Language Optionality and Variation
Festschrift in Honor of Anna Cardinaletti
edited by Chiara Branchini, Gianluca Lebani,
Lara Mantovan, Francesca Volpato

The Consequences of Large Language Models for Linguistic Theory

Alessandro Lenci
Università di Pisa, Italia

Abstract Large Language Models (LLMs) have revolutionized the relationship between linguistics and artificial intelligence. This paper argues that, despite their remarkable performance, LLMs are not linguistic theories but tools that can both challenge and advance them. I will show that current models diverge from human cognition and that linguistics remains essential not only for interpreting how these systems work but for guiding the design of more cognitively plausible and interpretable language models.

Keywords Language models. Linguistic theory. Artificial intelligence. Distributional learning. Cognitive plausibility.

Summary 1 Introduction. – 2 Linguistics and Machines Before and After LLMs. – 3 They Speak Like Us, but They Are Not Like Us. – 4 What Linguistics Can (and Must) Do for Large Language Models. – 5 Conclusions.

1 Introduction

Language Models – neural networks trained to predict the most probable word given the preceding context – have emerged as key players in the contemporary landscape of artificial intelligence (AI) and cognitive science. Their extraordinary performances, often comparable to human abilities and in some cases even 'superhuman' (especially in their larger versions, the so-called Large Language Models, LLMs), have sparked an intense debate on the nature, acquisition, and processing of human language (Chesi 2025; Futrell, Mahowald 2025).

Edizioni Ca'Foscari

LiVVaL 9
e-ISSN 2974-6574 | ISSN 2974-6981
ISBN [ebook] 979-12-5742-086-4 | ISBN [print] 979-12-5742-121-2

Peer review | Open access
Submitted 2025-11-10 | Accepted 2026-01-28 | Published 2026-11-17
© 2026 Lenci | CC BY 4.0
DOI 10.30687/979-12-5742-086-4/008

For decades, natural language processing (NLP), the field of AI devoted to building machines endowed with all aspects of human intelligence connected to the use and understanding of natural language, sought to construct 'talking machines' solely through the implementation of linguistic theories, either in the form of manually crafted rules or via linguistically annotated data used to train supervised machine-learning algorithms. Today, for the first time, we are witnessing a new era: A generation of computational models has emerged, the LLMs, that acquire highly sophisticated and unprecedented linguistic abilities exclusively from the distributional statistics extracted from massive amounts of unannotated data. A kind of 'Copernican revolution' has thus taken place: rather than implementing linguistic theories, computational models are increasingly conceived as linguistic theories themselves (Piantadosi 2024). This shift is accompanied by the diffusion of a scientific paradigm in which the linguistic abilities of computational models are seen simply as emergent properties arising from the ever-growing mass of training data and the scale of neural architectures.

In this article, I will outline several arguments in favor of a different view of the relationship between Language Models and linguistic theory, which can be summarized as follows:

1. Language Models are not, in themselves, linguistic theories (Katzir 2023), but they possess extraordinary potential as tools for the investigation and validation of scientific hypotheses in linguistics and cognitive science;[1]
2. Existing Language Models, especially the largest ones, are not yet ready to fully assume such a role. Although they are technologically powerful systems with enormous applicative potential, significant gaps remain relative to the requirements that a scientifically valid model of human language acquisition and processing should satisfy (Frank 2023);
3. Linguistics can and must offer a fundamental contribution to the evolution of Language Models, transforming them into tools for the study of natural language and guiding the development of a new generation of models more directly inspired by the functioning and acquisition of language in humans, models capable of overcoming the structural limitations of current LLMs and their dependence on 'superhuman' amounts of training data.

1 Goldberg 2024; Portelance, Jasbi 2024; Baggio 2025; Binz et al. 2025; Futrell, Mahowald 2025.

2 Linguistics and Machines Before and After LLMs

To fully grasp the scope and implications of the ongoing transformation in the relationship between linguistics and AI, consider the following example:

(1) Yesterday evening Ann asked Mary if she wanted to go to the cinema with her. Mary replied: "I have to get up early tomorrow." What did Mary mean by that sentence?

