AI-Linguistics: A New Method or a New Paradigm? LLM-Based-, LLM-Driven- and LLM-Native-Linguistics

Daniel Słapek (Jagiellonian University in Kraków, Poland)

Submitted 23/09/2026, accepted 09/11/2026, published 12/11/2026
https://doi.org/10.62408/ai-ling.v6i1.107

daniel.slapek(at)uj.edu.pl

Abstract

In this paper, I deal with the growing role of Large Language Models (LLMs) in linguistic research, asking if their use represents just a methodological innovation or a broader paradigmatic change. Drawing on the distinction between corpus-based and corpus-driven linguistics, I propose three approaches to AI-assisted language research: LLM-based, LLM-driven, and LLM-native linguistics. In LLM-based research, models serve as tools for analyzing linguistic data within previously defined questions and hypotheses. In LLM-driven research, patterns identified by the model provide the basis for formulating new research questions and hypotheses. Both approaches constitute methodological innovations, since the object of inquiry remains human language. LLM-native linguistics, by contrast, treats LLM-generated texts and the models themselves as autonomous objects of linguistic investigation. This approach raises important questions about the status of AI-generated texts and the boundaries of linguistic inquiry. I believe that LLM-native linguistics may change what linguistics recognizes as its object of study.

Keywords

AI-linguistics, LLM-based linguistics, LLM-driven linguistics, LLM-native linguistics, linguistic paradigm

1 Introduction

It was less than two decades ago that Stefan Thomas Gries introduced an exchange on corpus linguistics in *Language and Linguistics Compass* by stating that "[c]orpus linguistics is one of the fastest-growing methodologies in contemporary linguistics" (2009: 1225). We can see today that Gries was very much on the mark back then. In recent years, corpus-based linguistic studies have robustly proliferated, and the corpus itself, whatever its definition, has become a basic tool of empirical inquiry into language. Vigorous advancements in corpus linguistics have not only improved the research toolkit, but also sparked changes in the ways we interpret language data, understand the relationship of theory and empirical evidence, and make sense of the nature of language as such.[1]

[1] What I mean is a shift from the notion of language as an abstract system to the concept of language as, also, a static, probabilistic, and usage-dependent object. While the latter idea was not bred by corpus linguistics alone, corpora definitely gave it a very strong empirical support.

AILing

Słapek, Daniel. 2026.
AI-Linguistics: A New Method or a New Paradigm?
LLM-Based-, LLM-Driven- and LLM-Native-Linguistics.
AI-Linguistica, Vol.6 No.1
DOI: 10.62408/ai-ling.v6i1.107

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses ISSN: 2943-0070
© 2026 Daniel Słapek, licensed under CC BY-NC-SA 4.0.

As a result of Elena Tognini-Bonelli's (2001) work, the distinction between *corpus-based* and *corpus-driven* methodological approaches has become firmly entrenched in corpus linguistics. In the former, the corpus is a tool for verifying or falsifying prior hypotheses informed by linguistic theories, and the corpus data serve to substantiate initial assumptions. In the latter approach, the corpus itself provides a starting point for proposing research insights, generalisations, and hypotheses, with data analysis not so much corroborating/disproving an existing theory as helping researchers derive it from the language data and patterns they identify (on this issue, see e.g. Biber 2009; Storjohann 2005). While the two strategies may be complementary and are sometimes combined in research practice, the distinction has proven methodologically useful and epistemologically important by highlighting the bidirectional interplay of theory and empirical data in contemporary linguistics.

This reference to corpus linguistics is not a random one. The point is that today we are in the midst of another methodological transformation whose scale may eventually equal, if not surpass, that of the corpus revolution. With the dynamic development of Large Language Models (LLMs), they are more and more frequently used as a tool supporting language analysis, at the same time becoming an autonomous environment in which language data are generated, modelled, and interpreted. Paraphrasing Gries, LLM-linguistics (LLM-based/driven/native linguistics, see below)—or, to apply a frequently used metonymy, AI-linguistics—is one of the fastest-growing fields in contemporary linguistics, and there is every reason to believe that this process is only starting.

My aim in this paper is to outline differences between LLM-based, LLM-driven, and LLM-native linguistics. I will seek to locate these three approaches against a broader backdrop of linguistic methodology and theory. I believe that it is a matter of some relevance today to ponder whether LLMs are just new, more effective, and faster instruments for linguists or whether their application entails a deeper, paradigmatic change that impacts the very way we study and make sense of language.[2] The focal issue of my paper is encapsulated in its title, with my central question being whether we are faced with a new method or a new paradigm of linguistics (in Thomas Kuhn's understanding of science paradigms; see below).

