World Journal of Corpus Linguistics
Vol. 3, No. 1 - 2024, Page 56 - 64
Available online at https://pusatpublikasi.com/index.php/corpus/

CORPUS LINGUISTICS IN DICTIONARY CONSTRUCTION AND LANGUAGE TEACHING: A SYSTEMATIC REVIEW OF CURRENT TRENDS AND APPLICATIONS

Otomosi Gea
Universitas Nias, Nias, Indonesia
E-mail: ottogea2@gmail.com

Abstract
Corpus linguistics has transformed lexicography and language pedagogy through empirical, data-driven approaches. This systematic review examines recent developments in corpus-based methodologies for dictionary construction and language teaching between 2023-2025. Using systematic literature review methodology, this study analyzed 10 key publications from major linguistics journals to identify current trends, challenges, and future directions. Findings reveal that corpus-based approaches significantly enhance dictionary accuracy and language teaching effectiveness, though implementation challenges persist regarding teacher training and resource accessibility. The study identifies three primary application areas: corpus-driven lexicography utilizing NLP tools, data-driven learning in EFL contexts, and corpus-based teacher education. Results indicate growing integration of computational techniques with corpus linguistics, though gaps remain in making corpus methods accessible to mainstream educators. This review provides insights for researchers, lexicographers, and language educators seeking to leverage corpus methodologies in their practice.

Keywords: corpus linguistics, lexicography, language teaching, data-driven learning, systematic review.

INTRODUCTION

Corpus linguistics represents a fundamental paradigm shift in linguistic research and applied linguistics, moving from intuition-based approaches to empirical analysis of authentic language data. Over the past decades, this methodology has profoundly impacted two critical domains: dictionary construction and language teaching. The field has experienced remarkable growth since the early 2000s, driven by advances in computational technology and the proliferation of digital text collections (Hilpert, 2024).

The application of corpus linguistics to lexicography dates back centuries, though systematic corpus-based approaches emerged more recently (Götz, S., etc., 2024). Historical records show that corpus methodologies were employed as early as 1775 for dictionary development, with the Oxford English Dictionary representing a landmark corpus-based project in the nineteenth century (Mair, 2024). Contemporary lexicography now routinely employs sophisticated corpus analysis techniques, utilizing Natural Language Processing (NLP) tools to process massive text collections and extract patterns relevant for dictionary entries (Maachi & Khamar, 2025).

In language pedagogy, corpus-based instruction has gained traction as researchers and educators recognize the value of exposing learners to authentic language patterns. Data-driven learning (DDL), introduced by Johns in 1991, encourages learners to discover linguistic patterns through guided exploration of corpus data (Poehner & Lu, 2024). Recent studies demonstrate that corpus-based approaches enhance various language skills while promoting learner autonomy and metalinguistic awareness (Crosthwaite, 2024).

Despite these advances, significant gaps remain in understanding optimal implementation strategies and addressing practical challenges. Teacher education programs struggle to adequately prepare educators for corpus-based pedagogy, and many teachers report insufficient confidence in using corpus tools (Leńko-Szymańska, 2025). Furthermore, while substantial research focuses on English language contexts, corpus applications in other languages remain underexplored.

This systematic review addresses these gaps by synthesizing recent research on corpus linguistics applications in lexicography and language teaching. The study investigates three primary research questions: (1) How are corpus methodologies currently applied in dictionary construction? (2) What are the documented effects of corpus-based instruction on language learning outcomes? (3) What challenges impede wider adoption of corpus methods in language education, and what solutions have been proposed?

RESEARCH METHODS

Research Design

This study employs systematic literature review methodology to synthesize current research on corpus linguistics applications in lexicography and language teaching. The review follows established protocols for systematic reviews in applied linguistics, focusing on recent high-quality publications that address the research questions.

Search Strategy and Data Sources

The literature search was conducted in October 2024 using multiple academic databases and search engines including Web of Science, Google Scholar, and publisher-specific platforms (De Gruyter, Routledge, Springer, Wiley, Cambridge Core). The search strategy employed combinations of keywords including "corpus linguistics," "lexicography," "dictionary construction," "language teaching," "data-driven learning," "corpus-based instruction," and "teacher education." To capture the most current developments, the search was limited to publications from 2023-2025. This timeframe was selected to focus on contemporary trends and recent innovations in corpus methodology and application. The search prioritized peer-reviewed journal articles, edited volumes from major academic publishers, and systematic reviews published in high-impact linguistics journals.

