Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11

WYDAWNICTWO UMCS

ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN – POLONIA

VOL. XI SECTIO N 2026

ISSN: 2451-0491 • e-ISSN: 2543-9340 • CC-BY 4.0 • DOI: 10.17951/en.2026.11.77-86

# Corpus Studies in the Age of Generative AI: Uncertainty, Curation, and the Future of Language Materials

# Badania korpusowe w erze generatywnej sztucznej inteligencji: niepewność, selekcja i przyszłość materiałów językowych

UMCS

*Ferit Kılıçkaya*

Burdur Mehmet Akif Ersoy University
İstiklal Campus, 15200 Burdur, Türkiye
ferit.kilickaya@gmail.com
https://orcid.org/0000-0002-3534-0924

**Abstract.** The emergence of generative artificial intelligence calls for a rethinking of the assumptions underlying corpus-based research and pedagogy. This paper argues for a reconceptualization of corpus studies as a critical, reflexive, and socially situated practice. This entails viewing corpora not as inert containers but as multimodal and theoretically informed objects, curated with attention to provenance, modality, and communicative context. Pedagogically, greater transparency in the development of corpus-informed materials should be promoted so that learners understand issues of authorship, intentionality, and linguistic variation. Language education in the age of AI must move beyond communicative competence towards a critical awareness of how digital systems shape language use, cognition, and learners' socialization into various discourse communities globally.

**Keywords:** artificial intelligence; corpus; language materials; resources

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



**Abstrakt.** Pojawienie się generatywnej sztucznej inteligencji wymusza przemyślenie założeń leżących u podstaw badań i pedagogiki opartych na korpusach. Autor artykułu opowiada się za rekonceptualizacją badań korpusowych jako praktyki krytycznej, refleksyjnej i osadzonej w kontekście społecznym. Wymaga to ponownego wyobrażenia sobie korpusów nie jako statycznych repozytoriów, lecz dynamicznych, multimodalnych i opartych na teorii obiektów, selekcjonowanych z dbałością o pochodzenie, modalność i kontekst komunikacyjny. Z pedagogicznego punktu widzenia materiały oparte na korpusach muszą być opracowywane w sposób przejrzysty, zachęcający uczniów do analizowania autorstwa, intencjonalności i różnic językowych. Edukacja językowa w erze sztucznej inteligencji musi wykraczać poza kompetencje komunikacyjne i budować krytyczną świadomość tego, jak systemy cyfrowe kształtują użycie języka, poznanie i socjalizację uczniów w globalnych społecznościach dyskursywnych.

**Słowa kluczowe:** sztuczna inteligencja; korpus; materiały językowe; zasoby

## INTRODUCTION

There has been a paradigm shift in the production of texts and the use of language in the 2020s (Chapelle 2025). Generative artificial intelligence (AI) systems are at the core of this epistemic change and have become widely used to produce grammatically, stylistically, and contextually appropriate texts across virtually all known genres (Moorhouse, Wong 2025). Generative models are increasingly leveraged to produce texts across a wide range of discursive fields, in formats ranging from poetry and student essays to technical reports and academic articles, many of which were previously thought to be the preserve of human production. However, a particular challenge is that contamination of the corpora used for research and teaching can occur unnoticed. This paper's intervention is critical and proposes principled curation and textual genealogy as new standards of data integrity in lieu of AI detection. The ubiquity of generative AI, therefore, poses not only a practical challenge to the epistemology of corpus compilation but also a fundamental question about the very kind of language data that remains worth collecting, analyzing, and, ultimately, teaching (Boulton 2021).