2 LLM-Based and LLM-Driven Linguistics

In this context, Tognini-Bonelli's distinction between corpus-based and corpus-driven linguistics is indeed a natural analogy. To elucidate what LLM-based and LLM-driven linguistics are, I will rely on my own research practice. Importantly, the case studies described below are not fully fledged linguistic analyses of selected issues; instead, they are meant to exemplify the difference between the LLM-based and LLM-driven approaches.

I understand LLM-based linguistics as research in which the use of an LLM is preceded by posing research questions, selecting categories, and/or formulating

[2] Technically, LLMs are trained to model statistical regularities in large collections of texts (e.g., Brown et al. 2020). Their ability to generate contextually appropriate linguistic output and perform a wide range of language-related tasks makes them particularly useful in linguistic research.

a preliminary hypothesis. The model itself functions as a tool that makes a given operation on the language material possible or faster by classifying examples, annotating data, identifying constructions that meet the predefined criteria, extracting certain information from large sets of texts, etc. The logic of research procedure, though, remains basically unchanged (hypothesis → LLM → data). Like in corpus-based studies, researchers know what they are looking for or what they want to test in the corpus. Novelty lies in the application of a new instrument and, primarily, in the previously unattainable extent of the analysis this instrument makes possible, rather than in putting forward the research problem as such.

An illuminating example is provided by the syntactic-semantic properties of the Italian infinitive (*infinito*) preceded by the preposition *in* combined with the definite article (a so-called articled preposition, *preposizione articolata*) in sentences such as *Mi sembra che esageri* nel dare *senso a tutto ciò* (It seems to me that you're going too far in trying to make sense of all this). Such structures are traditionally discussed in grammar courses as adverbial clauses of time. Without going into the particulars of syntactic classification, which as a matter of fact should be specified in more detail (see Słapek 2026), students often—and rightly—ask: "Which verbs can be used in such structures?" and "Are there any restrictions on verbs that can appear in this position?"

In classic corpus analysis, structures of this kind can be relatively easily searched out by means of a proper query and extracted from the corpus for further study. What remains unsolved, though, is the size problem of the material obtained in this way. For instance, in *Italian Web 2020*, a corpus containing almost 15 billion tokens, the CQL query comprising *nel*/*nello*/*nell'* + *infinito* retrieves more than 1.8 million examples. Besides, the frequency record indicates that this schema is highly productive, given that the material encompasses as many as 1,006 different verbs from very diverse semantic classes. To classify them manually would be extremely time-consuming, and, worse even, might not lead to the identification of semantic categories that could provide a basis for formulating a clear hypothesis about possible restrictions on the construction.

The point is that the material contains a broad range of verbs: verbs that denote particular activities (*fare* [do/make], *prendere* [take/grab]), perception (*vedere* [see], *sentire* [feel/hear/sense]), cognitive processes (*pensare* [think], *comprendere* [understand]), and communication (*parlare* [speak], *rispondere* [answer]), as well as modal verbs (*volere* [will], *potere* [can], *dovere* [must]).

For their part, LLMs analyse such diversified resources very quickly, making it possible to classify verbs semantically and to present this classification in detailed charts, if only this instruction is properly formulated in the prompt. In my case study, such an analysis did not reveal any clear semantic-lexical restrictions on verbs that could appear in the 'in + art. det. + infinito' structure. An LLM (ChatGPT-5 in my case) made it possible to quickly falsify the preliminary hypothesis that there were such restrictions.[3]

[3] The distinctiveness of this structure is determined, among other factors, by its syntactic and semantic heterogeneity and the ambiguous status of the article-preceded infinitive, which straddles the line between verbal and nominal categories. For a closer discussion, see Słapek 2026.

LLM-driven linguistics works quite differently. A study does not begin from a closed set of issues or structures to be identified by the model in the material. Rather, the material itself and the patterns an LLM identifies in it provide a starting point for research (data → LLM → identified pattern → hypothesis). An example from my experience below illustrates this.

