JOURNAL OF ADVANCED LINGUISTIC STUDIES
VOL. 12, NO. 2, JUL-DEC 2025 (ISSN 2231-4075)

Advantages and Limitations of Corpus-Based Discourse Analysis: A Case Study of Learner Conversations with AI Chatbots

NARGIZA ISOMITDINOVNA ASRorova
Uzbek State World Languages University, Uzbekistan

ABSTRACT

The integration of corpus linguistics and discourse analysis has given rise to corpus-based discourse analysis (CBDA), a methodological synergy that offers powerful tools for investigating language use in context. This article examines the advantages and limitations of CBDA through a case study of English as a Second Language (L2) spoken conversations with AI chatbots. Drawing on a specialized corpus compiled from interactions between Uzbek learners and two AI interlocutors, the study demonstrates how CBDA enables the identification of recurring syntactic-lexical patterns, pragmatic strategies, and prosodic features that might elude traditional qualitative discourse analysis. Key advantages include enhanced objectivity through quantitative validation, scalability for handling large datasets, and the ability to triangulate findings, thereby reducing researcher bias. However, the study also highlights significant limitations of CBDA such as failure to capture the full contextual and paralinguistic dimensions of discourse, challenges in annotating and analyzing pragmatic phenomena automatically, and constraints by the representativeness and design of the corpus itself. The article concludes that while CBDA is an indispensable methodology for profiling the unique discourse genre of human-AI interaction, researchers must take into account of its inherent constraints, advocating for a mixed-methods approach that complements computational rigor with qualitative, context-sensitive interpretation.

Keywords: Corpus-based discourse analysis, learner corpora, AI chatbots, L2 speech, method.

INTRODUCTION

Discourse analysis (DA) is the study of language use beyond the sentence level and its embeddedness in social context and it has traditionally relied on qualitative, interpretative methods (Gee, 2014). While rich in detail, such approaches can be limited by researcher subjectivity and the challenge of generalizing from small-scale data. Concurrently, corpus linguistics (CL) has emerged as a field dedicated to the analysis of large, systematically compiled collections of electronic texts, employing quantitative methods to identify patterns of language use (McEnery & Hardie 2012). For some time, these two fields were seen as epistemologically divergent; DA focused on the functional and contextual aspects of language, while CL was often critiqued for treating text as a decontextualized product (Widdowson 2000).

However, the last two decades have witnessed a fruitful convergence, leading to the development of corpus-based discourse analysis (CBDA). CBDA leverages the empirical, data-driven strengths of CL to ground and enrich the insights of DA, creating a more robust and replicable framework for studying language in use (Baker et al. 2008). As Flowerdew (2023) notes, this integration allows researchers to "uncover linguistic patterns and their frequencies which often go unnoticed during the traditional manual discourse analysis." This synergy is particularly valuable in applied linguistics, where understanding the nuances of language acquisition and use requires both detailed description and empirical validation.

A compelling and contemporary domain for applying CBDA is the analysis of second language (L2) learner discourse, especially, in technologically mediated environments. The proliferation of AI-powered chatbots has created a new genre of discourse: real-time, spoken interaction between human learners and artificial conversational agents. Platforms like ChatGPT and specialized tools like Flow developed bySpeechmatics are increasingly used by L2 learners for language practice, generating vast amounts of spoken data (Koç&Savaş 2025; Rassaei 2023). This context presents a unique opportunity and

challenge for discourse analysts. How do learners adapt their language, interactional strategies, and pragmatic competence when the interlocutor is a non-human AI? What are the distinctive features of this AI-mediated discourse genre?

Answering these questions necessitates a method that can systematically identify patterns across numerous interactions while still attending to the functional and contextual meanings of language. CBDA is ideally suited for this task. It allows researchers to compile a dedicated learner corpus, annotate it for various linguistic features, and use computational tools to identify trends that inform a broader discourse analysis of the interactional dynamics at play.

This article explores the advantages and limitations of CBDA through the lens of a specific case study: a corpus-based discourse analysis of L2 spoken conversations between Uzbek learners of English and AI chatbots. The case study is drawn from a doctoral dissertation that designed, compiled, and analyzed a specialized corpus for this purpose (Asrorova, 2025). By reflecting on the processes and findings of this study, we will critically evaluate the capacity of CBDA to provide meaningful insights into this novel communicative context, while also acknowledging the methodological boundaries it encounters. The central argument is that while CBDA dramatically enhances the scope, objectivity, and empirical grounding of discourse studies, its effectiveness is contingent on a critical awareness of its limitations, particularly regarding context, pragmatics, and the very nature of the data it can process.

