# AI linguistics, *quo vadis?* Introduction to the special issue *Natural Language and AI. New Perspectives for Linguistic Studies*

*Nicholas Catasso* (Bergische Universität Wuppertal)

catasso(at)uni-wuppertal.de

## 1 Background

In recent years, the study of language has undergone a profound transformation driven by rapid advances in artificial intelligence (AI) and in particular by the emergence of large-scale neural language models. Systems based on deep learning architectures – most prominently transformer-based models – have achieved unprecedented performance across a wide range of language-related tasks, from text generation and translation to inference and dialogue. As a result, AI has moved from being primarily a tool for engineering applications to becoming a relevant object of inquiry within linguistic research itself. This shift has given rise to a rapidly expanding interdisciplinary field in which computational modeling and linguistic theory increasingly inform one another.

At the core of these developments lies the question of how natural language can be represented in a form that is accessible to computational systems. Natural Language Processing (NLP), situated at the intersection of computer science and linguistics, addresses precisely this issue by developing formal and statistical methods for modeling linguistic structure and meaning. The overarching goal is often described as enabling human-like language processing (Young et al. 2018), that is, the extraction and interpretation of coherent information from textual data (Khan et al. 2010). This task is non-trivial, given the inherent properties of natural language, including cross-linguistic variation, synonymy and polysemy, as well as pragmatic phenomena such as negation, irony and implicature.

Since computational systems cannot operate directly on linguistic input, language must be transformed into representations that are machine-readable. Early approaches relied on relatively simple vector-space models, in which texts are encoded as distributions over lexical items. Binary bag-of-words representations or frequency-based weighting schemes such as term frequency-inverse document frequency (TF-IDF) (Salton and Buckley 1988; Ramos 2003) provided the first widely used numerical encodings of textual data. While such models enabled basic forms of information retrieval and text classification, they remained limited in their ability to capture structural dependencies and contextual meaning.

More recent approaches, in particular those based on neural networks and contextualized representations, aim to overcome these limitations by modeling

AILing

Catasso, Nicholas. 2026.
AI linguistics, *quo vadis?*
Special Issue: *Natural Language and AI*. Vol. 3 No. 1
DOI: 10.62408/ai-ling.v3i1.88

*AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* ISSN: 2943-0070

language as a highly structured and context-sensitive system. Transformer-based architectures, in particular, enable the processing of vast amounts of human-generated textual data, allowing models to acquire statistical regularities of language use at an unprecedented scale (Ananthaswamy 2023). As a result, contemporary systems are capable of generating highly coherent and contextually appropriate text, often without relying on explicitly encoded rules or external knowledge sources. Notably, these models can generalize to previously unseen tasks or input types, a property commonly described as "zero-shot " (Radford et al. 2019). These models are designed to capture not only distributional regularities but also aspects of syntactic structure and semantic composition, thereby addressing long-standing challenges in computational modeling of language. Against this background, contemporary large language models (LLMs) can be understood as attempts to approximate aspects of linguistic competence through large-scale data-driven learning, raising the question of what kinds of linguistic knowledge these systems encode and how this knowledge relates to theoretical models of language. At the same time, the nature of this apparent competence remains contested. While some authors argue that such systems merely simulate understanding on the basis of statistical pattern matching, variously described as "superficial" language processing (Chomsky et al. 2023), linguistic "simulation" (Berins 2023) or the behavior of "stochastic parrots" (Bender et al. 2021: 617), others interpret their performance as an emergent property of scale, pointing to the role of increasingly large datasets and parameter spaces in giving rise to qualitatively new capabilities (Bommasani et al. 2021; Ganguli et al. 2022; Tamkin et al. 2021; Wei et al. 2022).

Within linguistics, interest in AI is no longer confined to NLP as an applied domain. Rather, it extends to fundamental theoretical questions concerning the nature of linguistic knowledge, representation and use. A growing body of research investigates the extent to which contemporary language models encode, approximate or simulate linguistic competence across different levels of analysis. Much of this research, however, remains strongly centered on English, reflecting both the predominance of English-language training data and the uneven availability of linguistic resources across languages. These developments raise both methodological and epistemological questions: Can AI systems serve as models of human linguistic cognition? To what extent do they capture structural generalizations as opposed to statistical regularities? And how can they be productively integrated into empirical linguistic research?

Across core subfields of linguistics, recent work has begun to systematically explore these issues. In syntax, a substantial line of research has examined whether neural language models acquire abstract structural knowledge or rely on surface-level heuristics. Early probing studies (e.g. Rogers et al. 2020; Wilcox et al. 2020; van Schijndel and Linzen 2021) and large-scale evaluations of syntactic generalization (Hu et al. 2020; Warstadt et al. 2020) have provided evidence that such models capture certain hierarchical dependencies, while also revealing persistent limitations. More recent work has framed these questions explicitly in the

context of LLMs, addressing issues of robustness, multilingual transfer and human-like processing effects (Zhou et al. 2023; Hale and Stanojević 2024; Thomas and Joseph 2025). Dedicated evaluation platforms such as SyntaxGym (Gauthier et al. 2020) and benchmark suites like BliMP (Warstadt et al. 2020) have further enabled fine-grained assessment of syntactic competence.