To prepare the academic staff for challenges caused by the expansion of artificial intelligence (AI), the Jagiellonian University's Faculty of Philology held a series of sessions titled *AI Philologist Café*, which discussed how generative AI tools could be used in teaching languages and literatures. A person involved in running the sessions asked me to help them develop training materials of interest to university teachers and, most importantly, applicable in their daily work.[4] Like in the case described above, we chose issues related to building and analysing a language corpus. In brief, the study carried out by an LLM (the Claude model was used) under the user's constant direction and control proceeded as follows:

(1) The initial prompt that defined the research task was: "I am doing linguistic research in order to analyse communication patterns on online forums in Polish and Italian. I want to build a language corpus consisting of authentic posts by online forum users covering various subject areas. Suggest five thematic categories that could be a basis for further comparative analysis."

(2) The model asked some guiding questions, including: a) "What is the main aim of your language analysis? Lexical analysis (vocabulary), pragmatic/discursive analysis, syntactic/stylistic analysis, sociolinguistic analysis?"; b) "Do you prefer controversial categories (more emotions and polemics) or neutral ones (more information and advice)?"; c) "What corpus size do you have in mind for each category (in both languages jointly)?"

(3) The subject areas suggested by the model were: politics, religion, family, sport, and health.

(4) The model generated code for collecting and processing forum posts, which was subsequently used to compile a corpus of texts from several Polish and Italian online forums, including, among others, forum.gazeta.pl, kibice.net, termometropolitico.it, alfemminile.com and milanworld.net.[5]

(5) Data were retrieved (selected results are presented below).

Given that we wanted to produce training resources and enhance participants' interest, we opted for 'controversial categories,' specifically choosing the use of vulgarisms on online forums dedicated to the subject areas listed in point 3. For the study, the LLM collected the corpus of 20,469 posts (1,331,601 words)

[4] For details of the *AI Filolog Café* project, see the Faculty's website www.filg.uj.edu.pl/ai-filolog-cafe; for information on the training, see www.lawrocki.pl.

[5] Citing a passage from Claude: "anonymize.py – hashing of usernames; http_client.py –a session with throttling and retries; jsonl.py –writing/reading of records," etc.

for Polish and 9,440 posts (638,967 words) for Italian; Polish and Italian sources were selected for each of the five subject areas specified in point 3. Some of the hypotheses posed by the LLM on the basis of that material were (the most obscene examples have been removed):

I. Polish swears narrowly; Italian swears broadly. The Polish repertoire of vulgarisms in politics is almost entirely concentrated in three or four lexemes from the intellectual and ethical field. The Italian repertoire is distributed across four separate semantic fields (physiology, evaluation of the act, mental deficit, and personal slur). Presumably, the Italian literary and street tradition (Boccaccio, Pasolini, Roman cabaret) has spawned a broader conventional inventory of vulgar forms used in informal writing.

II. Polish attacks a person; Italian attacks a speech act. The Polish political polemic tends to qualify the interlocutor's features—their intellect, honesty, and common sense. The Italian one tends to qualify what they say as worthless (*stronzate*, *cazzate* [bullshit, crap]). Both approaches are pragmatically aggressive, but they target different aspects of communication. Presumably, this reflects two different polemical traditions, with the Polish being *ad personam*-based, and the Italian operating by categorising statements.

III. Polish vulgarisms in sport fulfil various functions, not all of them aggressive. The *jeb*.* [Eng.: fuck] family, which is traditionally classified as very strong, primarily functions as an intensifier (*zajebisty*, *zajebiście* [fucking]) in PL sport.

IV. The Italian polemical register is lexically sharper despite its lower frequency. Although its token frequency is almost half as high, Italian politics far more often uses physiological vulgarisms (*cazzo* [dick/shit], *merda* [crap], *culo* [arse]) and nominal personal insults (*stronzo* [asshole], *figli di puttana* [sons of bitches]). Polish rarely goes beyond nominal insults from intellectual and ethical fields (*idiota* [idiot], *kłamca* [liar]). Presumably, this difference indicates different acceptability norms in the two forum-discursive cultures.

Of course, at this stage this is only a preliminary material that can be a basis of proper linguistic analysis. Nevertheless, these observations seem interesting and unobvious enough (e.g. II. "Polish attacks a person; Italian attacks a speech act") to be a good point of departure for posing research problems which might not occur at all to a researcher when undertaking analysis with a pre-set list of questions.