MATERIALS AND METHODS

The study adopted a contextual, bottom-up approach to CBDA. Recognizing that existing large-scale corpora (e.g., BNC, COCA) did not contain AI-mediated learner conversations, the researcher compiled a specialized, small-scale spoken learner corpus.The discourse community consisted of university-level L2 English learners in Uzbekistan. A pre-task survey established participant identities, ensuring they were frequent users of AI chatbots for language practice. The final corpus included 54 speakers,

predominantly female (89%), with L1 backgrounds in Uzbek (95%), and proficiency levels ranging from B1 to C1 on the CEFR scale. Naturally occurring spoken conversations between the learners and AI chatbots, primarily flow by Speech matics and ChatGPT were recorded. The Flow application was used for its integrated ASR and conversational AI capabilities. A total of 54 samples were collected, transcribed, and compiled into a single corpus of approximately 13,450 words. Audio recordings were preserved for prosodic analysis.

The corpus underwent a multi-stage annotation and analysis process using a suite of computational tools. Stanford TreeTagger and the Penn Treebank Tagset were used for part-of-speech tagging and lemmatization. AntConc was used to generate frequency lists, n-grams, and collocations. The Lexical Complexity Analyzer (Ai & Lu, 2010) was employed to measure lexical density, sophistication, and variation. The L2 Syntactic Complexity Analyzer (Lu, 2010) was used to calculate 14 indices of syntactic complexity. Praaline was used to mark and analyze dysfluencies (pauses, fillers, repetitions, false starts) and repair strategies. Pragmatic analysis required qualitative manual coding due to the limitations of automated tools. The analysis covered conversational acts, discourse markers using Schiffrin's framework, politeness strategies using Brown & Levinson's framework, and register variation.

The annotated corpus was then subjected to discourse analysis. Using a bottom-up method, the text was first segmented into Vocabulary-Based Discourse Units (VBDUs) using the TextTiling program, which identifies topic shifts based on lexical cohesion (Biber, Connor & Upton 2007). These units were then analyzed for language patterns, subjected to a move analysis using a custom-designed rubric, and investigated for interactional strategies related to face, framing, and politeness.

RESULTS AND DISCUSSION

The CBDA approach yielded a rich, multi-layered profile of L2 learner discourse in AI chatbot conversations: Regarding Syntactic-lexical patterns, the analysis revealed a discourse

characterized by "formulaic chunking" and "syntactic scaffolding." Learners heavily relied on formulaic expressions to frame topics and structure responses. N-gram analysis showed a high frequency of conversational bigrams and trigrams. Syntactic complexity indices indicated moderate sentence length and clause density, with a notable overuse of low-cost connectors like "and" and "because" for clausal linkage, a pattern typical of developing interlanguage.

Pragmatic and interactional strategies were analyzed based on the corpus findings. The politeness strategy analysis revealed a predominance of "bald-on-record" directive ssuggesting learners perceived the AI as a tool without a "face" to threaten. However, they also displayed pragmatic awareness by using positive politeness, as an instance, calling the AI "dude" or "friend" to create a comfortable learning environment and hedging to protect their own face when discussing limitations. Move analysis identified seven key rhetorical moves, with a very high frequency of "respond to the question" and a significant frequency of proactive "Frame the task" moves, indicating that learners actively directed the interaction rather than passively responding.

Prosodic analysis quantified a high frequency of dysfluencies. Fillers were the most common, occurring 185 times, followed by repetitions in 70 instances and pauses in 45 instances. Repair strategies were overwhelmingly self-initiated and self-completed (SISR), highlighting a learning environment where learners felt free to monitor and correct their own output without external pressure.

The register was found to be a hybrid, unstable mix of informal, spoken language such as colloquialisms, contractions, fragments and occasional formal, academic features consisting of structured discourse markers, impersonal constructions. Learners frequently shifted between identities, acting as a "director" – giving commands, a "peer" – engaging in casual chat, an "expert" – sharing knowledge, and a "confidant" – revealing personal information which showcases the AI as a "pedagogical sandbox" for identity exploration.

The application of CBDA in the case study above underscores several profound advantages this methodology holds over traditional, purely qualitative approaches to discourse. The highlighted advantage of CBDA is reduced researcher bias. One of the most significant criticisms of traditional discourse analysis is its susceptibility to researcher subjectivity. Interpretations can be influenced by the analyst's own beliefs, theoretical commitments, or selective attention to data that confirms pre-existing hypotheses. CBDA directly addresses this issue by anchoring claims in quantitative, empirical evidence derived from the entire corpus. As Biber et al. (1998) argue, corpus analysis "provides a solid basis for generalizations about language use" (p. 4). In the context of the case study, it allowed the researcher to move from stating that learners "often" use formulaic language to specifying the exact n-gram distributions that constitute this "often," thereby enhancing the validity and reliability of the discourse analysis.

