# Do we still need corpora and corpus analysis platforms? Discourse analysis in times of LLMs

**Julia Krasselt, Dolores Lemmenmeier-Batinić, Philipp Dreesen**

Zurich University of Applied Sciences, Institute of Language Competence
Theaterstrasse 17, 8400 Winterthur
{krss, leme, dree}@zhaw.ch

**Abstract**

Corpus-based discourse analysis investigates the linguistic construction of societally shared knowledge by iterating between quantitative pattern detection and qualitative interpretation in large text collections. Large Language Models (LLMs) promise to lower practical barriers to such work (e.g., natural-language querying, qualitative coding), yet they also introduce risks that are especially consequential in discourse-analytic settings, where fluent summaries can encourage ungrounded interpretation. This position paper argues that integrating LLMs into corpus analysis platforms is appropriate only insofar as it remains compatible with three epistemic premises of corpus research: (1) transparency of the data basis and traceability of analytical operations; (2) interpretability as evidence-constrained sense-making; and (3) seriality and patternedness as distributional structure and variation. In this opinion paper, we contribute a platform-oriented requirements perspective that translates these premises into design constraints for tool-calling/RAG-style integration, and we outline implementation directions that treat LLMs as an interaction layer over inspectable corpus retrieval and platform-based analysis.

**Keywords:** corpus-based discourse analysis, corpus platforms, large language models (LLMs)

## 1. Background and Motivation

Corpus-based discourse analysis approaches the study of societally shared knowledge by iterating between quantitative pattern detection and qualitative interpretation of language use (lexical choices, argumentative patterns, stance-taking, etc.) across large text collections (cf. Baker, 2023; Baker & McEnery, 2015; Bubenhofer, 2009). Over the past two decades, this approach has been adopted across the humanities and social sciences, including public health sciences, political science, and media studies (e.g., (Grimmer & Stewart, 2013; Krasselt et al., 2022; O'Halloran, 2010), and it also informs applied communication tasks such as discourse-informed message design and stakeholder-oriented communication (Cooren, 2015).

This development has been supported by advances in corpus infrastructure. Curated corpora and platform ecosystems (web-based and local) have made large-scale discourse analysis accessible beyond corpus linguistics. Tools such as Sketch Engine (Kilgarriff et al., 2014), AntConc (Anthony, 2024), and #LancsBox X (Brezina & Platt, 2025) as well as initiatives such as ParlaMint (Erjavec et al., 2023), the Leipzig Corpus Collection (Goldhahn et al., 2012) and Swiss-AL (Krasselt et al., 2023) exemplify this expansion of accessible data and methods.

At the same time, effective corpus-based discourse analysis remains demanding because it is inherently iterative. Robust studies move between quantitative indicators and qualitative inspection—for example, by relating collocation profiles to concordance evidence (Baker, 2023)—and require consequential decisions about discourse modelling, corpus construction, analytical settings, and criteria of interpretive relevance. Difficulties often arise where statistical outputs must be connected to defensible discourse claims and where multiple analytical steps need to be integrated into a coherent, documentable workflow (cf. also McEnery & Brezina, 2022).

Large Language Models (LLMs) may reduce some of these frictions and are beginning to shape expectations about how corpus resources can be accessed and explored by researchers (Brezina, 2025). They can support natural interaction with corpora, assist with query formulation, and guide users through analytical options (e.g., Anthony, 2025; AI integration on english-corpora.org). However, current research demonstrates substantial risks, particularly when generative AI is used for qualitative analysis. Studies report that semantic categorisation of keywords is often generic – especially when items are presented without context – shows only marginal overlap with human-produced categorisations, and may even introduce fabricated assignments during the categorisation process. Reproducibility is an additional concern, given the non-deterministic behaviour of general-purpose models (Curry et al., 2024; Gillings et al., 2024; Morgan, 2023).

Incentives for speed and simplification are even stronger outside academia. Democratic societies depend on the formation of public opinion, which makes the analysis of public meaning-making processes attractive not only for research but also for public-facing monitoring and communication. Organizations such as political parties, public authorities, associations, and NGOs therefore have strong motivations to seek fast "insights" into what can be said, by whom, and with which effects. A current risk is that LLMs will be adopted for these tasks because they generate fluent, plausible-sounding accounts on demand, while concealing data choices and interpretive steps. In contrast, corpus-based discourse analysis provides suitable data, methods and tools for



*Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026*, pages 67–72

producing more accountable analyses of public meaning-making.