Importantly, the LLM offered these hypotheses on the basis of detailed statistical data fully accessible to the user (I do not cite them here, seeing that my argument is purely illustrative). To build the corpus and analyse it, the model needed 313 min (= 5 h 13 min), while the estimated time of corpus building and analysis by a human being (as calculated by Claude) would be 70–105 h, that is, 9–

13 work days (about 2–3 weeks of working at the full-time-job rate of 8 hours a day).[6]

3 LLM-Native Linguistics

I propose to call the third approach to LLM-generated data *LLM-native linguistics*, where *native* does not only mean 'having its source/born in' (from Latin *nativus*; for the two approaches above might also be described in this way) but also denotes 'embedment' in LLM-generated data (by analogy to *cloud-native*, which refers to constructing and operating applications in cloud environments). Accordingly, LLM-native linguistics is the study of structures, patterns, and phenomena in the LLM-generated models where LLM productions are treated as an autonomous object of linguistic research and, interestingly, LLMs themselves are regarded as being native language users. This approach seems to stir great interest among linguists today, if not to be the most widespread branch of research using LLM-generated data currently.[7]

The success of this approach is perhaps best exemplified by *AI-Linguistica: Linguistic Studies on AI-Generated Texts and Discourses*, a journal published by PUBLIA–SLUB Open Publishing (Saxon State and University Library Dresden) since 2024 with Anna-Maria De Cesare Greenwald as Editor-in-Chief. The journal's website states that its aim is "to reflect on generated texts from a variety of perspectives: theoretical, descriptive, and applied", where the very term *generated texts* is quite broadly understood as comprising "formats as diverse as texts generated by Large Language Models, AI-powered smart agents (i.e. chatbots, voice assistants, social bots, etc.), writing assistance tools, template-based software, and neural machine translation systems" (ai-ling.publia.org). Contributions to *AI-Ling* explore issues such as:

a) The grammatical competence of models, with linguists focusing on, for example, the syntactic competence of models (e.g. *Syntactic Patterns of*

[6] Interestingly, the LLM exhibited extraordinary methodological 'maturity and honesty' by listing the 'Pilot study limitations' in its report, such as "1. The asymmetry of forums: PL politics/religion/relations/health from one forum (gazeta.pl), IT from three forums. This can affect style (one forum = one micro-language). 2. The lack of an IT Catholic forum: all the tested Catholic forums were dead (the community migrated to Facebook). IT religion comes from the (secularly slanted) philosophical-religious forum *termometropolitico*. 3. PL sport heavily rescaled: kibice.net yielded >700k words (vs ~50–100k target). Sub-sampling recommended for balance analyses. 4. Dates widely scattered: 2002–2026. The filter "date >= 2020" would make sense in researching contemporary language. 5. No lemmatisation: frequencies contain all inflected forms. spaCy PL/IT linked to "tools/" in the next iteration."

[7] An important distinction should be made here between an LLM as a computational system and a conversational AI based on LLMs. The former can be investigated in terms of its internal representations, while the latter can be studied through the linguistic properties of its output. These two perspectives are related, but they are not methodologically identical. In the present work, I propose to regard both as falling within the broader category of LLM-native linguistics, since in both cases the LLM or its linguistic output become the object of study rather than just an instrument for analysing language data. Within this category, however, we could distinguish between model-oriented and output-oriented research.

Italo-Romance CP-layers in Transformers, ChatGPT and DeepSeek: Two Case Studies from Romansh and Neapolitan [Samo and Massaro 2025]) and assessment of the grammatical correctness and acceptability of AI-produced texts (e.g. *Benchmarking AI Acceptability and Grammaticality in German: A Study of ChatGPT and Human Judgments* [Catasso 2026]). In this regard, the fundamental questions are what grammatical knowledge LLMs 'possess' and how far it resembles native speakers' competence. Such research can be regarded as today's version of classic linguistic experimentation with a language model, instead of a human being, as its object. In this framework, LLMs are treated not only as 'producers' of texts but also as language informants to be asked questions just like respondents are interviewed in psycholinguistic research studies.[8]

b) The pragmatic competence of models, with researchers investigating pragmatic phenomena such as presuppositions (e.g. *Politicians vs ChatGPT: A Study of Presuppositions in French and Italian Political Communication* [Garassino et al. 2024]) and implicature (e.g. *How do DeepL and ChatGPT Process Information Structure and Pragmatics? An Exploratory Case Study on Topicalized Infinitives in Spanish (and Portuguese)* [Gerhalter 2024]). Their shared concern is whether LLMs indeed operate with pragmatic meaning or whether they only reproduce statistical patterns.

c) LLM-generated language as an autonomous language variety (its style, genres, and lexical and syntactic traits), with scholars investigating, among other issues, grammatical characteristics of AI-created texts (e.g. *Grammatica generata: analisi di alcune strutture presenti nei testi di ChatGPT in lingua italiana* [Tavosanis 2025]) and AI-produced simplifications of administrative writings (e.g. *Exploiting ChatGPT to Simplify Italian Bureaucratic and Professional Texts* [Paci et al. 2024]). Scholars observe that certain stylistic, lexical, and syntactic patterns occur across LLM-generated texts.