The correct answer is that Mary politely declined Ann's invitation, justifying herself by saying that she needed to wake up early the next day. In other words, Mary performed an indirect speech act: she used an assertion to implicitly communicate a refusal.

Before the advent of LLMs, designing an artificial system capable of answering the question in (1) involved formalizing and then implementing the knowledge and processes underlying human linguistic competence. For instance, linguistic theory tells us that to answer question (1) correctly, one must activate several types of knowledge:

1. reconstructing the morphological and syntactic structure of the sentence: for example, recognizing that the implied subject of the infinitival clause *to go* is Mary, that the pronoun her refers to Ann, that the subordinate clause governed by ask is an indirect question, and so forth;
2. correctly identifying the meanings of lexical items: for instance, interpreting the phrasal verb *get up* as rising from the bed after sleeping;
3. integrating this information with encyclopedic and pragmatic knowledge: for instance, that going to an evening movie implies returning home late, which might conflict with the need to get up early the next morning;
4. inferring the communicative intent associated with Mary's utterance on the basis of its meaning and all contextual information.

Until the 1990s, such knowledge was primarily formalized through rule-based systems operating on symbolic representations (such as context-free grammars, feature structures, semantic networks, etc.), manually crafted by computational linguists (Lenci, Vestrucci 2024). With the rise of the data-driven statistical paradigm, linguistic knowledge began to be represented as annotated data labeled with categories and structures, used to train supervised machine-learning algorithms.

In this context, a typical system designed to answer question (1) might consist of a classifier of communicative intents trained on

annotated corpora to distinguish, for instance, between different types of illocutionary acts expressed by sentences. Such a classifier would use syntactic and semantic information provided by specialized models designed to perform tasks such as syntactic parsing, anaphora resolution, and word-sense disambiguation, each trained, in turn, on corpora annotated with the relevant categories via machine-learning algorithms.

The shift from rule-based models to statistical and deep learning methods based on neural networks (with multiple hidden layers) represented a significant advancement, providing artificial systems with greater robustness, generalization capacity, and adaptability. The role of linguistics shifted from creating "grammars" to creating "annotated data" from which computational models could learn linguistic representations. However, models of this generation, including neural networks, continued to process language through explicit representations of categories supplied a priori by linguistic theory.

With the advent of LLMs, a substantial paradigm change occurred. These models demonstrate the ability to interact in natural language with a level of quality significantly superior to previous generations, despite not being explicitly trained to recognize or manipulate formal linguistic categories. For instance, when presented with a prompt consisting of question (1), a LLM such as ChatGPT provides the following answer:

(2) Mary's reply — "I have to get up early tomorrow" — is an indirect way of saying she can't (or doesn't want to) go to the cinema because she needs to wake up early the next day. In other words, she's politely refusing Ann's invitation and giving a reason for it: she has an early commitment in the morning.

An interesting point is that ChatGPT not only provides the correct answer but also the appropriate metalinguistic explanation, explicitly labeling Mary's response as an "indirect refusal". Similarly, the system can correctly identify the implicit subject of a subordinate clause, while also providing a detailed syntactic analysis of the sentence in linguistic terms:

(3) Prompt: What is the subject of "to go" in the sentence "Yesterday evening Ann asked Mary if she wanted to go to the cinema with her"?
ChatGPT: In the sentence "Yesterday evening Ann asked Mary if she wanted to go to the cinema with her," the subject of "to go" is Mary.
Here's why: The clause "if she wanted to go to the cinema with her" depends on "asked." The verb "wanted" has the subject "she" (which refers to Mary). The infinitive "to go" is part of what Mary wanted — so the understood (logical) subject of "to go" is the same as the subject of "wanted", i.e. Mary.

These answers are quite surprising, at least at first glance, given the way LLMs work. They consist of deep neural networks – Transformers (Vaswani et al. 2017) – which operate on continuous vector representations (known as embeddings). These embeddings *implicitly* encode knowledge about linguistic structures entirely derived from statistical regularities extracted from massive quantities of linguistic data (Lenci, Sahlgren 2023; Lenci et al. 2025).