In morphology, research has focused on the capacity of language models to generalize beyond memorized forms and to handle productive inflectional systems. Studies on systematic generalization (Ismayilzada et al. 2025; Xiong and Wu 2025) highlight persistent difficulties in capturing rule-like behavior, particularly in morphologically rich languages. At the same time, the role of subword tokenization has sparked an ongoing debate about whether morphological structure is genuinely learned or merely approximated through distributional patterns. Cross-linguistic investigations further demonstrate that performance varies significantly depending on typological properties, underscoring the importance of morphology as a testing ground for linguistic generalization in AI systems.

In semantics, LLMs are increasingly treated as repositories of structured knowledge and as tools for semantic parsing. Building on established annotation frameworks such as AMR (Banarescu et al. 2013), UCCA (Abend and Rappoport 2013) and FrameNet (Baker et al. 1998), recent work has shown that pretrained models can achieve strong performance in mapping text to formal meaning representations. At the same time, studies on knowledge extraction and inference (Marvin and Linzen 2018; Petroni et al. 2019) suggest that such models encode substantial amounts of relational and world knowledge. Benchmarks such as GLUE (Wang et al. 2018), SuperGLUE (Wang et al. 2019) and BIG-bench (Srivastava et al. 2022) have played a crucial role in evaluating semantic capabilities, including entailment, compositionality and reference resolution. More recent developments, such as chain-of-thought prompting and tool-augmented models, have further extended the scope of semantic reasoning in AI systems. Meanwhile, a number of studies have been focused on the quite intricate limitations of semantic competence in LLMs, in particular their lack of grounding, instability in interpretation and difficulties in relating linguistic form to real-world meaning (e.g. Bender and Koller 2020; Ma 2024; Schüle 2025; Rondini 2026 and literature therein)

Pragmatics has likewise emerged as a key domain in which AI systems are evaluated and conceptualized. A substantial body of recent work has focused on assessing the extent to which AI systems exhibit human-like performance in core pragmatic phenomena, e.g. irony and sarcasm detection, the interpretation of context-dependent meaning and the integration of background knowledge (Guo et al. 2023; Bommasani et al. 2023; Chang et al. 2024; Park et al. 2024; Kastrati et al. 2025). Computational approaches grounded in frameworks such as Rational Speech Act theory (Degen 2023) have provided formal models of pragmatic inference, implicature and speaker-listener reasoning. In parallel, work on instruction tuning and alignment (Ouyang et al. 2022; Rafailov et al. 2023) has foregrounded the importance of communicative intent and cooperative principles in shaping language model behavior. Dialogue systems and conversational agents have become central

testbeds for pragmatic competence, encompassing turn-taking, contextual appropriateness and the resolution of underspecified meaning. At the same time, critical perspectives have emphasized the role of presupposition, bias and social meaning, highlighting the broader implications of AI-mediated communication (for an overview, cf. Ma et al. 2025).

Beyond synchronic analysis, AI-based methods are increasingly being applied to questions in historical linguistics and language change. In particular, contextualized word representations and transformer-based models have enabled novel approaches to modeling diachronic semantic change, allowing researchers to trace shifts in meaning across time with a level of granularity that was previously unattainable. Recent work has also explored the adaptation of language models to historical corpora, addressing challenges such as orthographic variation and temporal contamination. The growing relevance of this line of research is reflected in the emergence of dedicated scholarly venues, including workshops and conferences specifically focused on the intersection of historical linguistics and AI (e.g., Historical Languages and AI, Humboldt-Universität zu Berlin, Germany, March 2026).At the same time, AI systems are increasingly being applied to tasks such as manuscript digitization and the translation of ancient languages (Moutsis et al. 2025), further expanding the empirical basis for diachronic research. Taken together, these developments open up new perspectives on long-standing questions concerning semantic change, grammatical evolution and the dynamics of linguistic systems. They also point to a rapidly evolving research landscape whose methodological and theoretical implications are likely to shape the field in the years to come. Finally, research on language variation has demonstrated that AI systems both reflect and amplify patterns of sociolinguistic diversity (among many others, Kelly-Holmes 2022, 2024; Székely, Miniota and Hejná 2025; Ocumpaugh, Liu and Zambrano 2025; Li and de Winter 2026). For example, recent work on dialect bias demonstrates that LLMs systematically disadvantage non-standard varieties: studies on African American English, as well as on a number of other non-national varieties show increased stereotyping, reduced comprehension and degraded response quality compared to standard varieties (Fleisig et al. 2024). These disparities extend beyond English, with evidence that models exhibit bias against regional dialect speakers, associating them with negative traits and making biased decisions based on linguistic variation (Bui et al. 2025; Platzgummer, McCrae and Ahmadi 2026). More broadly, recent evaluations of multilingual and cross-linguistic performance show that LLMs tend to normalize toward dominant language standards, for instance exhibiting English-centric patterns even when generating other languages (Guo et al. 2025; Schut, Gal and Furquhar 2025). At the same time, emerging work in computational sociolinguistics argues that language models inherently model socially meaningful variation, including identity-linked linguistic features, but do so unevenly and often in ways that reproduce existing hierarchies (Grieve et al. 2025). These findings highlight both the capacity of AI

systems to capture linguistic variation at scale and the risks they pose in terms of bias, fairness and representational adequacy.