Against this background, we return to the question raised in the title: Do we still need corpora and corpus analysis platforms? We argue for a conditional integration of LLM-based workflows into corpus platforms for discourse analysis. Here, *conditional* means that such integration must remain compatible with the epistemic logic of corpus analysis. By *corpus platforms* we mean integrated, access-controlled environments that stabilise corpus definitions and metadata, support concordance-level inspection, and log platform-side computations and parameters. We propose three epistemic premises: (1) corpus workflows must ensure transparency of the data basis and traceability of analytical operations; (2) interpretation must be supported by inspectable concordance-level evidence and triangulation across measures; and (3) claims must be grounded in seriality and patternedness, evidenced by distributions and structured variation, for example across time, genres, and languages. Building on these premises, we discuss which forms of LLM integration can support corpus-based discourse analysis while remaining anchored in accountable procedures and transparent evidence.

## 2. Epistemic premises of corpus analysis

### 2.1 Transparency and Traceability

In corpus analysis, transparency and traceability are essential for justifying interpretations and claims based on empirical evidence (Baker, 2023: 225; Bednarek et al., 2024). Transparency concerns the data basis: it should be clear which corpus (or which parts of it) were analysed and on what textual instances a claim rests. Traceability concerns the process: analytical steps and settings (from preprocessing to statistical measures) should be documented so that results can be reconstructed when the same procedures are applied again. Together, these requirements support interpretability as making sense of patterns in light of a research question (see Section 2.2).

A central advantage of corpora is that they are explicitly delimited and enriched with metadata (e.g., time, genre, speaker, region), enabling controlled subsetting and auditing. This differs from LLMs, whose training data and selection principles are generally not inspectable at the level of individual documents and whose outputs do not provide provenance for the instances that would support a claim (Heersmink et al., 2024). Consequently, LLM responses cannot function as evidence on their own.

Transparency and traceability translate into platform-level documentation of corpus modelling, processing, and analysis. Platforms should record corpus composition and metadata, key processing choices (e.g., deduplication, tokenization, tagging/lemmatization, language identification), and analytical parameters (e.g., measures, thresholds, window sizes, dispersion metrics). Because tools often present aggregated outputs (e.g., keyword or collocation lists), transparency also requires drill-down to concordance lines and contextual evidence.

LLM integration introduces specific risks for transparency. Interfaces may silently reformulate queries or apply undocumented defaults, and generated summaries can remain detached from the empirical evidence if they are not explicitly linked to the retrieved data and the relevant quantitative outputs. For this reason, LLM functionality should be constrained by corpus retrieval and platform-side calculation: if implemented, summaries should be based on computed results, and any claim should link back to concordance-level evidence and distributional information so that users can reconstruct how it was derived.

### 2.2 Interpretability

In corpus-based discourse analysis, interpretability is a central epistemic requirement because analyses substantiate claims about how shared understandings are distributed and become prevalent in society (Keller, 2024). Since such objects are inseparable from social context and may inform public debate or institutional practice, it is not sufficient that an analysis yields plausible statements; interpretations must be justified in relation to a research question and constrained by the available evidence.

Interpretation denotes the methodological step of relating distributional observations (e.g., collocation profiles, concordance patterns, shifts over time) to an analytical focus and, where relevant, to an applied problem. What counts as salient and how a pattern is understood depends on the question and theoretical perspective; accordingly, there is rarely a single “correct” reading. Interpretability in discourse analysis is thus supported by practices that keep such justification empirically and analytically constrained. This includes contextualisation (reading patterns against relevant co-text and situational knowledge), testing whether an observation is stable or driven by particular sources or periods, and triangulation across alternative operationalisations and measures (Baker, 2023: 44; Bednarek, 2009; Marchi & Taylor, 2009). Equally important is attention to structured variation: differences across genres, time periods, or speaker groups are often analytically central and should remain visible rather than being collapsed into a single narrative.

LLM-assisted analysis introduces specific risks at this interpretive stage. Models may hallucinate

examples or explanations not supported by the corpus, and fluent summaries can invite narrative overreach (Ji et al., 2023). Conversational interaction may also amplify confirmation dynamics and encourage premature closure, smoothing analytically consequential variation across time, genres, or groups.