Corpus linguistics as a field has largely relied on the assumption that the data analyzed in its research do indeed reflect "authentic" linguistic use, that is, language produced by humans in a real social environment for the purpose of communication (Gilmore 2019; Pérez-Paredes, Boulton 2025; Yoon, Jo 2014). The assumption also needs to be carefully reconsidered, given that it was always somewhat "permeable" to begin with. This epistemological foundation is further destabilized by the free circulation of AI-generated texts on various public platforms, some of which are used as data sources for corpus compilation. How, then, should authentic language use be defined? How "authentic" is a corpus

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



as a representation of a speech community when the "texts" it contains can be produced not only by humans but also by non-human agents? Could teaching materials extracted from such corpora end up perpetuating meaningless language use? These questions are not peripheral but central to the legitimacy of future corpus-informed pedagogical practices (Mishan 2005, 2017).

These dilemmas cannot be resolved by relying on traditional definitions or long-established practices. The concept of the "naturally occurring text" is no longer valid, an idea that long served as a cornerstone in corpus methodology. It is no longer in question that there are certain terminological issues that represent a paradigmatic rupture. This calls for a methodological and conceptual reconsideration of corpora in an age of digitally redefined semiotic space. Of course, this is not to suggest that paradigmatic shifts have never occurred before, since language has always been mediated by technological advancements, from the printing press to the smartphone. However, generative AI can now readily mimic patterns of human linguistic behavior with great fluency, thereby blurring the line between actual simulation and linguistic socialization.

This is the epistemological crisis in corpus linguistics brought about by generative AI that this position paper seeks to address. This paper also attempts to change the discussion from technical apprehension to a reflexive methodology of the digital age in a systematic manner. While generative AI impacts all facets of digital communication, for the scope of this discussion, its meaning is limited to corpus-informed language education and the methodological sanctity of linguistic data sets. The discussion here is strictly on the epistemological relocation from "authentic" (human-only) corpora to "stratified" corpora, which accepts hybridity in the making of texts.

To ensure analytical consistency, this paper operationalizes five core constructs: authenticity, intentionality, provenance, communicative context, and epistemic pluralism. Authenticity is not a binary state of "humanness" but interactional authenticity, the degree to which a text manifests situated, socially embedded purposes of communication. Operationalized intentionality is the distributed agency manifested in human–AI co-configurations that negotiate meanings in human geosocial prompting metonymically realized in algorithmic output. Provenance is the documentation of a textual genealogy of a text, tracing its point of origination – human, AI, or hybrid – to delivery or adding to a corpus. Communicative context is the sociotechnical context of textual production, from the generative models and prompts to editing interventions. Finally, epistemic pluralism is a view that dismisses the "purity" of the sole human corpus in favor of a stacked archive that accepts machine-mediated language as a valid, albeit different linguistic object of inquiry.

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



## THE EPISTEMOLOGICAL CRISIS: RETHINKING AUTHENTICITY AND INTENTIONALITY

The traditional reliance on "authentic" human text is fundamentally destabilized by the saturation of AI-produced content; thus, the concept of intentionality requires re-evaluation. This paper, therefore, defines intentionality as the purposeful, situated, and socially meaningful use of language. The multilingual teacher is unlikely to face difficulties in understanding AI-generated texts because such materials often lack contextual and communicative adequacy (Hymes 1972), despite their high degree of fluency. AI-generated texts may pass linguistic tests but fail to fully capture the social rules of language use when they are produced without sufficient grounding in interactional contexts. There is fluency, but this is ultimately a "synthetic" simulation of linguistic form rather than an intentional act by a speaker seeking to achieve a particular social outcome.

The blurring of authorship that occurs when humans subsequently edit AI-generated content raises the question of whether such entries can still be regarded as "authentic" or whether they instead introduce amplified distortions into corpus-based approaches to linguistic description. The authorship agnosticism invoked by the post-edited hybrid model presents the question of whether such "authentic" entries distort the descriptive potential of corpus-based realizations, which otherwise relied on principled curation during development, such as balanced sampling of genres, registers, and speaker populations, explicit annotation of metadata, and deliberate exclusion of non-representative content (Flowerdew 2012). Grants on such criteria are increasingly untenable in a digital ecosystem marked by human-machine co-authorship, often to an indistinguishable degree (Sardinha 2024).