Identification of patterns and frequencies allow for more accurate results. The human brain is not optimized for accurately perceiving frequency distributions across large datasets. We tend to notice what is salient or unusual, often overlooking subtle but pervasive patterns. Computational corpus tools excel at this task, revealing linguistic patterns that are statistically significant but may be invisible to the naked eye. The tools used in the case study, such as AntConc and the Lexical Complexity Analyzer, automatically identified the most frequent words, collocations, and n-grams. This revealed that the discourse was built on a foundation of conversational formulae and scaffolds. Without frequency lists and n-gram analysis, the researcher might have noted a few instances of "I think that" or "there is/are," but the corpus tools demonstrated that these were not isolated occurrences but central, structural elements of the learners' discourse.

This advantage is what Gumperz (1982) alludes to when he states that CBDA focuses on the "thread of language... used in the situation network" (p. 85). It allows the analyst to see the recurring "threads" – the lexical bundles, syntactic frames, and discourse markers – that weave the fabric of the discourse

community's communication. This profiling is essential for understanding the linguistic "fingerprint" of a specific genre or community.

Moreover, CBDA holds an upper hand in scalability and handling of large datasets**.** While the case study involved a "small-scale" corpus of 13,450 words, the methodologies employed are inherently scalable. The same annotation and analysis pipeline could, in principle, be applied to a corpus ten or a hundred times larger with minimal additional manual effort. This scalability is a cornerstone of corpus linguistics and a major advantage of CBDA.

As research on AI chatbot interactions grows, researchers will likely compile much larger corpora. Manual analysis of thousands of conversations would be prohibitively time-consuming and inconsistent. CBDA, however, can process vast amounts of data consistently. Automated POS taggers, syntactic parsers, and n-gram retrievers operate with the same rules regardless of corpus size. This allows for research that is both deep and broad, capable of identifying macro-trends across a global population of learners while still drilling down to specific linguistic features.

This scalability also promotes replicability and comparability. Another researcher can apply the same tools and methods to a different learner-AI corpus, allowing for cross-contextual comparisons (e.g., comparing Uzbek learners with Korean learners) and the testing of hypotheses on a wider scale. CBDA allows for triangulation of research, the use of multiple methods or data sources to study a phenomenon, which is a key principle of rigorous qualitative research. CBDA inherently facilitates triangulation by combining quantitative and qualitative approaches. In the case study, the research design was a mixed-methods approach where quantitative corpus findings informed and were interpreted through qualitative discourse analysis.

This synergy addresses the critique that corpus data is "decontextualized." The qualitative discourse analysis provides the necessary context to interpret the quantitative patterns. As a result, the findings are more credible, well-rounded, and robust. The corpus data curbs the potential for speculative qualitative

interpretations, while the qualitative analysis ensures that the numbers are meaningfully connected to communicative functions and speaker intentions. The empirical findings generated by CBDA are not merely of academic interest; they have direct practical applications. The detailed profile of learner discourse generated in the case study offers valuable insights for language pedagogy and the design of educational technology.

Despite its powerful advantages, CBDA is not a methodological panacea. The case study also brings into sharp focus several inherent limitations that researchers must acknowledge and navigate. Perhaps the most enduring critique of corpus-assisted methods is their struggle to fully capture context. As Widdowson (2000) asserted, computers can only analyze the "textual traces" of discourse, not the dynamic process of meaning negotiation. While the qualitative component of CBDA mitigates this, the core corpus data often lacks crucial contextual information.

A Corpus-Based Discourse Analysis (CBDA) of transcribed spoken conversations, while powerful, provides only a partial picture of the communicative event. The methodology is inherently limited in capturing crucial contextual layers such as non-verbal communication, multimodality, and the nuanced emotional information conveyed by prosody (Ai & Lu 2010). Furthermore, computational tools struggle with the automatic annotation of pragmatic phenomena, which are central to discourse analysis. While tools can tag syntax and lexis, they cannot reliably identify speech acts or politeness strategies, forcing researchers to rely on manual, qualitative interpretation. This re-introduces an element of subjectivity and underscores that CBDA efficiently handles the "what" of language but stumbles on the "why" and "how" of its use (Brown & Levinson 1987).

Finally, CBDA faces considerable practical hurdles. The methodology depends on a complex suite of software tools, each with its own technical requirements and learning curve, and some high-performance tools are not freely available (Lu 2014). More critically, the process is immensely time-consuming. Although automated components are scalable, the essential tasks of manual

error-checking and qualitative coding are not, placing a practical limit on the depth and scale of analysis possible within the constraints of a typical research project.

CONCLUSIONS

Corpus-based discourse analysis represents a significant methodological advancement in the study of language in use. As demonstrated through the case study of L2 learner interactions with AI chatbots, CBDA offers a powerful set of advantages: it grounds discourse analysis in empirical evidence, thereby reducing researcher bias; it reveals recurrent linguistic patterns that are otherwise difficult to discern; it is scalable and capable of handling large datasets; it facilitates the triangulation of quantitative and qualitative findings; and it generates knowledge with direct pedagogical and technological relevance.