Inclusion and Exclusion Criteria

Publications were included if they met the following criteria:

- Published between 2023-2025
- Focused on corpus linguistics applications in lexicography or language teaching
- Published in peer-reviewed journals or by reputable academic publishers
- Written in English
- Provided empirical data, theoretical frameworks, or systematic reviews relevant to the research questions Publications were excluded if they:
- Focused exclusively on computational linguistics without pedagogical or lexicographic applications
- Did not directly address corpus methodologies
- Were conference abstracts, working papers, or non-peer-reviewed materials
- Focused on languages or contexts with insufficient detail for meaningful synthesis

Data Extraction and Analysis

From each included publication, the following information was extracted:

- Publication details (authors, year, journal/publisher)
- Research focus and objectives
- Methodologies employed
- Key findings and contributions
- Implications for practice
- Identified challenges and limitations

The extracted data was analyzed thematically, with findings organized around three main themes: (1) corpus applications in lexicography, (2) corpus-based language teaching and learning, and (3) teacher education and implementation challenges. Within each theme, patterns, trends, and recurring findings were identified and synthesized.

Quality Assessment

All included publications underwent quality assessment based on criteria including methodological rigor, clarity of reporting, relevance to research questions, and contribution to the field. Publications from high-impact journals with rigorous peer review processes were prioritized. Studies with clear methodology, adequate sample sizes (where applicable), and explicit reporting of limitations received higher quality ratings.

Limitations

This review has several limitations. First, the focus on recent publications (2023-2025) provides current insight but may miss important foundational work. Second, limiting the search to English-language publications may introduce linguistic and geographic bias. Third, while systematic in approach, the relatively small number of included publications reflects the specific focus on very recent work. Finally, rapid technological advancement means some findings may quickly become outdated, necessitating ongoing review and synthesis.

RESULTS AND DISCUSSION

Corpus Linguistics in Dictionary Construction

Integration of NLP Tools in Lexicography

Recent research demonstrates significant advances in integrating Natural Language Processing tools with corpus linguistics for dictionary construction. Maachi and Khamar (2025) investigated the development of school lexicography using corpus-based approaches combined with NLP tools. Their study showed that textbook-based corpora, when analyzed using computational techniques, enable creation of contextually relevant and pedagogically appropriate dictionaries for learners.

The effectiveness of this approach lies in its ability to process large volumes of authentic educational texts, extracting vocabulary that students actually encounter in their learning materials. This contrasts with traditional dictionary development, which often relies on expert intuition rather than systematic analysis of learner-relevant texts. The integration of NLP tools automates many labor-intensive aspects of corpus analysis, including part-of-speech tagging, collocation extraction, and frequency analysis.

Mair (2024) provides a comprehensive review of digital corpora in language study, characterizing corpus linguistics as a "success story" in recent linguistics research history. He notes that excellent corpus resources now exist for major world languages, though significant gaps remain for less-resourced languages. The availability of these digital corpora has fundamentally changed lexicographic practice, enabling evidence-based documentation of language use at unprecedented scales.

Word Selection Methodologies

A critical aspect of corpus-based lexicography involves establishing systematic criteria for word selection in specialized wordlists. Contemporary research has identified nine primary word selection criteria: frequency, range, specialized occurrence, dispersion, expert judgment, dictionary checking, keyness, discipline measure, and ratio. Each criterion employs specific parameters and threshold settings that determine which words merit inclusion in specialized dictionaries or wordlists.

Frequency analysis measures how often words appear across a corpus, providing basic information about word importance. However, frequency alone proves insufficient, as words may be frequent in limited contexts but rare in others. Range analysis addresses this limitation by examining how widely distributed words are across different texts or subcorpora. Dispersion measures provide additional nuance, assessing whether words appear evenly throughout a corpus or cluster in specific sections.

These sophisticated selection criteria enable lexicographers to create specialized dictionaries that accurately represent vocabulary relevant to specific domains, proficiency levels, or learning purposes. For instance, academic wordlists developed using these criteria help learners prioritize vocabulary essential for academic success, while specialized technical dictionaries can focus on terminology specific to particular professional domains.