Unlike previous computational paradigms, LLMs are not trained to manipulate grammatical categories or syntactic structures defined by linguistic theory. They are trained exclusively to solve the task of predicting the most probable next word given a preceding sequence of words. The linguistic abilities displayed by LLMs are therefore considered emergent abilities (Wei et al. 2022), since they do not stem from pre-specified linguistic knowledge but arise from the inductive learning process inherent in word-sequence prediction.

However, the advanced performance of models like ChatGPT cannot be attributed solely to basic predictive training (the so-called pretraining phase), but also to subsequent refinement processes (post-training) such as instruction tuning and reinforcement learning from human feedback (Tie et al. 2025), which play a crucial role in aligning the model's responses with human communicative and social expectations. It is important to note, however, that these refinement procedures do not involve explicit instruction about grammatical categories or linguistic rules: no model is ever explicitly taught what a grammatical subject is, for example.

LLMs therefore represent the first 'talking machines' developed without any direct contribution from traditional linguistic theories about the organization and functioning of human language. They mark a significant epistemological discontinuity relative to previous computational paradigms, raising new questions both in computational modeling and in our theoretical understanding of linguistic competence.

3 They Speak Like Us, but They Are Not Like Us

Although LLMs may astonish us with their linguistic abilities, these machines are profoundly different from humans in how such abilities are acquired and realized.

First, the training data available to LLMs differ from those accessible to humans, both quantitatively and qualitatively. LLMs are trained on datasets that are vastly larger than the linguistic input human speakers are ever exposed to. For instance, Lan et al. (2024) estimate that the data used for the first version of ChatGPT correspond to roughly 36,540 human years of linguistic exposure, assuming eleven million words as the average yearly linguistic input

for a person (Hart, Risley 1995). Moreover, the training corpora consist almost entirely of written texts scraped from the web (Liu et al. 2024). Therefore, they lack key aspects of the data humans use to acquire language, such as oral and visual input, conversational dynamics, and references to extra-linguistic and social contexts.

Second, LLMs rely exclusively on probabilistic prediction mechanisms. While distributional learning undoubtedly plays an important role in linguistic development (Lenci, Sahlgren 2023), other data, cognitive mechanisms, and structures are indispensable as well (Resnik 2025). This crucially includes extralinguistic experiences in multimodal communicative contexts, as well as learning biases that help the learner to rapidly converge towards the target linguistic structures in the face of highly noisy input data. Language is an extremely sophisticated function that has evolved to optimize the use of the limited cognitive and physiological resources humans possess for acquiring it. It is highly plausible that the outcome of this evolution is a system of cognitive constraints that guide the acquisition, representation, and organization of linguistic knowledge, enabling humans to generalize effectively even from limited input.

The same holds true for how LLMs solve any task, either linguistic or otherwise. ChatGPT produces the answer in (2) because that string of text is the most probable continuation of the text string in (1) used as a prompt. Word-sequence probability is, in fact, the sole behavioral criterion guiding the system. McCoy et al. (2023) show that even the mathematical problem-solving abilities of LLMs depend on the frequency of occurrence of the relevant functions in their training data. These systems are therefore constrained by their purely probabilistic nature, being models trained simply to predict word sequences. This is true even for the more recent reasoning models, a new generation of LLMs optimized specifically for reasoning, and yet they still exhibit substantial limitations in problem-solving capabilities (McCoy et al. 2024; Shojaee et al. 2025).

These findings are consistent with the view that, while probabilistic processes play an important role in human cognition, they are not the only form of reasoning of which we are capable (Kahneman 2011). Human reasoning, in its more complex forms, clearly relies on inferential schemes that may not be learnable solely from the statistical co-occurrence patterns on which LLMs are based. Therefore, the fact that ChatGPT produces human-like answers does not mean it does so using human-like mechanisms. As Shanahan (2024) puts it, although they seem to speak like us, LLMs are new entities that are, in reality, not like us, exhibiting "a patchwork of less-than-human with superhuman capacities, of uncannily human-like with peculiarly inhuman behavior" (p. 79).