However, a critical application of CBDA must also confront its limitations. The methodology struggles to capture the full richness of contextual and paralinguistic information, it is poorly equipped to automatically handle the nuances of pragmatics, it can impose a static view on dynamic discourse, its findings are constrained by corpus design, and it faces practical computational and resource challenges.

The key to harnessing the power of CBDA lies in recognizing it not as a replacement for qualitative discourse analysis, but as its essential partner. The two approaches exist in a dialectical relationship. The corpus provides the "what" – the broad, empirical landscape of language use. The discourse analyst provides the "why" and "how" – the interpretative, context-sensitive understanding of that landscape. The case study exemplified this synergy: the quantitative data on politeness strategies was meaningless without the qualitative interpretation of the "AI politeness paradox," and the qualitative observations about learner agency were given weight by the quantitative frequency of topic-initiation moves.

For researchers embarking on the study of new discourse genres, particularly in digitally mediated environments like human-AI communication, CBDA is an indispensable

methodology. It provides the rigorous, data-driven framework needed to map this uncharted territory. However, its application must be guided by methodological humility. Researchers must be transparent about the constraints of their corpora, deliberate in their use of mixed methods, and vigilant in complementing computational findings with deep, qualitative interpretation. In doing so, CBDA will continue to evolve as a robust and critical tool for understanding the complexities of discourse in the modern world.

REFERENCES

Ai, H. & Lu, X. 2010. A web-based system for automatic measurement of lexical complexity. *CALICO Symposium 2010*, Amherst, MA, United States. Available online: <https://aihaiyang.com/software/lca/>.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T. & Wodak, R. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19/3, 273-306. Available online: <https://doi.org/10.1177/0957926508088962>.

Biber, D. 1993. Representativeness in corpus design. Literary and *Linguistic Computing*, 8/4, 243-257.

——., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

——., Connor, U. & Upton, T. A. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. John Benjamins Publishing.

Brown, P. & Levinson, S. C. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Flowerdew, L. 2023. Corpus-based discourse analysis. In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 275-289). 2nd ed. Routledge.

Gee, J. P. 2014. *An Introduction to Discourse Analysis: Theory and Method*. 4th ed. Routledge.

Granger, S. 2002. A bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3-33). John Benjamins Publishing.

Gumperz, J. J. 1982. *Discourse Strategies*. Cambridge University Press.

Hyland, K. & Paltridge, B. (Eds.) 2011. *Continuum Companion to Discourse Analysis*. Continuum.

IBM. 2021. What is a chatbot? IBM. 25 Jun 2024. Available online: <https://www.ibm.com/topics/chatbots>.

Jia, J. 2009. CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. Knowledge-Based Systems, 22/4, 249-255. Available online: <https://doi.org/10.1016/j.knosys.2008.12.001>.

Koç, F. Ş. & Savaş, P. 2025. The use of artificially intelligent chatbots in English language learning: A systematic meta-synthesis study of articles published between 2010 and 2024. *ReCALL*, 37/1, 4-21. Available online: <https://doi.org/10.1017/S0958344024000163>.

Lee, D. 2008. Corpora and discourse analysis: New ways of doing old things. In V. Bhatia, J. Flowerdew & R. H. Jones (Eds.), *Advances in Discourse Studies* (pp. 86-99). Routledge.

Lee, H. & Lin, Y. 2023. Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL*, 35/1, 48-64. Available online: <https://doi.org/10.1017/S0958344022000169>.

Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15/4, 474-496. Available online: <https://doi.org/10.1075/ijcl.15.4.02lu>.

——. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45/1, 36-62. Available online: <https://doi.org/10.5054/tq.2011.240859>.

——. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96/2, 190-208. Available online: <https://doi.org/10.1111/j.1540-4781.2011.01232.x>.

——. 2014. *Computational Methods for Corpus Annotation and Analysis*. Springer.

McEnery, T. & Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

——. Xiao, R. & Tono, Y. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. Routledge.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24/4, 492-518. Available online: <https://doi.org/10.1093/applin/24.4.492>.

Rassaei, E. 2023. The effects of text-based and audio-based dynamic glosses on L2 vocabulary learning: A dynamic assessment approach. *The Language Learning Journal*, 51/4, 509-522. Available online: <https://doi.org/10.1080/09571736.2021.2025420>.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge University Press.

NARGIZA ISOMITDINOVNA ASRAROVA
PHD RESEARCHER,
UZBEK STATE WORLD LANGUAGES UNIVERSITY,
TASHKENT, UZBEKISTAN.
E-MAIL: <N.ASRAROVA@TIFT.UZ>