Corpus-Based Language Teaching and Learning

Theoretical Frameworks

Recent scholarship has explored theoretical foundations for corpus-based language teaching, particularly examining connections between corpus linguistics and second language acquisition theory. Poehner and Lu (2024) investigate the intersection of sociocultural theory and corpus-based English language teaching. They argue that corpus linguistics provides valuable resources for concept-based language instruction (C-BLI), helping teachers understand learner language abilities and providing authentic examples that illustrate target linguistic concepts.

Their framework proposes three phases where corpus resources prove beneficial: (1) assessing learners' current understanding of target concepts through learner corpus analysis, (2) providing authentic examples from native speaker corpora that illustrate target concepts, and (3) tracking learner development over time through repeated corpus analysis. This integration of sociocultural theory with corpus methods offers a principled approach to corpus-based pedagogy grounded in established learning theory.

Lu (2023) provides comprehensive coverage of corpus linguistics applications in second language acquisition research. His work synthesizes decades of research examining how corpus-based approaches inform understanding of interlanguage development, acquisition sequences, and factors influencing second language learning outcomes. This body of research demonstrates that corpus methodologies contribute both to theoretical understanding of language learning processes and to practical pedagogical applications.

Effectiveness of Corpus-Based Instruction

Empirical evidence increasingly supports the effectiveness of corpus-based instruction for language learning. Crosthwaite (2024) presents diverse perspectives on corpus-assisted language learning, bridging research-practice gaps. His edited volume showcases international applications of data-driven learning using tools like AntConc, WordSmith Tools, and

CorpusMate. Studies included in this volume demonstrate that corpus-based approaches enhance learners' understanding of collocations, grammatical patterns, and discourse features.

Specific benefits identified across multiple studies include improved accuracy in using multi-word expressions, enhanced awareness of register variation, and better understanding of authentic language patterns that differ from textbook presentations. Learners exposed to corpus data develop stronger metalinguistic awareness, becoming more conscious of patterns in language use and better able to notice and learn from input.

However, research also reveals that corpus-based instruction proves most effective with intermediate and advanced learners. Beginning learners may find concordance lines overwhelming and struggle to identify relevant patterns in corpus data. This suggests the need for carefully scaffolded approaches that introduce corpus methods gradually, perhaps beginning with simplified corpus data or heavily guided exploration activities.

Construction Grammar and Corpus Analysis

Hilpert (2024) examines the convergence of corpus linguistics, historical linguistics, and construction grammar.

Since the early 2000s, researchers have increasingly employed corpus-based methods to investigate language change from constructionist perspectives. This work employs sophisticated analytical techniques including collostructional analysis, multivariate modeling, distributional semantic analysis, and network analysis.

Li, Szmrecsanyi, and Zhang (2024) exemplify this approach in their diachronic corpus study of Chinese theme-recipient constructions. Using corpus data spanning from the 14th to 20th centuries, they demonstrate how corpus methods can reveal long-term patterns of grammatical change (Biber, D., etc., 2024). Their multivariate analyses show that while some conditioning factors remain stable over time, others exhibit significant shifts, and individual constructions follow distinct developmental trajectories.

These studies illustrate how corpus linguistics contributes to theoretical linguistics by providing empirical evidence for or against proposed theoretical models. The ability to test theoretical predictions against large-scale authentic language data strengthens linguistic theory while identifying phenomena that require theoretical refinement.

Teacher Education and Implementation Challenges

Corpus Literacy in Teacher Education

Leńko-Szymańska (2025) addresses the critical issue of teacher education for corpus-based pedagogy. She identifies three primary approaches to integrating corpora into teacher education programs: (1) using corpora as tools for developing teachers' language awareness, (2) providing pedagogically-focused instruction on data-driven learning methods, and (3) offering comprehensive corpus linguistics courses.

Each approach offers distinct advantages and limitations. Using corpora for language awareness helps teacher trainees recognize patterns in their target language, potentially improving their own linguistic competence.

However, this approach may not adequately prepare teachers to implement corpus methods in their own classrooms. Pedagogically-focused instruction emphasizes practical applications but may lack theoretical depth. Comprehensive courses provide thorough grounding but require substantial time commitments that may not be feasible in all teacher education programs.