There are also studies examining automatic translation (e.g. *DeepL traduce Maraini: le poesie metasemantiche 'sotto gli occhi' della traduzione automatica* [Saccà 2024]) and papers that discuss both a model as such and its application in linguistic pursuits [e.g. *Vocabolari d'autore e intelligenza artificiale: è possibile creare il vocabolario della lingua di Boccaccio con ChatGPT?* [Palmieri 2025]).

The scale of interest in the LLM-native approach to language data is vividly illustrated by developments in my own research field. At the recent 18th Congress of the International Society of Italian Linguistics and Philology (SILFI), as many as twelve talks out of about one hundred papers and posters concerned LLMs, and

[8] This line of inquiry can also be situated within the broader context of computational approaches to linguistic theory. For example, Wilcox, Futrell, and Levy (2024) demonstrate how language models can be used to investigate questions of syntactic learnability. Although research on LLMs is not identical to research on conversational LLMs (as discussed above), this study illustrates the intersection between NLP research and linguistic theory.

ten of those fell under the LLM-native linguistics category.[9] This was also true of Anna Maria De Cesare's keynote lecture titled *Quanto è naturale l'italiano "sintetico" prodotto dall'IA generativa?* ('How natural is the "synthetic" Italian language generated by generative AI?'; see her publications on this topic: e.g., De Cesare 2023; 2025; 2026). In other words, more than ten percent of presentations at a congress of a scholarly association whose name references philology as traditionally understood in the spirit of the Italian university tradition not only used LLM-generated data but also embraced LLM-native linguistics.[10]

4 Conclusions

Let us revisit the question of method and paradigm. In Kuhn's classic account, a paradigm is not one theory or a single research method; instead, it is an ensemble of commonly endorsed premises, models, practices, issues, and standard ways of solving them. In other words, shared by the community of scholars, a paradigm defines how research is carried out, what counts as a valid research problem, and primarily how the very object of knowledge is understood (Kuhn 1970 [1962]: 10–11; 40–46; 111–121). Transferring Kuhn's notion of paradigm to linguistics is not straightforward. Kuhn himself highlighted the difference between the natural sciences and the social sciences, recognising a far greater scale of conflict around what are legitimate research issues and methods in the latter (1970 [1962]: viii; for disciplinary boundaries, among others, between linguistics and the social sciences, see e.g. Bouterse and Karstens, 2015). Besides, there is no consensus either on the number or on the scope and nature of disciplinary paradigms within linguistics itself, or rather within metalinguistics[11] (see e.g. Słapek 2017; Jędrzejko 2018; Stalmaszczyk 2018; the applicability of Kuhn's notion of paradigm to linguistics was discussed by Percival 1976 and Winston 1976). Yet, there is also a broader concept of a paradigm, where a paradigm is a certain way of understanding the object of research in conjunction with a particular ontology of language, mode of research, and answers to two fundamental questions: What is language? and thus: What is the object of linguistic research? (Stalmaszczyk 2018). From this perspective, consecutive linguistic paradigms have differed not only in the methods they have used but also in the status they have ascribed to language and the aspects of language they have foregrounded (as in comparative-historical, structural, general, and cognitive linguistics; on this issue, see e.g. Bobrowski 2015).

[9] The conference, which I had the pleasure of co-organising, was held in Cracow on 27–29 May 2026; for the conference programme, see *silfi.org*.

[10] While writing these words, I received news of the publication of *AI-Generated Texts: Issues, Challenges and Operational Proposals* (Nitti 2026). Such studies are likely to proliferate soon. Of course, there is already a substantial amount of research on AI-generated texts across a wide range of languages. To mention just a few examples, see Mizumoto et al. (2024) for English, Drackert et al. (2025) for German, De Cesare (2026) for Italian, Mazur (2024) for Polish, and Li and Zhang (2025) for Chinese.