The issue of interpretability also becomes evident when collecting and evaluating ethically sensitive data or data that is protected by copyright, for example. Discourse corpora often contain copyrighted material and potentially sensitive data (e.g., political extremism), the processing of which requires specialist knowledge. Corpus platforms can provide secure research environments and transparent, documented data-handling procedures that generic chat interfaces layered on top of an LLM typically do not. In sum, the interface is changing, but the need for accountable, corpus-based infrastructures for discourse analysis remains.

### 2.3 Seriality, Patternedness, and Distributional Grounding

In corpus-based discourse analysis, the epistemic goal is not the description of isolated events but the identification of seriality and patternedness in public discourse (Foucault, 1981, 1982). Discourses become socially powerful not because a statement occurs once, but because formulations, evaluations, and argumentative moves recur across texts and sediment into shared assumptions about how the world can or should be understood. This repetition structures what appears normal, plausible, or sayable in society. From a corpus perspective, seriality becomes observable as patterned language use (e.g., recurring lexical choices, frames, metaphors, actor configurations, or stance profiles) that can be shown to be stable or systematically shifting across time, contexts, and communities.

Importantly, discourse-analytic seriality is not simply frequency. A pattern is discourse-relevant when it exceeds the idiosyncrasies of single authors, outlets, or formats and when it appears as a structured distribution across the discursive field. This includes, for example, patterns that are dispersed across sources rather than concentrated in a few texts; patterns that recur across genres while taking genre-specific forms; or patterns that differentiate speaker groups, political camps, or linguistic communities. Seriality therefore combines recurrence with structured variation.

Corpora and corpus platforms support the identification of seriality by making such distributional questions testable. They allow researchers to delimit the discursive space through corpus design and metadata, detect candidate patterns via aggregated measures (e.g., keywords, collocations, association profiles), and inspect instances through concordance evidence and contextual reading. Combined with transparency and traceability, this enables analysts to evaluate whether an observation is genuinely serial by checking dispersion across sources, stability across subcorpora or time periods, and sensitivity to operationalisation (e.g., alternative queries, reference corpora, or window settings).

LLM-assisted workflows pose particular risks to establishing seriality and patternedness, especially when fluent, abstractive summaries become the primary analytic output. Such summaries can smooth heterogeneity and obscure analytically relevant variation. Majority bias may privilege dominant frames and paraphrases, masking marginal positions. In addition, LLM outputs often neglect dispersion and distribution, making it difficult to distinguish broadly shared patterns from event-driven spikes or source-specific effects. In the worst case, the epistemic target shifts from demonstrating seriality as distributional structure to producing plausible accounts of "what the discourse is about." Conversely, when LLMs are used to support corpus-anchored tasks (e.g., schema-driven, custom annotation, candidate retrieval, or consistency checks across coded instances, cf. Yu et al., 2024) they may strengthen rather than weaken the establishment of patterned variation, provided results remain traceable and are validated through distributional analysis.

## 3. Implications for the integration of LLMs into platforms for corpus-based discourse analysis

### 3.1 Criteria for LLM integration

Building on the epistemic premises outlined above, we propose the following criteria for LLM integration into corpus platforms for discourse analysis:

**Outputs must remain corpus-grounded and evidence-linked.** LLM assistance should operate on an explicitly defined corpus and maintain a clear separation between model text and corpus evidence; analytic statements should link to inspectable contexts (concordances/documents) and the relevant computed outputs.

**Workflows must be traceable and reversible.** Platforms should log and export executed queries, preprocessing state, parameters, and statistical measures, and make any automated query reformulations or defaults explicit and reversible rather than silently applied.

**Computation must remain platform-side.** Quantitative results (e.g., keywords, collocations, dispersion) should be computed by the platform; the LLM may guide exploration and explanation but must not substitute for, or fabricate, analytical results.

**Interpretation must be supported without narrative smoothing.** LLM interaction should encourage contextual inspection, counterevidence, and alternative framings, while preserving structured variation and enabling checks of dispersion and stability across time, genres, groups, and languages.

**Hallucination resilience is required.** When the available evidence is insufficient, the system should signal these limits and prompt further retrieval and inspection rather than producing confident, ungrounded claims.