These strands of research illustrate the extent to which AI has become deeply embedded in contemporary linguistic inquiry. AI systems are not only objects of evaluation but also tools for annotation, modeling and hypothesis testing. This includes not only theoretical and descriptive domains, but also applied linguistic research engaging with AI in contexts such as language learning, assessment and teacher education (Chapelle et al. 2024).

The studies gathered here build on these developments and explore, from a range of perspectives, what AI can do for linguistics – and, conversely, what linguistic theory can contribute to the understanding and advancement of AI.

## 2 Contributions in this special issue

### 2.1 Overview of the contributions

The contributions in this special issue collectively explore the linguistic capacities and limitations of LLMs across different domains, including discourse, grammar, pragmatics and language change. Taken together, they provide a multifaceted perspective on how AI systems engage with linguistic structure and variation.

FRANZ MEIER's paper investigates intersubjectivity in AI-generated and human-written editorials. The study shows that while language models can reproduce stance-taking strategies, they do so in systematically different ways. In particular, AI-generated texts rely more heavily on explicit stance markers, suggesting a compensatory strategy for their lack of contextual grounding and pragmatic awareness. Meier's study highlights that LLMs can simulate discourse-level phenomena, but often through surface-level approximation rather than deeply contextualized positioning.

A complementary perspective is offered by NICHOLAS CATASSO's paper, which examines grammaticality and acceptability judgments in German. The findings indicate a high degree of alignment between ChatGPT and human speakers overall, but also reveal systematic divergences in cases involving gradience, sociolinguistic markedness and contextual appropriateness. Importantly, the study demonstrates that LLMs do not merely encode grammatical rules, but rather reflect probabilistic patterns shaped by usage, exposure and social norms.

Focusing on pragmatics, NICOLA BROCCA, ELENA NUZZO and JOSEPH WANG-KATHREIN evaluate the performance of different AI systems in speech act annotation. The results show that ChatGPT achieves near-human accuracy in annotating pragmatic categories, but raises concerns regarding reproducibility and consistency over time. This contribution underscores both the methodological potential of LLMs for corpus annotation and the challenges associated with their opacity and instability.

The study by PETRA SLEEMAN explores the role of ChatGPT as a linguistic informant in translation tasks. The study finds that AI-generated translations closely approximate human outputs and can provide useful observational data for linguistic

analysis. However, the model's limitations become evident when deeper analytical explanations are required, indicating a gap between surface performance and underlying linguistic competence.

GOHAR RAHMAN examines the use of GPT-4 for semantic annotation in Urdu, a low-resource language. The study shows that, through instruction-based prompting, the model achieves strong performance across tasks such as named entity recognition, semantic similarity and sentiment analysis. At the same time, it identifies limitations in handling culturally specific expressions, idioms and sarcasm, highlighting the continued need for careful prompt design and human-AI collaboration. By focusing on a low-resource language, this study extends the scope of the special issue beyond high-resource settings and demonstrates that the opportunities and challenges of LLMs are equally pronounced in multilingual and underrepresented linguistic contexts.

Finally, MARINUS WIEDNER and MATTHIAS SCHÖFFEL's article adopts a computational modeling perspective on language change, demonstrating that machine learning approaches can capture systematic patterns in the evolution of grammatical gender. While not based on LLMs in the narrow sense, this study illustrates the broader potential of AI methods to model diachronic linguistic processes and uncover distributional regularities in language data.

### 2.2 Overall observations

The papers in this special issue bring together a set of empirically grounded studies that examine how contemporary AI and computational approaches engage with linguistic structure, use and variation across a range of domains. By focusing on areas such as discourse organization, grammaticality judgments, pragmatic annotation, semantic processing and language change, the papers offer a differentiated view of what these systems can and cannot do when confronted with the complexity of natural language.

A consistent observation across these studies is that AI-based systems approximate human linguistic behavior through large-scale statistical patterning rather than through an underlying system of linguistic competence. This is evident, for example, in the way models reproduce discourse-level phenomena such as stance-taking, where they rely on overt and often repetitive markers, or in their handling of acceptability judgments, where high overall agreement with human speakers coexists with systematic deviations in gradient or context-sensitive cases. Similarly, in semantic annotation and translation tasks, models are able to produce outputs that closely resemble human performance, yet these outputs are best understood as the result of distributional alignment rather than of rule-based or conceptually grounded linguistic knowledge.

At the same time, the results presented in these different lines of work make clear that model behavior is not uniform, but varies in systematic ways across dimensions of linguistic variation. Differences emerge with respect to register and

genre, for instance when models handle formal versus informal language, as well as in relation to pragmatic phenomena such as speech acts or implicatures, where performance depends on subtle contextual cues. Variation also becomes visible across linguistic systems and resource conditions. The study on Urdu demonstrates that, in low-resource settings, model performance is closely tied to the availability and distribution of training data, including differences between scripts and degrees of standardization. In this sense, linguistic variation is not simply captured by the models, but filtered through the uneven representation of languages, varieties and usage contexts in the data on which they are trained.

Some of the investigations show that computational models encode socially and distributionally mediated norms of language use. Rather than operating with stable, context-sensitive representations of meaning, they reproduce patterns that reflect frequency, co-occurrence and conventionalization in the training data. This becomes particularly apparent in cases involving gradience, idiomatic expressions or culturally embedded meanings, where models tend to default to more literal, prototypical or statistically dominant interpretations. As a result, the linguistic behavior of these systems cannot be separated from the social and distributional structures that shape their input.