[11] In Polish linguistics, the term *metalinguistics* has a dual meaning. In the broad sense, it means all meta-scientific reflection on linguistics (e.g. Grucza 1983; Słapek 2017), and in the narrow sense it means a critique of linguistic ideas and texts (e.g. Żmigrodzki 2003).

Importantly, linguistics does not follow a linear pattern of replacing one paradigm with another; indeed, linguistics is multi-paradigmatic today, with different models of language and traditions of research coexisting, intersecting, and interacting (Bobrowski 2015; Słapek 2017; Jędrzejko 2018).

I believe that LLM-based and LLM-driven linguistics should primarily be regarded as new approaches in the methodology of language research. In both cases, LLMs change the way in which linguistic findings are arrived at as they provide new procedures for analysing language materials and/or novel modes of formulating hypotheses, yet they do not lead to redefining the object of linguistics. Research still focuses on human language, and LLMs are just tools for exploring it (this is also how corpus-based and corpus-driven linguistics tend to be viewed).[12] However, LLM-native linguistics is a different matter since change is not limited to the methodology of research in it; crucially, the ontology of the research object shifts as well. An LLM-generated text no longer serves to study human language, but becomes an autonomous language fact, and, more importantly perhaps, the model itself enters the picture as a new subject in linguistic research. If one criterion of paradigm change is a revised understanding of the object of research (language), something more than merely a methodological change is at hand in LLM-native linguistics.

LLM-native linguistics promotes asking new questions in linguistics, such as: Can a product of a non-biological system be an autonomous object of linguistic analysis? Can language competence be attributed to an LLM as well? Should an LLM-generated language be regarded as an imitation of human language, as its unique variety, or rather as a distinct kind of language action? Of course, these new questions alone do not mean that a new paradigm (as defined by Kuhn) has been constituted. This would probably take both time and some consolidation in the linguistic community, including the formation of a relatively cohesive set of research issues and practices (for now, LLM-native linguistics is still somewhat of a novelty). Nevertheless, it might be argued that while LLM-based and LLM-driven linguistics are a new research method, for they transform the way language is studied, LLM-native linguistics recasts what we are prepared to recognise as language and the research object of linguistics (the distinction between ‘how we study’ vs ‘what we study’). Hence, in the latter case, the question whether a new paradigm is at stake is no longer a purely rhetorical one.

My brief outline of the three manners in which LLMs are used in language research—LLM-based, LLM-driven, and LLM-native linguistics—shows that linguistics is going through an essential transformation. Today, LLMs can serve as an analytical tool, function as a source of data, and be an object of research in its own right. Obviously, research at the intersection of these three approaches can be expected to develop, such as the study of AI-generated texts (LLM-native) by means of LLMs (LLM-based) with elements of the LLM-driven framework (when the model itself identifies patterns the researcher has not anticipated). Other intriguing developments include linguistic analyses of the ways in which people talk of and metaphorically conceptualise AI (e.g. Stanič 2025) and examination of

[12] See e.g.: “I think Corpus Linguistics is definitely a method(ology) or a “methodological paradigm,” no more but also no less” (Gries 2011: 83).

individuals' communication with AI models, that is, of human-machine interactions (e.g. Lehner 2025). Captivating though they may be, such studies are not exactly part of the branch of linguistics discussed in this paper because what they research are not productions of LLMs but, so to speak, traditionally conceived methods and products of people's language actions, even if limited to a certain thematic area or context.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

Acknowledgments

I acknowledge the use of OpenAI's GPT-5 model for the analyses presented in the first part of Section 2 and Anthropic's Claude model for the analyses presented in the second part of Section 2. However, I ensured that all contributions adhered strictly to the standards and ethical guidelines of academic writing.

References

Biber, Douglas. 2009. Corpus-based and corpus-driven analyses of language variation and use. In Heine, Bernd & Narrog, Heiko (eds), *The Oxford Handbook of Linguistic Analysis*, 159–192. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199544004.013.0008

Bobrowski, Ireneusz. 2015. *Problems of Methodology and Philosophy in Linguistics*. Frankfurt am Main: Peter Lang Edition.

Bouterse, Jeroen & Karstens, Bart. 2015. A diversity of divisions: Tracing the history of the demarcation between the sciences and the humanities. *Isis. A Journal of the History of Science Society* 106(2). 341–352. https://doi.org/10.1086/681995

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33. 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Catasso, Nicholas. 2026. Benchmarking AI acceptability and grammaticality in German: A Study of ChatGPT and Human Judgments. *AI-Linguistica* 3(1). https://doi.org/10.62408/ai-ling.v3i1.35

De Cesare, Anna-Maria. 2023. Assessing the quality of ChatGPT's generated output in light of human-written texts. A corpus study based on textual parameters. *CHIMERA. Romance Corpora and Linguistic studies* 10. 179–210.

