language processing tools, ensuring machine learning reflects natural human communication.

However, this study has certain limitations, as it focuses primarily on written text corpora and relies heavily on Western linguistic frameworks. For further research in the future, it is highly recommended that scholars investigate multimodal corpora, combining visual and spoken data to better capture digital communication. Future studies should also explore how novice researchers in diverse geographic contexts adapt corpus design rules to local language teaching needs, ensuring continuous innovation in the discipline.

References

- Ali, M., Ali, N., & Khuram, M. (2026). Corpus Studies, Current Trends and Future Directions. *Contemporary Journal of Social Science Review*, 4(1), 122-136. <https://doi.org/10.63878/cjssr.v4i1.1876>
- Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4), 243-257. <https://doi.org/10.1093/llc/8.4.243>

- Biber, D. (2015). Corpus-Based and Corpus-Driven Analyses of Language Variation and Use. In B. Heine, & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (2nd ed., pp. 193-224). Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780199677078.013.0008>
- Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). *Grammar of Spoken and Written English*. John Benjamins Publishing Company. <https://doi.org/10.1075/z.232>
- Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press. <https://doi.org/10.1017/9781316410899>
- Chang, J.-Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, 26(2), 243-259. <https://doi.org/10.1017/S0958344014000056>
- Crosthwaite, P. (Ed.). (2024). *Corpora for Language Learning: Bridging the Research-Practice Divide* (1st ed.). Routledge. <https://doi.org/10.4324/9781003413301>
- Egbert, J., Biber, D., & Gray, B. (2022). Corpus Design and Representativeness in Practice – With Daniel Keller. In J. Egbert, D. Biber, & B. Gray (Eds.), *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness* (pp. 177-219). Cambridge University Press. <https://doi.org/10.1017/9781316584880.007>
- Florescu, C. M., & Ohniwa, R. L. (2025). On the replicability of corpus-derived medical word lists. *Applied Corpus Linguistics*, 5(2), 100130. <https://doi.org/10.1016/j.acorp.2025.100130>
- Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.
- Friginal, E., Walker, M., & Randall, J. B. (2014). Exploring mega-corpora: Google Ngram Viewer and the Corpus of Historical American English. *EuroAmerican Journal of Applied Linguistics and Languages*, 1(1), 48-68. <https://doi.org/10.21283/2376905X.1.4>
- Gauthier, L. (2025). Digital Humanities and the Notion of Corpus in Ancient History. *Digital Studies/Le champ numérique*, 15(1), 1-28. <https://doi.org/10.16995/dscn.18169>
- Hakima, A., & Iroda, A. (2025). Types of Corpora: General, Specialized, and Parallel. *THEORY OF SCIENTIFIC RESEARCHES OF WHOLE WORLD*, 2(5), 36-39. Retrieved June 11, 2026, from <https://esiconf.org/index.php/TOSROF/article/view/6038/6026>
- Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Routledge.
- Kholkovskaia, O. (2017). Role of the Brown Corpus in the History of Corpus Linguistics. *Poster 2017 Proceedings*. Prague: Sciencesconf. Retrieved June 11, 2026, from https://poster.fel.cvut.cz/poster2017/proceedings/Poster_2017/Section_HS/HS_018_Kholkovskaia.pdf
- Li, L. X. (2024). Developing Strategies to Improve Textbook Design Using Synergy of Native and Learner Corpora. *Journal of Psycholinguistic Research*, 53, 76. <https://doi.org/10.1007/s10936-024-10116-2>
- Lu, R., Ali, A. M., & Ghani, C. A. (2021). A Comparative Study of Corpus-based and Corpus-driven Approaches. *Linguistics International Journal*, 15(2), 119-132. Retrieved June 11, 2026, from https://www.researchgate.net/publication/358125006_A_Comparative_Study_of_Corpus-based_and_Corpus-driven_Approaches
- Ly, C. K. (2025). Language Teaching Materials Evaluation: Principles from Theoretical Perspectives. *World Conference on Emerging Science, Innovation and Policy 2025*. 3. Poland: Futurity Research Publishing. <https://doi.org/10.5281/zenodo.15961001>

- Ly, C. K. (2026a). Artificial Intelligence in Language Teaching Materials: A Review of Current Trends and Future Directions. *Conference Proceedings The First SVAIL International Conference: Language Research and Teaching with AI*. Ho Chi Minh City, Vietnam: Social Sciences Publishing House. Retrieved from <https://drive.google.com/file/d/1YWCYr4UtPKeBTy5c8w0dT7YLMJVMWEm-/view>
- Ly, C. K. (2026b). Revisiting Linguistic Evaluation of ELT Coursebooks in the Era of AI. *International Multidisciplinary Conference on Innovation, Technology and Sustainability*. Poland: Futurity Research Publishing . <https://doi.org/10.5281/zenodo.18835734>
- McEnery, T., & Brezina, V. (2022). Everyday Linguistics: Form and Function. In T. McEnery, & V. Brezina, *Fundamental Principles of Corpus Linguistics* (pp. 154-183). Cambridge University Press. <https://doi.org/10.1017/9781107110625.007>
- McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511981395>
- Meyer, C. F. (2002). Corpus analysis and linguistic theory. In C. F. Meyer, *English Corpus Linguistics: An Introduction* (pp. 1-29). Cambridge University Press. <https://doi.org/10.1017/CBO9780511606311.002>
- Meyer, C. F. (2014). Corpus-based and corpus-driven approaches to linguistic analysis: one and the same? In I. Taavitsainen, M. Kytö, C. Claridge, & J. Smith (Eds.), *Developments in English: Expanding Electronic Evidence* (pp. 14-28). Cambridge University Press. <https://doi.org/10.1017/CBO9781139833882.004>
- Ne'matova, D., & Hakima, A. (2025). Differences between Written and Spoken Corpora. *Models and Methods in Modern Science*, 4(16), 93-95. <https://doi.org/10.5281/zenodo.17974462>
- Sinclair, J. (2004). Corpus and Text – Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow Books. Retrieved June 11, 2026, from <https://users.ox.ac.uk/~martinw/dlc/chapter1.htm>
- Sinclair, J. (2004). *Trust the Text: Language, Corpus, and Discourse*. Routledge.
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333-339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- Szmrecsanyi, B., & Engel, A. (2023). A variationist perspective on the comparative complexity of four registers at the intersection of mode and formality. *Corpus Linguistics and Linguistic Theory*, 19(1), 79-113. <https://doi.org/10.1515/cllt-2022-0031>
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing Company. <https://doi.org/10.1075/scl.6>
- Tran, T. N., & Ly, C. K. (2025). *Integrating Information Technology in Foreign Language Education: Theoretical Foundations and Pedagogical Applications*. Nhà xuất bản Đại học Sư Phạm TP Hồ Chí Minh.
- Wynne, M. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books. Retrieved June 11, 2026, from <https://users.ox.ac.uk/~martinw/dlc/index.htm>
- Xiao, R. Z. (2008). Theory-driven corpus research : using corpora to inform aspect theory. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (Vol. 1). Mouton de Gruyter.
- Zhao, Q., & Nazir, S. (2022). English Multimode Production and Usage by Artificial Intelligence and Online Reading for Sustaining Effectiveness. *Mobile Information Systems*, 2022, 6780502. <https://doi.org/10.1155/2022/6780502>