In addition to these descriptive findings, the contributions raise a set of methodological issues that are directly relevant for the use of AI systems in linguistic research. Questions of reproducibility and consistency arise, for instance, when model outputs vary across prompts or over time. The use of LLMs as annotators or linguistic informants introduces further challenges, including the opacity of their decision processes and the difficulty of interpreting their outputs in theoretically meaningful terms. At the same time, the high level of performance observed in tasks such as semantic annotation or speech act classification highlights their practical usefulness, particularly in contexts where traditional resources are limited.

What emerges from these studies is not a single, unified picture, but a set of converging insights into the nature of AI-based language processing. The analysis show that such systems are capable of capturing a wide range of linguistic patterns across tasks, domains and languages. At the same time, they make visible the extent to which this capacity is grounded in distributional learning, shaped by data availability and bias and constrained in cases that require deeper contextual, cultural or pragmatic interpretation.

In this sense, the papers provide concrete empirical support for the broader claim that AI systems both reflect and amplify patterns of linguistic variation. They demonstrate that variation can be modeled at scale, but also that what is captured, how it is represented and where the limits lie depend crucially on the structure and composition of the underlying data. Rather than approximating a human-like linguistic system, these models instantiate a particular form of data-driven generalization, one that reproduces existing regularities while also making their unevenness more visible.

## 3 Outlook and future directions

The studies in this special issue make clear that current AI-based approaches to language are already capable of modeling a wide range of linguistic phenomena. What is now at stake is less whether these systems can produce linguistically plausible output and more how their behavior can be understood, controlled and critically evaluated in relation to linguistic theory.

One central direction concerns the systematic study of linguistic variation in AI systems. While variation is clearly present in model outputs, it is not yet well understood how it is internally represented, how stable it is across prompts and contexts or how it interacts with social and contextual factors such as register, identity and communicative setting. Future work needs to move beyond isolated case studies toward more controlled, comparative designs that make variation itself the object of investigation.

A second key area lies in the relationship between data, representation and linguistic knowledge. Current models derive their behavior from large-scale, unevenly distributed data, which raises the question of how different types of input shape what is learned and what remains inaccessible. This is particularly relevant for low-resource languages and non-standard varieties, where gaps in data lead directly to gaps in model performance. More work is needed on data curation, controlled input manipulation and the development of evaluation frameworks that make such asymmetries visible.

Third, the role of AI systems as tools in linguistic research requires closer methodological scrutiny. Using models as annotators, informants or generators of linguistic data opens new possibilities, but also introduces problems of reproducibility, stability and interpretability. Small changes in prompting can lead to different outputs and model updates can shift behavior over time. Establishing more transparent and standardized ways of working with these systems will be crucial if they are to become reliable instruments in empirical research.

Another promising direction concerns the interaction between surface performance and underlying generalization. The empirical investigations in this special issue show that models perform well on tasks such as annotation, translation or classification, yet struggle with phenomena that require deeper contextual or culturally grounded interpretation. This raises the question of what exactly is being learned and whether current architectures can be pushed beyond distributional approximation toward more robust forms of generalization.

Finally, there is growing scope for integrating computational approaches more closely with linguistic theory. Rather than treating AI systems as black-box tools, future work can use them to test hypotheses about language structure, variation and change, while at the same time using linguistic theory to better interpret model behavior. This two-way interaction has the potential to move the field beyond purely performance-driven evaluation toward a more explanatory understanding of language in AI.

What emerges from these directions is a shift in focus: from building ever more capable models to developing a clearer account of how these systems represent, reproduce and transform language. This includes not only their technical performance, but also the linguistic and social assumptions embedded in their design and training. In this sense, the study of AI and language is no longer only about improving systems, but about understanding what kind of language these systems make possible.

**References**

Abend, Omri, & Rappoport, Ari. 2013. Universal Conceptual Cognitive Annotation (UCCA). In Schuetze, Hinrich & Fung, Pascale & Poesio, Massimo (eds): *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 228–238. Sofia: Association for Computational Linguistics. https://aclanthology.org/P13-1023/ (last accessed on 10.05.2026).

Ananthaswamy, Anil. 2023. In AI, is bigger better? *Nature* 615. 202–205. https://doi.org/10.1038/d41586-023-00641-w

Baker, Collin Francis & Fillmore, Charles John &. Lowe, John B. 1998. In Boitet, Christian & Whitelock, Pete (eds): The Berkeley FrameNet Project. *Proceedings of COLING-ACL 1998*, 86–90. https://doi.org/10.3115/980845.980860

Banarescu, Laura & Bonial, Claire & Cai, Shu & Georgescu, Madalina & Griffitt, Kira & Hermjakob, Ulf & Knight, Kevin & Koehn, Philipp & Palmer, Martha & Schneider, Nathan. 2013. Abstract meaning representation for Sembanking. In Pareja-Lora, Antonio & Liakata, Maria & Dipper, Stefanie (eds): *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186. Sofia: Association for Computational Linguistics. https://aclanthology.org/W13-2322/ (last accessed on 08.05.2026)