De Cesare, Anna-Maria. 2025. LLM referential chain generation. A qualitative case study based on Italian biographies produced by GPT-4. *Linguistik Online* 136(4). 25–52. https://doi.org/10.13092/8tppdv47

De Cesare, Anna-Maria. 2026. *L'italiano sintetico dell'intelligenza artificiale generativa*. Firenze: Cesati.

Drackert, Anastasia & Horbach, Andrea & Peters, Anja. 2025. How good are LLMs in generating input texts for reading tasks in German as a foreign language? *Annual Review of Applied Linguistics* 45. 222–252. https://doi.org/10.1017/S0267190525000066

Garassino, Davide & Masia, Viviana & Brocca, Nicola & Delorme Benites, Alice. 2024. Politicians vs ChatGPT: A study of presuppositions in French and Italian political communication. *AI-Linguistica* 1(1). https://doi.org/10.62408/ai-ling.v1i1.5

Gerhalter, Katharina. 2024. How do DeepL and ChatGPT process information structure and pragmatics? An exploratory case study on topicalized infinitives in Spanish (and Portuguese). *AI-Linguistica* 1(1). https://doi.org/10.62408/ai-ling.v1i1.8

Gries, Stefan Th. 2009. What is corpus linguistics? *Language and Linguistics Compass* 3(5). 1225–1241. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Gries, Stefan Th. 2011. Methodological and interdisciplinary stance in corpus linguistics. In Viana, Vander & Zyngier, Sonia & Barnbrook, Geoff (eds), *Perspectives on Corpus Linguistics*, 81–98. Amsterdam & Philadelphia: John Benjamins.

Grucza, Franciszek. 1983. *Zagadnienia metalingwistyki. Lingwistyka – jej przedmiot, lingwistyka stosowana*. Warszawa: PWN.

Jędrzejko, Ewa. 2018. Jeszcze o wieloparadygmatyczności i dylematach lingwistyki na progu trzeciego tysiąclecia. In Hawrysz, Magdalena & Uździcka, Marzanna (eds), *Wokół dziejów używania polszczyzny – wspólnoty, potrzeby, zachowania komunikatywne*, 107–120. Zielona Góra: Uniwersytet Zielonogórski.

Kuhn, Thomas S. 1970 [1962]. *The Structure of Scientific Revolutions* (2nd edition, enlarged). Chicago, London: University of Chicago Press.

Lehner, Sabine. 2025. The design and discursive construction of a "speaking" vacuum cleaning robot for assistive purposes: Findings on communication ideologies from a current research and development project. *AI-Linguistica* 2(1). https://doi.org/10.62408/ai-ling.v2i1.16

Li, Ziqi & Zhang, Qi. 2025. Linguistic differences between AI and human comments in Weibo: Detect AI-generated text through stylometric features. *Proceedings of the 24th China National Conference on Computational Linguistics*. 842–851. https://aclanthology.org/2025.ccl-1.64.pdf (last accessed on 09/11/2026).

Ławrocki, Łukasz. 2026. Szkolenie z zakresu AI w ramach cyklu AI Filolog Café. https://lawrocki.pl (last accessed on 09/11/2026).

Mazur, Rafał. 2024. O poprawności językowej tekstów generowanych przez SI na przykładzie ChatuGPT. *LingVaria* 19(1). 119–138. https://doi.org/10.12797/LV.19.2024.37.08

Mizumoto, Atsushi & Yasuda, Sachiko & Tamura, Yu. 2024. Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics* 4(3). 100106. https://doi.org/10.1016/j.acorp.2024.100106

Nitti, Paolo (ed.). 2026. *AI-Generated Texts: Issues, Challenges and Operational Proposals*. Firenze: Franco Cesati.