For example, blog entries, once heralded as a pinnacle of immediate "digital word-of-mouth", are now increasingly AI-authored or at least AI-edited, while learner essays submitted through platforms with integrated writing-assistant functionalities may no longer be indicative of learners' actual linguistic competence (Marzuki, Widiati, Rusdin, Darwin, Indrawati 2023). Textbooks based on such corpora may thus end up normalizing language structures and collocations that, while grammatically correct, lack social or pragmatic contextualization. For instance, a language teacher may select model texts from an online review corpus without realizing that many of these reviews have been generated by AI systems optimized to promote sales. The resulting material can be a lexically rich and syntactically diverse text, but often, critically, it is detached from the sociocultural context that has imbued the language with its rich meanings. Therefore, it tends to be what has been described as "synthetic fluency", superficially fluent

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



and grammatically perfect but without the "communicative-pragmatic baggage" that comes with authentic language use. This is not simply a matter of textual surface feature accuracy of form, but also a matter of ethical teaching and learning languages outside real interlocution.

## METHODOLOGICAL INTERVENTION: PROVENANCE METADATA AND STRATIFIED CURATION

To treat computer-written texts and human-written copies as identical is to blur the distinction between language form (which can be reproduced, e.g. lexical items) and language function, namely communicative use and the meaning underlying the text (Fuertes Gutiérrez 2004), with theoretical implications. For these reasons, it would be necessary to reconsider term selection, not only in the specific sense of the operational procedures involved in preparing any given corpus, but also in terms of the epistemological stance it represents. Instead of treating any kind of copy, be it human-copied, computer-copied, or human--computer-copied, as identical, genealogical information should be added to the annotation, making each copy selectable and valuable in the process of research, study, preparation, and selection of teaching materials (Muñoz-Basols, Fuertes Gutiérrez 2025). Rather than seeking "purity", curation must acknowledge the hybridity of 21st-century language, which is increasingly mediated by non-human agents that simulate human-like participation in discourse.

In these circumstances, curation is both a technical and an ethical issue. In this context, it means a theoretically motivated, purposeful, and transparent selection, annotation, and maintenance of texts in a corpus. Not only should metadata regarding register, genre, and source be provided, but, importantly, metadata concerning the modality of authorship should also be included. Was the text human-written, co-written by a human and AI, or AI-written and subsequently human-edited? These distinctions are not only of theoretical interest but should also be addressed if the enterprise of corpus linguistics is to remain empirically valid and educationally useful.

What is urgently needed is a provenance metadata system that would enable texts incorporated into corpora to carry explicit markers of their origin. This, however, is easier said than done. Fewer platforms now require or display disclosure of AI-generated content, as such texts are becoming increasingly indistinguishable from human-authored ones. This may leave the responsibility for detection, disclosure, and verification to corpus compilers. This is not only a technical challenge but also a disciplinary one, requiring corpus studies to reassert their standards for data integrity and engage in a broader

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



conversation about what kinds of language use should be documented, modeled, and taught.

To move beyond mere theoretical advocacy, operationalizing provenance requires developing a standardized metadata schema that classifies each text by its "generative modality". That way, researchers and educators can filter datasets according to the extent of machine intervention. Metadata for "posthuman" corpora may include the elements presented in Table 1.

Table 1. Proposed metadata framework

| Metadata field | Description | Categorization values |
|---|---|---|
| Authorship mode | the primary agent of text creation | human-produced, AI-generated, AI-assisted (hybrid) |
| Generative tool | the specific system used to produce or assist the text | among others, GPT-4o, Claude 3.5, Gemini 1.5 |
| Intervention level | the extent of human editing post-generation | none, minimal (proofreading), heavy (substantive) |
| Verification method | the process used to confirm the text's origin | self-disclosure, AI detection, blockchain/watermark |

Source: Author's own elaboration.