Bender, Emily M. & Koller, Alexander. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Jurafsky, Dan & Chai, Joyce & Schluter, Natalie & Tetreault, Joel (eds): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, Emily M. & Gebru, Timnit & McMillan-Major, Angelina & Shmitchell, Shmargaret. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Berins, Laura. 2023. Judith Simon über Chatbots: »ChatGPT versteht nicht, es simuliert nur Sprache«. 31 January 2023. https://www.fr.de/kultur/gesellschaft/judith-simon-ueber-chatbots-chatgpt-versteht-nicht-es-simuliert-nur-sprache-92060094.html (last accessed on 01.04.2026)

Bommasani, Rishi & Hudson, Drew A. & Adeli, Ehsan & Altman, Russ & Arora, Simran & von Arx, Sydney & Bernstein, Michael S. & Bohg, Jeannette & Bosselut, Antoine & Brunskill, Emma & Brynjolfsson, Erik & Buch, Shyamal & Card, Dallas & Castellon, Rodrigo & Chatterji, Niladri & Chen, Annie & Creel, Kathleen & Davis, Jared Quincy & Demszky, Dorottya & Donahue, Chris & Doumbouya, Moussa & Durmus, Esin & Ermon, Stefano & Etchemendy, John & Ethayarajh, Kawin & Fei-Fei, Li & Finn, Chelsea & Gale, Trevor & Gillespie, Lauren & Goel, Karan & Goodman, Noah & Grossman, Shelby & Guha, Neel & Hashimoto, Tatsunori & Henderson, Peter & Hewitt, John & Ho, Daniel E. & Hong, Jenny & Hsu, Kyle & Huang, Jing & Icard, Thomas & Jain, Saahil & Jurafsky, Dan & Kalluri, Pratyusha & Karamcheti, Siddharth & Keeling, Geoff & Khani, Fereshte & Khattab, Omar & Koh, Pang Wei & Krass, Mark & Krishna, Ranjay & Kuditipudi, Rohith & Kumar, Ananya & Ladhak, Faisal & Lee, Mina & Lee, Tony & Leskovec, Jure & Levent, Isabelle & Li, Xiang Lisa & Li, Xuechen & Ma, Tengyu & Malik, Ali & Manning, Christopher D. & Mirchandani, Suvir & Mitchell, Eric & Munyikwa, Zanele & Nair, Suraj & Narayan, Avanika & Narayanan, Deepak & Newman, Ben & Nie, Allen & Niebles, Juan Carlos & Nilforoshan, Hamed & Nyarko, Julian & Ogut, Giray & Orr, Laurel & Papadimitriou, Isabel & Park, Joon Sung & Piech, Chris & Portelance, Eva & Potts, Christopher & Raghunathan, Aditi & Reich, Rob & Ren, Hongyu & Rong, Frieda & Roohani, Yusuf & Ruiz, Camilo & Ryan, Jack & Ré, Christopher & Sadigh, Dorsa & Sagawa, Shiori & Santhanam, Keshav & Shih, Andy & Srinivasan, Krishnan & Tamkin, Alex & Taori, Rohan & Thomas, Armin W. & Tramèr, Florian & Wang, Rose E. & Wang, William & Wu, Bohan & Wu, Jiajun & Wu, Yuhuai & Xie, Sang Michael & Yasunaga, Michihiro & You, Jiaxuan & Zaharia, Matei & Zhang, Michael & Zhang, Tianyi & Zhang, Xikun & Zhang, Yuhui & Zheng, Lucia & Zhou, Kaitlyn & Liang, Percy. 2021. On the opportunities and risks of foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2108.07258

Bommasani, Rishi & Liang, Percy & Lee, Tony. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences* 1525(1). 140–146. https://doi.org/10.1111/nyas.14925

Brocca, Nicola & Nuzzo, Elena & Wang-Kathrein, Joseph. 2026. AI-driven speech act annotation: accuracy andreproducibilityacross ChatGPT, LadderWeb and

LlaMA. In Catasso, Nicholas & Scharinger, Thomas (eds): *Natural Language and AI. New Perspectives for Linguistic Studies*. Special issue in *AI Linguistica* 3(1). 1–30. https://doi.org/10.62408/ai-ling.v3i1.33

Bui, Minh Duc & Holtermann, Carolin & Hofmann, Valentin & Lauscher, Anne & von der Wense, Katharina. 2025. Large language models discriminate against speakers of German dialects. In Christodoulopoulos, Christos & Chakraborty, Tanmoy & Rose, Carolyn & Peng, Violet (eds): *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 8223–8251. Suzhou: Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2509.13835

Catasso, Nicholas. 2026. Benchmarking AI acceptability and grammaticality in German: A study of ChatGPT and human judgments. In Catasso, Nicholas & Scharinger, Thomas (eds): *Natural Language and AI. New Perspectives for Linguistic Studies*. Special issue in *AI Linguistica* 3(1). 1–41. https://doi.org/10.62408/ai-ling.v3i1.35

Chang, Yupeng & Wang, Xu & Wang, Jindong & Wu, Yuan & Yang, Linyi & Zhu, Kaijie & Chen, Hao & Yi, Xiaoyuan & Wang, Cunxiang & Wang, Yidong & Ye, Wei & Zhang, Yue & Chang, Yi & Yu, Philip S. & Yang, Qiang & Xie, Xing. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15(3). 1–45. https://doi.org/10.1145/3641289

Chapelle, Carol A. & Beckett, Gulbahar H. & Ranalli, Jim (eds). 2024. *Exploring AI in Applied Linguistics*. Ames: Iowa State University Digital Press. https://doi.org/10.31274/isudp.230

Chomsky, Noam & Roberts, Ian & Watumull, Jeffrey. 2023. Noam Chomsky: The false promise of ChatGPT. *The New York Times*, 8 March 2023.