Paci, Walter & Gregori, Lorenzo & Acerboni, Giovanni & Panunzi, Alessandro & Perugini, Maria Roberta. 2024. Exploiting ChatGPT to simplify Italian bureaucratic and professional texts. *AI-Linguistica* 1(1). https://doi.org/10.62408/ai-ling.v1i1.13

Palmieri, Claudia. 2025. Vocabolari d'autore e intelligenza artificiale: è possibile creare il vocabolario della lingua di Boccaccio con ChatGPT? *AI-Linguistica* 2(2). https://doi.org/10.62408/ai-ling.v2i2.20

Percival, W. Keith. 1976. The applicability of Kuhn's paradigms to the history of linguistics. *Language* 52(2). 285–294. https://doi.org/10.2307/412560

Saccà, Francesco. 2024. DeepL traduce Maraini: le poesie metasemantiche "sotto gli occhi" della traduzione automatica. *AI-Linguistica* 1(1). https://doi.org/10.62408/ai-ling.v1i1.4

Samo, Giuseppe & Massaro, Angelapia. 2025. Syntactic patterns of Italo-Romance CP-layers in transformers, ChatGPT and Deepseek: Two case studies from Romansh and Neapolitan. *AI-Linguistica* 2(2). https://doi.org/10.62408/ai-ling.v2i2.25

Słapek, Daniel. 2017. *Rozważania metajęzykoznawcze*. Łódź: Primum Verbum.

Słapek, Daniel. 2026. Osservazioni sul costrutto 'in + art. det. + infinito'. In Saržoska, Aleksandra & Trajkova, Mira (eds), *Intrecci disciplinari: linguistica, traduzione e didattica in dialogo*. Skopje: Univerzitet „Sv. Kiril i Metodij" vo Skopje.

Stalmaszczyk, Piotr. 2018. O paradygmatach w językoznawstwie (na przykładzie językoznawstwa generatywnego). *Linguistica Copernicana* 15. 37–59. https://doi.org/10.12775/LinCop.2018.004

Stanič, Daša. 2025. La concettualizzazione metaforica dell'intelligenza artificiale: Analisi del discorso giornalistico italiano contemporaneo. *Linguistica* 65(1). 185–203. https://doi.org/10.4312/linguistica.65.1.185-203

Storjohann, Petra. 2005. Corpus-driven vs. corpus-based approach to the study of relational patterns. In *Proceedings of the Corpus Linguistics Conference 2005.*, 1–20. Birmingham: University of Birmingham. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/5006/file/Storjohann_Corpus_driven_vs_corpus_based_approach_to_the_study_of_relational_patterns_2005.pdf (last accessed on 09/11/2026).

Tavosanis, Mirko. 2025. Grammatica generata: analisi di alcune strutture presenti nei testi di ChatGPT in lingua italiana. *AI-Linguistica* 2(2). https://doi.org/10.62408/ai-ling.v2i2.23

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins Publishing.

Wilcox, Ethan Gotlieb & Futrell, Richard & Levy, Roger. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry* 55(4). 805–848. https://doi.org/10.1162/ling_a_00491

Winston, Morton E. 1976. Did a (Kuhnian) Scientific revolution occur in linguistics? In Suppe, Frederick & Asquith, Peter D. (eds) *Proceedings of the Biennial Meeting of the Philosophy of Science Association, Vol. I: Contributed*

Papers, 25–33. Chicago: The University of Chicago Press. https://www.jstor.org/stable/192348 (last accessed on 09/11/2026).

Żmigrodzki, Piotr. 2003. Metalingwistyka jako krytyka idei i tekstów językoznawczych. *Polonica* 22–23. 5–18.

License


This article is licensed under the *Creative Commons Attribution 4.0 International* license, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format (including for commercial purposes) if certain conditions are met. In particular, you must give appropriate credit to the original author and source, provide a link to the Creative Commons license, and indicate if you have made any modifications. Images or other material from third parties included in this publication are also covered by this Creative Commons license.

An easy-to-understand description of this license can be found at https://creativecommons.org/licenses/by/4.0/deed.en. The legally binding, complete text of the license is also linked there.

About the journal

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses is a Diamond Open Access journal supported by PUBLIA – SLUB Open Publishing.



DOI https://doi.org/10.62408/ai-ling

ISSN 2943-0070

Disclaimer

The statements, opinions and data contained in this paper are solely those of the respective authors and not of *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity disclaim any responsibility for injury to persons or damage to property resulting from the ideas, methods, instructions or products mentioned in the content.

The authors of this paper are solely responsible for ensuring that all components of this article, including images or other material from third parties, can be published under the CC BY 4.0 license of the article (see above). *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity decline any responsibility for damages resulting from any copyright infringements by the authors. If the paper was written by only one person, the term "authors" also applies to that person.