UMCS

To ensure methodological clarity, these categories are operationalized according to the degree of agentive intervention in the creation of the text. This starts with human-authored materials, where the cognitive and linguistic labor remains exclusively human and digital assistance is limited to orthographic or grammatical checking. Moving toward a collaborative model, AI-assisted (hybrid) texts can be described as high-level conceptual prompting by a human, with substantive rewriting or fact-checking post-generation editing by the same person, even though the initial drafting is done by an AI system. AI-generated texts, on the other hand, embody the other extreme in this continuum, having been generated with a minimum of human input beyond selecting a prompt and exhibiting a complete absence of situational social accountability and a maximum degree of statistical typicality.

## RECONCEPTUALIZING THE CORPUS IN A POSTHUMAN ERA

In this emergent sense, corpus literacy must surpass the knowledge of querying for concordance lines and interpreting them, and must include a critical sense of data, evaluation of sources, and an appreciation of the sociotechnical infrastructural architectures that frame the conception of language production.

Pre- and in-service teachers need training not only in the use of corpora but also in how to scrutinize their contents, question their construction, and assess

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



their suitability for pedagogic use. The same applies to materials writers and curriculum designers, who must also become aware of the origins and communicative integrity of the language models which increasingly mediate access to corpora and facilitate the creation of multimodal texts (O'Halloran, Tan, Smith 2026). This is not an argument for the wholesale abandonment of AI-produced language but rather a call to epistemic pluralism, meaning an acceptance that more than one form of language production is available to us and is now interacting within our communicative ecologies.

Following Douglas Biber's (1988) framework for register analysis, which identifies variation based on situational constraints, AI-generated text emerges as a "disembodied" register. It is characterized by high lexical density and grammatical perfection, yet it lacks the register-specific markers of stance and interactional nuance typical of human-to-human communication. Treating AI as a unique register allows the field to systematically map its "stylistic fingerprints" against human-authored benchmarks. To operationalize this pluralism, corpus developers should move toward stratified corpora, i.e., a dataset composed of explicitly labeled sub-corpora for AI-generated, AI-assisted, and human-authored texts.

## PEDAGOGICAL REORIENTATION: CRITICAL DATA AWARENESS AND HYBRID LITERACY

The ubiquitous presence of AI-generated text in corpora poses a parallel crisis for language education, as datasets underpinning wordlists and textbook dialogues may be quietly inundated with statistically optimized but semantically weak AI-generated text. This risks ensuring that the usage patterns informing such pedagogical materials are pragmatically hollow, creating a form of "synthetic fluency" that is particularly evident in the travel blog genre, where the sensory markers and subjective qualia characteristic of human-authored narratives may be absent from AI-generated entries. Such entries may be grammatically impeccable yet lack the intangible qualities that mark a "register" as grounded in social realities and "interactional competence", particularly in the context of English for Academic Purposes.

In addition, when AI-generated texts are added to learner corpora, they pose the risk of misrepresenting the artificial fluency of algorithms as authentically developing interlanguage in learners. Not only would this lead to the distortion of accounts of developmental sequences, but it could also erode the empirical foundations for error analysis. As generative AI separates the author from the process of assistance, the authenticity of authorial intent becomes complex, and the "authenticity" of, e.g., a scientific document can no longer rely on the "humanness"

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



of prose alone. It can, instead, rely on the communicative soundness of a human-formulated prompt and on the ethical honesty of the validation of the text.

There may be materials in audio, video, or text formats that are beneficial and have been created with the assistance of AI (Kapp, Briskin 2025; Kern 2024). However, if learners are exposed solely or predominantly to AI-curated or AI-dominated inputs, the resulting materials may normalize syntactic and lexical forms that are not grounded in lived experience or shaped through interaction with actual interlocutors. More seriously, they may flatten the interpersonal and sociocultural dimensions of communication, reducing it to a series of predictable yet disembodied correct forms. Such an outcome is antithetical to the development of interactional competence, pragmatic nuance, and intercultural communicative intent as instructional goals in contexts such as English for Academic Purposes and English for Specific Purposes.