### 3.2 Implementation directions

These criteria matter because corpus-based discourse analysis advances by iteratively testing candidate patterns against textual evidence—a process that is time-consuming and cognitively demanding. Consider studies of argumentative structures in thematic discourses (so-called *topoi*, Wengeler, 2012). Identifying topoi typically involves inspecting keywords, n-grams, and collocates, reading concordances (and, where necessary, full texts), and iteratively building an annotation scheme until higher-order categories stabilise (Kalwa, 2013). The process produces many false positives, since only a small subset of retrieved items is relevant to the analytical aim. LLMs seem to offer a shortcut by answering questions like "Which topoi appear in discourse X?" even before a corpus is delimited. Yet such outputs are not discourse analysis but plausible summaries that bypass corpus definition, distributional testing, and evidence inspection.

Corpus platforms remain central in the LLM era because they stabilise the data basis, document preprocessing and statistical procedures, and preserve access to concordance-level evidence and structured variation. A robust integration strategy therefore treats the LLM primarily as an interaction layer, while retrieval and computation are executed by specialized platform functions and returned in inspectable form. This division of labour is increasingly reflected in tool-calling architectures (e.g., the Model Context Protocol, MCP and related function/tool-calling approaches), which standardise how models invoke external tools and incorporate their outputs (Anthropic, 2024; Hou et al., 2026).

Accordingly, two implementation directions are particularly compatible with the criteria outlined above. First, LLMs can be used to post-process platform results via predefined prompt templates— for instance to sort, filter, cluster, or categorise concordance lines or aggregated outputs (cf. Davies, 2025), while preserving links to evidence and keeping operations reversible. Second, Retrieval-Augmented Generation (RAG, cf. Lewis et al., 2020; Gao et al., 2023) can enable natural-language interaction with user-defined corpus slices by retrieving relevant passages or documents first and using them as grounded context for the model's response. In contrast to "pure" generation, RAG and tool-calling workflows make the evidential basis explicit and allow outputs to be tied back to retrievable sources. In both cases, the decisive requirement is that AI support remains constrained by corpus retrieval and platform-provided results, and that outputs preserve traceability, evidence access, and distributional visibility rather than replacing them with narrative closure.

## 4. Conclusion

Do we still need corpora and corpus platforms for discourse analysis in the age of LLMs? We argue that we do – because LLMs change the interface to text analysis without removing the epistemic requirements that make discourse-analytic claims accountable. Corpus-based discourse analysis relies on (i) transparency of the data basis and traceability of analytical operations, (ii) interpretability as evidence-constrained sense-making supported by contextualisation and triangulation, and (iii) the demonstration of seriality and patternedness as structured distributions and variation. These premises define what counts as a defensible discourse-analytic result.

The societal stakes sharpen this point. Discourses shape shared assumptions and orders of speech, influencing what becomes sayable and legitimate. This makes discourse analysis a particularly sensitive domain: LLMs can generate fluent, plausible accounts while obscuring data choices and interpretive steps, and are therefore likely to be adopted for public-facing "insights" into opinion formation and communicative dynamics. Without corpus-grounded procedures, such uses risk replacing evidence-based analysis with persuasive narrative.

For these reasons, corpus platforms may become more rather than less relevant by providing orientation amid fast access to unreliable knowledge. Their distinctive contribution is curated, documented corpora and robust views of distributions and structured variation with inspectable concordance evidence. The criteria proposed here translate these premises into design requirements: LLMs may support interaction and exploration, but results must remain corpus-grounded, workflows reconstructable, and variation visible. Tool-calling architectures point to a division of labour in which the LLM acts as an interaction layer while specialized platform tools handle retrieval and quantitative analysis.

## 5. Bibliographical References

Anthony, L. (2024). *AntConc* (Version 4.3.1)

[Computer software]. Waseda University. https://www.laurenceanthony.net/software/AntConc

Anthony, L. (2025). Integrating AI technology into corpus-based language learning through *ChatAI*. *Computer Assisted Language Learning*, 1–19. https://doi.org/10.1080/09588221.2025.2589747

Anthropic. (2024). *Introducing the Model Context Protocol*. https://www.anthropic.com/news/model-context-protocol

Baker, P. (2023). *Using Corpora in Discourse Analysis* (2nd edition). Bloomsbury Academic. https://doi.org/10.5040/9781350083783

Baker, P., & McEnery, T. (Eds). (2015). *Corpora and Discourse Studies. Integrating Discourse and Corpora*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137431738

Bednarek, M. (2009). Corpora and Discourse: A Three-Pronged Approach to Analyzing Linguistic Data. In M. Haugh, K. Burridge, J. Mulder, & P. Peters (Eds), *Selected proceedings of the 2008 HCSNet workshop on designing the Australian national corpus: Mustering languages*.