Degen, Judith. 2023. The Rational Speech Act Framework. *Annual Review of Linguistics* 9. 519–540. https://doi.org/10.1146/annurev-linguistics-031220-010811

Fleisig, Eve & Smith, Genevieve & Bossi, Madeline & Rustagi, Ishita & Yin, Xavier & Klein, Dan. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In Al-Onaizan, Yaser & Bansal, Mohit & Chen, Yun-Nung (eds): *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13541–13564. Miami: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.835

Ganguli, Deep & Hernandez, Danny & Lovitt, Liane & Askell, Amanda & Bai, Yuntao & Chen, Anna & Conerly, Tom & Dassarma, Nova & Drain, Dawn & Elhage, Nelson & El Showk, Sheer & Fort, Stanislav & Hatfield-Dodds, Zac & Henighan, Tom & Johnston, Scott & Jones, Andy & Joseph, Nicholas & Kernian, Jackson & Kravec, Shauna & Mann, Ben & Nanda, Neel &

Ndousse, Kamal & Olsson, Catherine & Amodei, Daniela & Brown, Tom & Kaplan, Jared & McCandlish, Sam & Olah, Christopher & Amodei, Dario & Clark, Jack.. 2022. Predictability and surprise in large generative models. In Isbell, Charles & Lazar, Seth & Oh, Alice & Xiang, Alice (eds): *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764. New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533229

Gauthier, Jon & Hu, Jennifer & Wilcox, Ethan & Qian, Peng & Levy, Roger. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In Celikyilmaz, Asli & Wen, Tsung-Hsien (eds): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 70–76. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.10

Guo, Zishan & Jin, Renren & Liu, Chuang & Huang, Yufei & Shi, Dan & Supryadi & Yu, Linhao & Liu, Yan & Li, Jiaxuan & Xiong, Bojian & Xiong, Deyi. 2023. Evaluating large language models: A comprehensive survey. *arXiv*. https://doi.org/10.48550/arXiv.2310.19736

Guo, Yanzhu & Conia, Simone & Zhou, Zelin & Li, Min & Potdar, Saloni & Xiao, Henry. 2025. Do large language models have an English accent? Evaluating and improving the naturalness of multilingual LLMs. In Che, Wanxiang & Nabende, Joyce & Shutova, Ekaterina & Pilehvar, Mohammad Taher (eds): *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1)*, 3823–3838. Vienna: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.193

Grieve, Jack & Bartl, Sara & Fuoli, Matteo & Grafmiller, Jason & Huang, Weihang & Jawerbaum, Alejandro & Murakami, Akira & Perlman, Marcus & Roemling, Dana & Winter, Bodo. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence* 7:1472411. 1–18. https://doi.org/10.3389/frai.2024.1472411

Hale, John & Stanojević, Miloš. 2024. Do LLMs learn a true syntactic universal? In Al-Onaizan, Yaser & Bansal, Mohit & Chen, Yun-Nung (eds): *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17106–17119. Miami: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.950

Kastrati, Muhamet & Imran, Ali Shariq & Hashmi, Ehtesham & Kastrati, Zenun & Daudpota, Sher Muhammad & Biba, Marenglen. 2025. Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with explainable artificial intelligence. *Engineering Applications of Artificial Intelligence* 149. 1–25. https://doi.org/10.1016/j.engappai.2024.108609

Hu, Jennifer & Gauthier, Jon & Qian, Peng & Wilcox, Ethan & Levy, Roger. 2020. A systematic assessment of syntactic generalization in neural language models. In Jurafsky, Dan & Chai, Joyce & Schluter, Natalie & Tetreault, Joel (eds): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1725–1744. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

Kelly-Holmes, Helen. 2022. Sociolinguistics in an increasingly technologized reality. *Sociolinguistica* 36. 99–110. https://doi.org/10.1515/soci-2022-0005

Kelly-Holmes, Helen. 2024. Artificial intelligence and the future of our sociolinguistic work. *Journal of Sociolinguistics* 28. 3–10. https://doi.org/10.1111/josl.12678

Khan, Aurangzeb & Baharudin, Baharum & Lee, Lam Hong & Khan, Khairullah. 2010. A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology* 1(1). 4–20. https://doi.org/10.4304/jait.1.1.4-20

Li, Wei & de Winter, Adrian. 2026. Where and how to improve English dialectal fairness. *arXiv*. https://doi.org/10.48550/arXiv.2603.15187

Ma, Bolei. 2024. Evaluating lexical aspect with large language models. In Kuribayashi, Tatsuki & Rambelli, Giulia & Takmaz, Ece & Wicke, Philipp & Oseki, Yohei (eds): *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 123–131. Bangkok: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.cmcl-1.11