In this sense, pedagogical risk is not restricted to misrepresentation of form; it can also be misrepresentation of function – a learner internalizing polite requests issued by a chatbot and reproducing them without knowing that the linguistic form lacks pragmatic calibration against human interface, or expectations about coherence of narration, genre construction, and lexical frequency, etc. established by perusing an AI-generated corpus does not correspond to empirical human encounter, thus requiring an urgent recalibration of corpus-literacy, especially in teacher training programs.

Such a reconceptualization of the corpus project has implications both for pedagogy and for disciplinary identity: linguists and teachers understand themselves not only as users of corpora and, of course, providers of education, but also as curators of possibility-agents within a nexus of language, technology, and community: open, clean, and pedagogy-friendly. In this view of language, the education that such a society expects can neither exclude nor disregard these issues, if for no other reason than to engage in a meaningful scholarly discussion concerning (un)addressed issues of creative AI: the ethical implications of, and ultimately education about, communication systems of the 21st century should continue to be part of effective communication per se.

## CONCLUSIONS

Generative AI calls into question some of the core assumptions of corpus-based language research and teaching. This is not just a case of new technologies bringing new affordances, but an ontological paradigm shift that underlies our understanding of the production, circulation, and reception (or consumption) of linguistic entities. A new epistemological paradigm arises, and our empirical

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



understanding of language, especially with regard to the antagonism between the empirical objectives and resources of corpus linguistics, on the one hand, and the epistemic nature of machine-produced language, on the other, must be redefined. Our pedagogical practice, as well as our corpus research, now has to tackle the concepts of authenticity, authorship, and intentionality in order to cope with the rapidly increasing number of linguistic phenomena we collect in our corpora.

However, amid these tensions lies an instructive possibility: to reconceive corpus studies as a critical, reflexive, and socially oriented activity. This would mean moving from viewing corpora as static collections of language to understanding them as curated, multimodal, dynamic, and theoretically informed objects. It would require substantial methodological rethinking of the curation and documentation of texts, including accurate information about textual provenance and generative modality, a definition of communicative context, and an ethical awareness of how AI-generated texts can reshape learners' understandings of language use, cognition, and socialization within language communities.

From a pedagogical perspective, learners should critically interrogate the language and contexts in which it occurs, and the design of corpus-informed materials should be an intentional and transparent process. Educators and materials writers should be encouraged to be transparent about the texts they use, and learners should be able to engage with those texts in ways that enable them to develop their own conceptualizations of linguistic variation across contexts of authorship. The aims of language pedagogy should encompass both effective communication and a critical evaluation of the systems and structures that mediate 21st-century communication.

## ACKNOWLEDGEMENTS

The manuscript has been revised and linguistically refined with the assistance of digital tools, including Grammarly. The author would also like to express his sincere gratitude to the Polish National Agency for Academic Exchange (NAWA) for their support under the Ulam NAWA Program.

## REFERENCES

Biber, D. (1988). *Variation Across Speech and Writing.* Cambridge: Cambridge University Press.

Boulton, A. (2021). Research in Data-Driven Learning. In: P. Pérez-Paredes, G. Mark (Eds.), *Beyond the Concordance: Corpora in Language Education* (pp. 9–34). Amsterdam: John Benjamins. DOI: 10.1075/scl.102.01bou.

Pobrane z czasopisma Annales N - Educatio Nova **http://educatio.annales.umcs.pl**
Data: 22/11/2026 17:27:11



Chapelle, C.A. (2025). Generative AI as a Game-Changer: Implications for Language Education. *System*, 132. DOI: 10.1016/j.system.2025.103672.