Research on the effectiveness of these approaches faces methodological challenges. Studies typically involve small sample sizes, rely heavily on self-reported data, and rarely track long-term impacts on classroom practice after teachers complete their training. Available evidence suggests that well-designed training positively influences teachers' corpus literacy and attitudes, though training alone may not suffice to ensure sustained implementation in practice.

Barriers to Adoption

Despite demonstrated benefits, corpus use among language teachers remains limited, particularly in primary and secondary education contexts. Multiple barriers impede wider adoption. First, many teachers lack adequate training in corpus linguistics and report insufficient confidence using corpus tools. The technical aspects of corpus consultation can seem daunting, particularly for teachers without strong computational skills.

Second, existing corpus tools often feature steep learning curves and interfaces designed for researchers rather than practitioners. Concordancers like AntConc, while powerful, require users to formulate explicit search queries and interpret dense concordance displays. Teachers facing substantial workload pressures may lack time to develop these skills, even when motivated to explore corpus-based approaches.

Third, resource constraints limit access to appropriate corpora and corpus tools. While major corpora exist for widely-taught languages like English, teachers of less commonly taught languages may find few suitable resources. Even for English, specialized corpora matching specific teaching contexts may not exist or may require institutional subscriptions beyond reach of under-resourced schools.

Fourth, curriculum constraints and standardized testing regimes may discourage adoption of corpus-based approaches. Teachers facing pressure to cover prescribed content and prepare students for specific examinations may perceive corpus exploration as too time-consuming or insufficiently aligned with assessed competencies.

Proposed Solutions

Researchers have proposed several solutions to address these barriers. Crosthwaite (2024) emphasizes developing user-friendly tools specifically designed for pedagogical purposes. Examples include ColloCaid, which provides corpus-based collocation suggestions integrated into writing environments, and Write & Improve, which offers corpus-informed feedback on learner writing. These tools make corpus insights accessible without requiring users to master traditional concordancing software.

Leńko-Szymańska (2025) advocates for creating accessible resource banks containing ready-to-use corpus-based materials and activities. Such resources reduce preparation time while providing models that teachers can adapt to their contexts. Open educational resource initiatives have produced valuable materials, including step-by-step guides for creating corpus-informed lessons and repositories of corpus-based activities.

Teacher education programs must also evolve to better prepare educators for corpus-based pedagogy. This includes not only technical training on corpus tools but also pedagogical training on how to integrate corpus activities effectively into curricula. Ongoing professional development opportunities enable practicing teachers to develop corpus literacy incrementally rather than requiring mastery before any classroom implementation.

Emerging Trends and Future Directions

Integration with Artificial Intelligence

The intersection of corpus linguistics with artificial intelligence and large language models presents both opportunities and challenges. Large language models trained on massive corpora encode patterns of language use that corpus linguists have traditionally studied through explicit analysis. However, the opacity of these models raises questions about how their "knowledge" relates to corpus-derived linguistic insights.

Future research must investigate how AI technologies can be leveraged to make corpus insights more accessible while maintaining the empirical grounding that characterizes corpus linguistics. Potential applications include AI-powered tools that provide learners with corpus-based feedback on their language production, adaptive systems that select appropriate corpus examples based on learner needs, and intelligent tutoring systems that scaffold learner exploration of corpus data.

Multilingual and Cross-Linguistic Perspectives

While English dominates corpus linguistics research, expanding corpus-based approaches to other languages represents a critical priority. This includes developing corpora for under-resourced languages, creating parallel and comparable corpora for cross-linguistic research, and investigating whether pedagogical approaches successful in English contexts transfer to other language learning situations.

Multilingual corpus research can illuminate language-universal and language-specific patterns, informing both linguistic theory and language pedagogy. The development of learner corpora for multiple languages enables comparative research on second language acquisition processes across different target languages and learner first language backgrounds.

Corpus-Based Genre Pedagogy

Growing interest in corpus-based genre pedagogy reflects recognition that effective communication requires not just grammatical accuracy but also understanding of genre conventions. Corpus analysis can reveal how linguistic features pattern in different genres, informing explicit instruction on disciplinary and professional discourse.

Researchers have begun developing corpus-based pedagogical approaches that help learners understand rhetorical moves, lexical bundles, and grammatical patterns characteristic of specific genres. This work connects corpus linguistics with genre analysis traditions, offering learners explicit knowledge of conventions governing important text types in academic and professional contexts.

