The European Journal of Literature and Linguistics 2025, No 3–4
ISSN 2310-5720

PREMIER Publishing
ppublishing.org

DOI:10.29013/EJLL-25-3.4-16-19

CORPUS-BASED DISCOURSE ANALYSIS: RULES, CONDITIONS AND RESEARCH APPLICATIONS

Dilrabo Kosheva [1]

[1] National Pedagogical University

Cite: *Kosheva D. (2025). Corpus-Based Discourse Analysis: Rules, Conditions and Research Applications. European Journal of Literature and Linguistics 2025, No 3–4.* https://doi.org/10.29013/EJLL-25-3.4-16-19

Abstract

Corpus-based discourse analysis has become a central methodological approach in modern linguistics, allowing researchers to examine naturally occurring language with empirical accuracy. Unlike traditional qualitative readings, corpus methods enable large-scale, replicable and quantifiable discourse interpretation. This article explores theoretical foundations of discourse analysis and demonstrates how corpus-driven methodology strengthens interpretive validity through representativeness, transparency, reproducibility and annotation.

Keywords: *modern linguistics, discourse interpretation, discourse research, corpus linguistics, discourse analysis, communication, linguistic material, cognitive environment*

Introduction

Discourse analysis, as a field, seeks to explore language beyond individual sentences by examining context, speaker intention and social interaction. With digitization and advancements in textual databases, corpus linguistics offers new possibilities for discourse research. Instead of relying on small manually collected samples, scholars now have access to millions of words of authentic spoken and written text – transcribed, tagged and searchable with software. Discourse, as a concept, has received extensive interpretation across linguistic scholarship, resulting in a rich plurality of definitions that reflect the evolution of language study. Early foundational work by Harris (Harris, Z., 1952) described discourse as language beyond the sentence, emphasizing continuity, cohesion, and meaning that emerges only when linguistic units are examined in extended stretches. Building on this structural perspective, Hymes (Hymes, D., 1972) shifted attention toward discourse as contextualized language use, suggesting that understanding communication requires consideration not only of linguistic form but also of how, when, and why utterances are produced. Later, van Dijk expanded the view further by framing discourse as the dynamic interaction of text and context, highlighting the inseparability of linguistic material from the social and cognitive environment in which it occurs. Fairclough (Fairclough, N., 1992), in turn, introduced a critical dimension by conceptualizing discourse as a form of social practice, arguing that every discursive event is shaped by – and simultaneously contributes to – underlying power relations and institutional structures. Together, these perspectives illustrate that discourse cannot be reduced to textual material alone; rather, it is a multidimensional phenomenon that bridges language, context, interaction, and society. A humorous classroom dialogue illustrates discourse as social interaction where meaning emerges pragmatically rather than solely syntactically:

Student: *"Can I be punished for something I didn't do?"*

Teacher: *"Of course not."*

Student: *"Good, because I didn't do my homework."*

The unintended pragmatic effect – manipulation of presuppositions – is visible only when discourse is treated contextually, not at sentence level. Why use corpora in discourse analysis? Corpora expand analytical precision by providing large, representative data instead of researcher's intuition. As the seminar materials note, corpora are pre-compiled, annotated, and allow empirical confirmation of patterns that may otherwise remain unnoticed. Scholars such as Sinclair, Biber et al. (Biber, D., Conrad, S., & Reppen, R., 1998), and McEnery and Hardie (Simpson, R., Briggs, S., Ovens, J., & Swales, J., 2002) highlight several key advantages that make corpora indispensable in contemporary linguistic research. First, corpus-based methods provide exceptional scale and efficiency, enabling researchers to access thousands or even millions of words with minimal effort compared to manual data collection. This broad access naturally enhances representativeness, reducing researcher bias and ensuring that conclusions are drawn from diverse and balanced samples rather than isolated excerpts. A third benefit is replicability, as corpora offer transparent documentation of data origin and analytical procedure, allowing other scholars to verify results and reproduce studies under the same conditions. Additionally, the annotation depth of many corpora – including part-of-speech tagging, speaker metadata, and discourse markers – facilitates detailed, layered analysis that would be difficult to achieve manually. Finally, corpora enable comparability across registers, genres, and even languages, opening opportunities for large-scale contrastive studies.

Materials and methods

The research procedure for analyzing evaluative discourse was conducted using the Michigan Corpus of Academic Spoken English (MICASE), selected for its rich representation of authentic university-level interactions, including lectures, seminars, advising sessions, and student discussions. To obtain a manageable yet representative dataset, ten transcripts were extracted from the corpus, yielding approximately 144,000 words of raw spoken material. These files were then combined into a single working text and subjected to an initial processing stage, during which timestamps, speaker labels, and paralinguistic noise markers such as [laughter] or [pause] were removed. This cleaning stage aimed to retain only linguistically relevant content, producing a raw corpus suitable for computational interrogation. The refined text was analyzed using AntConc 4.3.1, a concordance and frequency tool selected for its accessibility and compatibility with discourse-oriented research. A set of evaluative lexical items – *good, clear, excellent, should, might, maybe, and think* – was defined as search targets based on their relevance to academic assessment and feedback practices. Frequency lists were first generated to determine the distribution of these items across the corpus, after which collocation and key-word-in-context (KWIC) searches were performed to reveal recurrent syntagmatic patterns. Subsequent linguo-pragmatic interpretation focused on how these items functioned within interactional sequences, particularly in relation to praise and hedging strategies used to mitigate criticism. Results were summarized in a table linking keywords to collocations and discourse functions, demonstrating two dominant evaluative strategies within academic speech: assessment through positive adjectives (good, clear, excellent) and face-saving feedback through modal verbs and hedging expressions (should, might, maybe).