Flowerdew, L. (2012). *Corpora and Language Education*. London: Palgrave Macmillan.

Fuertes Gutiérrez, M. (2004). El papel de los misioneros en la descripción de lenguas asiáticas por Lorenzo Hervás y Panduro (1735–1809). In: O. Zwartjes, E. Hovdhaugen (Eds.), *Missionary Linguistics: Selected Papers from the First International Conference on Missionary Linguistics, Oslo, 13–16 March 2003* (pp. 233–252). Amsterdam–Philadelphia: John Benjamins. DOI: 10.1075/sihols.106.17fue.

Gilmore, A. (2019). Materials and Authenticity in Language Teaching. In: S. Walsh, S. Mann (Eds.), *Routledge Handbook of English Language Teacher Education* (pp. 299–318). Abingdon–New York: Routledge.

Hymes, D.H. (1972). On Communicative Competence. In: J.B. Pride, J. Holmes (Eds.), *Sociolinguistics: Selected Readings* (pp. 269–293). Harmondsworth: Penguin.

Kapp, K.M., Briskin, J. (2025). Transformative Teaching with AI. In: J.R. Corbeil, M.E. Corbeil (Eds.), *Teaching and Learning in the Age of Generative AI: Evidence-Based Approaches to Pedagogy, Ethics, and Beyond* (pp. 49–76). New York: Routledge. DOI: 10.4324/9781032688602.

Kern, R. (2024). Twenty-First Century Technologies and Language Education: Charting a Path Forward. *Modern Language Journal*, 108(2), 515–533. DOI: 10.1111/modl.12924.

Marzuki, Widiati, U., Rusdin, D., Darwin, Indrawati, I. (2023). The Impact of AI Writing Tools on the Content and Organization of Students' Writing: EFL Teachers' Perspective. *Cogent Education*, 10(2). DOI: 10.1080/2331186X.2023.2236469.

Mishan, F. (2005). *Designing Authenticity into Language Learning Materials*. Bristol: Intellect.

Mishan, F. (2017). Authenticity 2.0: Reconceptualising Authenticity in the Digital Era. In: A. Maley, B. Tomlinson (Eds.), *Authenticity in Materials Development for Language Learning* (pp. 10–24). Newcastle upon Tyne: Cambridge Scholars Publishing.

Moorhouse, B.L., Wong, K.M. (2025). *Generative Artificial Intelligence and Language Teaching*. Cambridge: Cambridge University Press. DOI: 10.1017/9781009618823.

Muñoz-Basols, J., Fuertes Gutiérrez, M. (2025). Opportunities for Artificial Intelligence (AI) in Language Teaching and Learning. In: J. Muñoz-Basols, M. Fuertes Gutiérrez, L. Cerezo (Eds.), *Technology-Mediated Language Teaching: From Social Justice to Artificial Intelligence* (pp. 297–315). Bristol: Multilingual Matters. DOI: 10.21832/MUNOZ9872.

O'Halloran, K.L., Tan, S., Smith, B.B.A. (2026). Multimodal Approaches to English for Academic Purposes. In: K. Hyland, P. Thompson (Eds.), *The Routledge Handbook of English for Academic Purposes* (pp. 274–286). New York: Routledge.

Pérez-Paredes, P., Boulton, A. (2025). *Data-Driven Learning In and Out of the Language Classroom*. Cambridge: Cambridge University Press. DOI: 10.1017/9781009511384.

Sardinha, T.B. (2024). AI-Generated vs Human-Authored Texts: A Multidimensional Comparison. *Applied Corpus Linguistics*, 4(1). DOI: 10.1016/j.acorp.2023.100083.

Yoon, H., Jo, J.W. (2014). Direct and Indirect Access to Corpora: An Exploratory Case Study Comparing Students' Error Correction and Learning Strategy Use in L2 Writing. *Language Learning & Technology*, 18(1), 96–117. DOI: 10125/44356.