CONCLUSION

This systematic review of recent corpus linguistics research reveals significant advances in both theoretical understanding and practical applications. In lexicography, integration of NLP tools with corpus methodologies has enhanced dictionary construction, enabling evidence-based documentation of language use at unprecedented scales. Contemporary lexicographers employ sophisticated word selection criteria, utilizing frequency, range, dispersion, and other measures to create specialized dictionaries that accurately reflect vocabulary relevant to specific domains and learner needs.

In language teaching, corpus-based instruction demonstrates effectiveness in enhancing various language competencies while promoting learner autonomy and metalinguistic awareness. Theoretical frameworks connecting corpus linguistics with second language acquisition theory provide principled foundations for corpus-based pedagogy. Studies document improvements in learners' use of collocations, grammatical structures, and genre-appropriate language when instruction incorporates corpus data.

However, significant implementation challenges persist. Teacher education programs must better prepare educators for corpus-based pedagogy, addressing both technical corpus literacy and pedagogical skills for integrating corpus activities into curricula. The field needs continued development of user-friendly tools that make corpus insights accessible without requiring extensive technical expertise. Resource development efforts should prioritize creating ready-to-use corpus-based materials that teachers can readily adapt to their contexts.

Several promising directions emerge for future research. First, investigation of how artificial intelligence technologies can be leveraged to make corpus methods more accessible while maintaining empirical grounding represents an important frontier. Second, expansion of corpus-based approaches to under-resourced languages and multilingual contexts will broaden the field's impact and enable cross-linguistic research. Third, development of corpus-based genre pedagogies that help learners master conventions of important academic and professional text types offers practical applications with clear learner benefits.

The integration of corpus-based approaches into mainstream language teaching remains incomplete, with significant gaps between research findings and classroom practice. Bridging this gap requires sustained effort across multiple fronts: continued innovation in tool development, expanded teacher education, creation of accessible resources, and institutional support for teachers implementing corpus-based approaches. As these challenges are addressed, corpus linguistics will increasingly fulfill its potential to transform language teaching and learning through evidence-based, data-driven approaches grounded in authentic language use.

REFERENCES

Crosthwaite, P. (Ed.). (2024). *Corpora for language learning: Bridging the research-practice divide*. Routledge. https://doi.org/10.4324/9781003413301

Götz, S., & Granger, S. (2024). Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research*, *10*(1), 1–38.

Hilpert, M. (2024). Corpus linguistics meets historical linguistics and construction grammar: How far have we come, and where do we go from here? *Corpus Linguistics and Linguistic Theory*, *20*(3), 481–504. https://doi.org/10.1515/cllt-2024-0009

Leńko-Szymańska, A. (2025). Teacher education for pedagogical uses of corpora. In *Handbook of language teacher education*. Springer. https://doi.org/10.1007/978-3-031-51447-0_89-1

Li, Y., Szmrecsanyi, B., & Zhang, W. (2024). Beyond dynasties and binary alternations: A diachronic corpus study of four-way variability in Chinese theme-recipient constructions. *Folia Linguistica*, *58*, 221–255. https://doi.org/10.1515/flin-2023-2026

Lu, X. (2023). *Corpus linguistics and second language acquisition: Perspectives, issues, and findings*. Routledge.

Maachi, H., & Khamar, H. (2025). The contribution of corpus linguistics and natural language processing tools in the development of school lexicography. In B. Hdioud & S. L. Aouragh (Eds.), *Arabic language processing: From theory to practice* (pp. 47–66). Springer. https://doi.org/10.1007/978-3-031-80438-0_4

Mair, C. (2024). Digital corpora in language study: Reviewing a success story in the recent history of linguistics research. *Research in English Language Pedagogy*, *12*(3), 469–477.

Poehner, M. E., & Lu, X. (2024). Sociocultural theory and corpus-based English language teaching. *TESOL Quarterly*, *58*(3), 1256–1263. https://doi.org/10.1002/tesq.3282

Biber, D., Douglas, L., Tove, L., & Hancock, G. R. (2024). The linguistic organization of grammatical text complexity: Comparing the empirical adequacy of theory-based models. *Corpus Linguistics* *and Linguistic Theory*, *20*(2), 347–373. https://doi.org/10.1515/cllt-2024-0008