Results and discussion

The computational analysis conducted through AntConc provided valuable insights into how evaluative language is distributed and pragmatically mobilized within academic interaction. One of the most noteworthy observations is that raw frequency alone only partially reflects evaluative tendencies; it is the patterns and environments in which lexical items occur that reveal their discourse function. The terms *good* and *clear,* for example, appeared frequently in the frequency list, yet frequency output did not explain how or why they were used. It was the KWIC concordance lines, generated through AntConc, that made interactional purpose visible, highlighting clusters such as *good point, good example, clear explanation, and clear idea*. These collocational structures suggest that evaluation is rarely expressed in isolation – it is embedded within feedback sequences that validate student contributions and highlight comprehension. Similarly, modal verbs *(should, might, could, maybe)* did not surface as dominant items numerically, but KWIC retrieval revealed their importance as subtle facilitators of hedged instruction. Instead of direct imperatives such as *change this* or *you must*, instructors frequently opted for softened alternatives: *you could develop this further, maybe try expanding this point, you might want to clarify this example*. The repeated appearance of such phrases indicates a preference for feedback delivery that is supportive rather than authoritative.

Another layer of the discussion lies in the sequential organization of evaluative discourse. AntConc analysis showed that positive adjectives often occurred before hedged suggestions, forming what could be described as a pedagogical praise–advise cycle. KWIC lines frequently revealed structures such as *that was a good point – maybe you could support it with data, or your explanation is very clear – you might want to add one more example here.* This sequencing strengthens the claim that evaluation in academic speech is not performed as binary judgement but rather as a collaborative meaning-building process. The corpus thus demonstrates a consistent use of positive reinforcement as an interactional strategy preceding corrective guidance, creating space for improvement without threatening the learner's self-perception.

Furthermore, the use of AntConc allowed for the identification of evaluative density – clusters of keywords appearing in close textual proximity. In discussions and seminar transcripts in particular, the software highlighted stretches of discourse containing multiple evaluative markers within short spans. For instance, lines such as *I think that's a really good idea, maybe you could expand it a bit*, represent dense evaluative packaging, combining stance expression *(I think)*, praise *(good) and* hedging *(maybe)* within a single feedback move. These dense clusters rarely appear in written academic texts, reinforcing the distinction between spoken and written evaluation conventions. Spoken feedback appears to prioritize immediacy and encouragement, whereas written feedback often favors precision, structure, and reduction of redundancy.

The findings generated through AntConc analysis therefore demonstrate that evaluative discourse in academic settings is multidimensional: it merges linguistic assessment with interpersonal management, oscillating between affirmation and cautious suggestion. The corpus revealed evaluation not as a simple act of approval or disapproval, but as a complex pragmatic negotiation that protects student agency while simultaneously guiding performance. The methodological implication here is clear: quantitative outputs such as frequency lists cannot independently reveal discursive attitudes – it is through concordance inspection, collocation mapping, and sequential interpretation that evaluative discourse becomes fully visible. AntConc thus functions not merely as a tool for counting words, but as an instrument for uncovering patterns of interaction, stance, and power dynamics in spoken educational settings.

Conclusion

In summary, the case study demonstrates that corpus-supported discourse analysis offers quantifiable evidence for pragmatic interpretation, allowing researchers to move beyond isolated examples and instead trace recurring patterns across hundreds of authentic interactions. This methodological shift substantially increases analytical reliability and opens new possibilities for comparative work, including contrasts between English and Uzbek academic discourse, variation in teacher versus student evaluative strategies, and exploration of gender- or register-based differences in feedback behaviour. Importantly, the study highlights a notable research gap – the absence of a large-scale English–Uzbek parallel corpus – which presents a valuable direction for future scholarship. Overall, corpus-based discourse analysis serves as an effective bridge between qualitative interpretation and quantitative verification: by integrating frequency statistics, collocational mapping and contextual function analysis, researchers obtain a multidimensional view of discourse as social action. The MICASE analysis confirms that such tools can uncover evaluative and interactional tendencies that traditional close reading may overlook, and as corpus resources continue to grow, their role in shaping discourse research, comparative linguistics and pedagogical application is likely to become increasingly influential.

References

Anthony, L. (2022). AntConc (Version 4.3.1) [Computer software]. Waseda University. URL: https://www.laurenceanthony.net/software/antconc

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

Fairclough, N. (1992). Discourse and social change. Polity Press.

Harris, Z. (1952). Discourse analysis. Language, – 28(1). – P. 1–30.

Hymes, D. (1972). Models of the interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), Directions in sociolinguistics (p. 35–71). Blackwell.

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

Simpson, R., Briggs, S., Ovens, J., & Swales, J. (2002). The Michigan Corpus of Academic Spoken English (MICASE) [Corpus]. The Regents of the University of Michigan. URL: https://micase.eecs.umich.edu

submitted 19.11.2025;
accepted for publication 03.12.2025;
published 19.12.2025

Contact: science_7777@mail.ru