Bednarek, M., Schweinberger, M., & Lee, K. K. H. (2024). Corpus-based discourse analysis: From meta-reflection to accountability. *Corpus Linguistics and Linguistic Theory*, *20*(3), 539–566. https://doi.org/10.1515/cllt-2023-0104

Brezina, V. (2025). Corpus linguistics and AI: #LancsBox X in the context of emerging technologies. *International Journal of Language Studies*, *19*(2). https://doi.org/10.5281/ZENODO.15250820

Brezina, V., & Platt, W. (2025). *#LancsBox X* [Computer software]. Lancaster University. http://lancsbox.lancs.ac.uk.

Bubenhofer, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. De Gruyter.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), 100082. https://doi.org/10.1016/j.acorp.2023.100082

Davies, Mark (2025). Comparing the predictions of Large Language Models to actual corpus data. (White papers). English-Corpora.org. https://www.english-corpora.org/ai-llms/

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., De Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., De Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0

Foucault, M. (1981). The order of discourse. In R. Young (Ed.), *Untying the text: A post-structuralist reader* (pp. 51–78). Routledge & Kegan Paul.

Foucault, M. (1982). *The archaeology of knowledge*. Pantheon Books.

Gillings, M., Kohn, T., & Mautner, G. (2024). The rise of large language models: Challenges for Critical Discourse Studies. *Critical Discourse Studies*, 1–17. https://doi.org/10.1080/17405904.2024.2373733

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Heersmink, R., De Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, *26*(3), 41. https://doi.org/10.1007/s10676-024-09777-3

Hou, X., Zhao, Y., Wang, S., & Wang, H. (2026). Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. *ACM Transactions on Software Engineering and Methodology*, 3796519. https://doi.org/10.1145/3796519

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kalwa, N. (2013). *Das Konzept 'Islam': Eine diskurslinguistische Untersuchung*. De Gruyter.

Keller, R. (2024). *The Sociology of Knowledge Approach to Discourse: Foundations, Concepts and Tools for a Research Programme*. Springer International Publishing. https://doi.org/10.1007/978-3-031-55114-7

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Krasselt, J., Dreesen, P., Fluor, M., & Rothenhäusler, K. (2023). Swiss-AL. Korpus und Workbench für mehrsprachige digitale Diskurse. In M. Kupietz & T. Schmidt (Eds), *Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022* (pp. 127–142). Narr Francke Attempto. 10.24053/9783823396024

Krasselt, J., Robin, D., Fadda, M., Geutjes, A., Bubenhofer, N., Suzanne Suggs, L., & Dratva,

J. (2022). Tick-Talk: Parental online discourse about TBE vaccination. *Vaccine*, *40*(52), 7538–7546. https://doi.org/10.1016/j.vaccine.2022.10.055

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds), *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Marchi, A., & Taylor, C. (2009). *If on a winter's night two researchers…: A challenge to assumptions of soundness of interpretation*. https://cris.unibo.it/handle/11585/792146

McEnery, T., & Brezina, V. (2022). *Fundamental Principles of Corpus Linguistics* (1st edn). Cambridge University Press. https://doi.org/10.1017/9781107110625

Morgan, D. L. (2023). Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. *International Journal of Qualitative Methods*, *22*, 16094069231211248. https://doi.org/10.1177/16094069231211248

O'Halloran, K. (2010). How to use corpus linguistics in the study of media discourse. In A. O'Keeffe & M. McCarthy (Eds), *The Routledge Handbook of Corpus Linguistics* (pp. 563–576). Routledge.

Wengeler, M. (2012). *Topos und Diskurs: Begründung einer argumentationsanalytischen Methode und ihre Anwendung auf den Migrationsdiskurs (1960-1985)*. Niemeyer. https://doi.org/10.1515/9783110913187

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, *29*(4), 534–561. https://doi.org/10.1075/ijcl.23087.yu