Ma, Bolei & Li, Yuting & Zhou, Wei & Gong, Ziwei & Liu, Yang Janet & Jasinskaja, Katja & Friedrich, Annemarie & Hirschberg, Julia & Kreuter, Frauke & Plank, Barbara. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In Che, Wanxiang & Nabende, Joyce & Shutova, Ekaterina & Pilehvar, Mohammad Taher (eds): *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 8679–8696. Vienna: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.482

Marvin, Rebecca & Linzen, Tal. 2018. Targeted syntactic evaluation of language models. In Riloff, Ellen & Chiang, David & Hockenmaier, Julia & Tsujii, Jun'ichi (eds): *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192–1202. Brussels: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1151

Ismayilzada, Mete & Circi, Defne & Sälevä, Jonne & Sirin, Hale & Köksal, Abdullatif & Dhingra, Bhuwan & Bosselut, Antoine & Ataman, Duygu & Van Der Plas, Lonneke. 2025. Evaluating morphological compositional generalization in large language models. In Chiruzzo, Luis & Ritter, Alan & Wang, Lu (eds): *Proceedings of the 2025 Conference of the Nations of the*

*Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1270–1305. Albuquerque: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.78

Meier, Franz. 2026. Marking intersubjectivity in human-written and AI-generated editorials published in *Il Foglio*. In Catasso, Nicholas & Scharinger, Thomas (eds): *Natural Language and AI. New Perspectives for Linguistic Studies*. Special issue in *AI Linguistica* 3(1). 1–27. https://doi.org/10.62408/ai-ling.v3i1.65

Moutsis, Stavros N. & Ioakeimidou, Despoina & Tsintotas, Konstantinos A. & Evangelidis, Konstantinos & Nastou, Panagiotis E. & Tsolomitis, Antonis. 2025. Artificial intelligence for historical manuscripts digitization. *Engineering Proceedings* 107(1). 8. https://doi.org/10.3390/engproc2025107008

Ocumpaugh, Jaclyn & Liu, Xiner & Zambrano, Andres Felipe. 2025. Language models and dialect differences. In Martínez Monés, Alejandra & Mills, Caitlin & Jovanovic, Jelena & Ochoa, Xavier (eds.): *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*, 204–215, New York. Association for Computing Machinery. https://doi.org/10.1145/3636555.3636891

Ouyang, Long & Wu, Jeff & Jiang, Xu & Almeida, Diogo & Wainwright, Carroll L. & Mishkin, Pamela & Zhang, Chong & Agarwal, Sandhini & Slama, Katarina & Ray, Alex & Schulman, John & Hilton, Jacob & Kelton, Fraser & Miller, Luke & Simens, Maddie & Askell, Amanda & Welinder, Peter & Christiano, Paul & Leike, Jan & Lowe, Ryan. 2022. Training language models to follow instructions with human feedback. *arXiv*. https://doi.org/10.48550/arXiv.2203.02155

Park, Dojun & Lee, Jiwoo & Park, Seohyun & Jeong, Hyeyun & Koo, Youngeun & Hwang, Soonha & Park, Seonwoo & Lee, Sungeun. 2024. MultiPragEval: Multilingual pragmatic evaluation of large language models. In Hupkes, Dieuwke & Dankers, Verna & Batsuren, Khuyagbaatar & Kazemnejad, Amirhossein & Christodoulopoulos, Christos & Giulianelli, Mario & Cotterell, Ryan (eds.): *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, 96–119, Miami: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.genbench-1.8

Petroni, Fabio & Rocktäschel, Tim & Riedel, Sebastian & Lewis, Patrick & Bakhtin, Anton & Wu, Yuxiang & Miller, Alexander. 2019. Language models as knowledge bases? In Inui, Kentaro & Jiang, Jing & Ng, Vincent & Wan, Xiaojun (eds): *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

*Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. Hong Kong: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1250

Platzgummer, Verena & McCrae, John & Ahmadi, Sina. 2026. Versteasch du mi? Computational and socio-linguistic perspectives on GenAI, LLMs, and non-standard language. *arXiv*. https://doi.org/10.48550/arXiv.2603.28213

Radford, Alec & Wu, Jeff & Child, Rewon & Luan, David & Amodei, Dario & Sutskever, Ilya. 2019. Language models are unsupervised multitask learners. OpenAI technical report. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (last accessed on 08.04.2026)

Rafailov, Rafael & Sharma, Archit & Mitchell, Eric & Ermon, Stefano & Manning, Christopher D. & Finn, Chelsea. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv*. https://doi.org/10.48550/arXiv.2305.18290

Rahman, Gohar. 2026. Automating semantic annotation in low-resource languages: Evaluating GPT-4 for Urdu NLP. In Catasso, Nicholas & Scharinger, Thomas (eds): *Natural Language and AI. New Perspectives for Linguistic Studies*. Special issue in *AI Linguistica* 3(1). 1–26. https://doi.org/10.62408/ai-ling.v3i1.40

Ramos, Juan. 2003. Using TF-IDF to determine word relevance in document queries. In Littman, Michael (ed): *Proceedings of the First Instructional Conference on Machine Learning*, 133–142. Piscataway: Rutgers University.

Rogers, Anna & Kovaleva, Olga & Rumshisky, Anna. 2020. A primer in BERTology. In Johnson, Mark & Roark, Brian & Nenkova, Ani (eds): *Transactions of the Association for Computational Linguistics* 8. 842–866. Cambridge, MA: MIT Press. https://doi.org/10.1162/tacl_a_00349

Rondini, Silvia. 2026. LLMs and meaning. https://philarchive.org/rec/RONLAM-2 (last accessed 08.04.2026)

Salton, Gerard & Buckley, Christopher. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5). 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Schüle, Martin. 2025. On the semantics of large language models. *Intellectica* 81. 15–36.

Schut, Lisa & Gal, Yarin & Farquhar, Sebastian. 2025. Do multilingual LLMs think in English? *arXiv*. https://doi.org/10.48550/arXiv.2502.15603

Sleeman, Petra. 2026. ChatGPT as a linguistic informant. A comparison of human and AI-generated translations. In Catasso, Nicholas & Scharinger, Thomas (eds): *Natural Language and AI. New Perspectives for Linguistic Studies*. Special issue in *AI Linguistica* 3(1). 1–28. https://doi.org/10.62408/ai-ling.v3i1.34

Srivastava, Aarohi & Rastogi, Abhinav & Rao, Abhishek & Shoeb, Abu Awal Md & Abid, Abubakar & Fisch, Adam & Brown, Adam R. & Santoro, Adam & Gupta, Aditya & Garriga-Alonso, Adrià & Kluska, Agnieszka & Lewkowycz, Aitor & Agarwal, Akshat & Power, Alethea & Ray, Alex et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* 5. 1–95. https://doi.org/10.48550/arXiv.2206.04615

Székely, Éva & Miniota, Jura & Hejná, Michaela. 2025. Will AI shape the way we speak? In Torres, Maria Ines & Matsuda, Yuki & Callejas, Zoraida & del Pozo, Arantza & D'Haro, Luis Fernando (eds): *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, 335–340. https://doi.org/10.18653/v1/2025.iwsds-1.40

Tamkin, Alex & Brundage, Miles & Clark, Jack & Ganguli, Deep. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2102.02503

Thomas, Binu & Joseph, Jeena. 2025. The syntax of power: how large language models recode social hierarchies. *AI & Society* 41. 1339–1340. https://doi.org/10.1007/s00146-025-02520-6

van Schijndel, Marten & Linzen, Tal. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science* 45(6). e12988. 1–31. https://doi.org/10.1111/cogs.12988

Wang, Alex & Singh, Amanpreet & Michael, Julian & Hill, Felix & Levy, Omer & Bowman, Samuel R. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, Tal & Chrupała, Grzegorz & Alishahi, Afra (eds): *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5446

Wang, Alex & Pruksachatkun, Yada & Nangia, Nikita & Singh, Amanpreet & Michael, Julian & Hill, Felix & Levy, Omer & Bowman, Samuel R. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in Neural Information Processing Systems 32*. https://doi.org/10.48550/arXiv.1905.00537

Warstadt, Alex & Parrish, Alicia & Liu, Haokun & Mohananey, Anhad & Peng, Wei & Wang, Sheng-Fu & Bowman, Samuel R. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In Ettinger, Allyson & Jarosz, Gaja & Pater, Joe (eds): *Proceedings of the Society for Computation in Linguistics 2020*, 437–438. New York: Association for Computational Linguistics. https://doi.org/10.7275/zejz-qs04

Wei, Jason & Tay, Yi & Bommasani, Rishi & Raffel, Colin & Zoph, Barret & Borgeaud, Sebastian & Yogatama, Dani & Bosma, Maarten & Zhou, Denny & Metzler, Donald & Chi, Ed H. & Hashimoto, Tatsunori & Vinyals, Oriol & Liang, Percy & Dean, Jeff & Fedus, William. 2022. Emergent abilities of large language models. *arXiv*. http://arxiv.org/abs/2206.07682

Wiedner, Marinus & Schöffel, Matthias. 2026. Simulating the evolution of grammatical gender from Latin to Old Occitan: A computational approach using LSTM with Attention. In Catasso, Nicholas & Scharinger, Thomas (eds): *Natural Language and AI. New Perspectives for Linguistic Studies*. Special issue in *AI Linguistica* 3(1). 1–24. https://doi.org/10.62408/ai-ling.v3i1.66

Wilcox, Ethan & Vani, Pranali & Levy, Roger. 2021. A targeted assessment of incremental processing in neural language models and humans. In Zong, Chengqing & Xia, Fei & Li, Wenjie & Navigli, Roberto (eds): *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Vol. 1)*, 939–952. Online: Association for Computational Linguistics. 10.18653/v1/2021.acl-long.76

Xiong, Ying & Wu, Shiyu. 2025. Do large language models learn like humans: Interleaved and spaced practice in morphological learning. *Acta Psychologica* 260. 105518. 1–15. https://doi.org/10.1016/j.actpsy.2025.105518

Young, Tom & Hazarika, Devamanyu & Poria, Soujanya & Cambria, Erik. 2018. Recent trends in deep learning-based NLP. *IEEE Computational Intelligence Magazine* 13(3). 55–75. https://doi.org/10.1109/MCI.2018.2840738

Zhou, Houquan & Hou, Yang & Li, Zhenghua & Wang, Xuebin & Wang, Zhefeng & Duan, Xinyu & Zhang, Min. 2023. How well do large language models understand syntax? *arXiv*. https://doi.org/10.48550/arXiv.2311.08287