Mitchell and Krakauer (2023) argue that Large Language Models constitute a new class of systems within the 'technological zoo,'

achieving linguistic competence through fundamentally non-human processes. From an engineering perspective, this is not a problem: technological innovations often diverge from biological mechanisms in the way they realize a given function, just as airplanes differ from birds in how they fly. Indeed, the history of human technological progress is full of examples of tools capable of performing tasks once reserved for humans or animals, even though they rely on mechanisms entirely different from those found in nature.

The case of flight is exemplary. As long as humans tried to imitate bird wings directly, every attempt failed. The Wright brothers' first successful flight at Kitty Hawk (North Carolina) on December 17, 1903, became possible precisely because of a machine built on principles far removed from those of animal flight: rigid wings, an engine for propulsion, and a structure entirely unlike that of a living organism. 'Flying machines,' therefore, share certain elements with birds but are radically different in how they realize the act of flying. Something similar occurred on November 30, 2022, the day OpenAI released ChatGPT. Thus, describing the emergence of LLMs and generative AI as a 'Copernican revolution' is entirely appropriate. Copernicus's scientific revolution displaced humanity from the center of the universe; similarly, LLMs have displaced humans from the center of language. Linguistic abilities now appear to manifest themselves in machines that, while inspired by human processes such as neural modeling and distributional learning, learn language in ways that greatly differ from human ones.

Likewise, linguistics, as the scientific study of human language, seems to have been removed from the center of research on 'talking machines.' Yet this decentering is only apparent, and hopefully only temporary. There are at least two reasons why linguistics can and must continue to contribute to AI, albeit in a new way compared to the past.

4 What Linguistics Can (and Must) Do for Large Language Models

The mainstream approach to LLM design assumes that the continual increase in training data and the size of neural architectures is sufficient to enable artificial systems to acquire linguistic and reasoning abilities comparable to those of humans. Such abilities are said to emerge from linguistic data once a critical mass is reached (Wei et al. 2022). The scaling up of models and datasets has indeed been decisive for recent progress, but it carries important theoretical and practical consequences:

- LLMs are now widely used to validate hypotheses in linguistic theory (a prominent case being the Poverty of the Stimulus argument; Lan et al. 2024; Piantadosi 2024). However, the

"superhuman" quantities of data used to train increasingly large systems raise significant doubts about their actual usability for cognitive and linguistic research, and about their validity as models of human language acquisition and use.

- The growth of LLMs is not unlimited, since the amount of textual data available is finite (leading to the increasing use of synthetic linguistic data produced by the models themselves; Villalobos et al. 2024). Moreover, their scaling is unsustainable, given the immense computational resources required both for their training and at inference, which entail extremely high environmental costs and a deeply concerning carbon footprint (Ren et al. 2024).

As discussed in Section 2, it is not uncommon for a machine to perform the same functions as a biological organism using different mechanisms, but one must ask whether it does so with equal efficiency. From this observation arises a first fundamental role for the language sciences in AI: to contribute to the development of a new generation of talking machines that are more sustainable and efficient, inspired by the constraints and principles governing human language.

A promising line of research focuses on the development of so-called Baby Language Models (BabyLMs), trained on data that are more similar, both quantitatively and qualitatively, to the input available to human learners. For instance, the BabyLM Challenge (Warstadt et al. 2023) created a small-scale dataset (approximately one hundred million words), roughly comparable to the linguistic input received by a ten-year-old child, constructed from transcriptions of spoken language, a significant portion of which consists of child-directed speech. Suozzi et al. (2025) developed BAMBI (BAby language Models Bootstrapped for Italian), a series of BabyLMs trained on samples of Italian data selected to model the linguistic input available to learners at different age ranges, from 0 to 24 years. In this context, the contribution of linguistics lies both in the selection and collection of cognitively plausible training data, and in the study of how to integrate the predictive mechanisms of Language Models with learning biases and architectural constraints (e.g., hierarchical structuring, locality constraints, minimum description length, etc.) capable of guiding the acquisition of linguistic structures even in environments with limited training data (Zamparelli 2025).

One of the most significant contributions of Language Models has been to demonstrate that the information available from the statistical distribution of linguistic input is far richer than previously imagined (Lenci 2023). This is an extremely important finding for the study of human language acquisition. However, the richness and power of distributional information do not rule out the possibility that human ontogenetic development also relies on a priori biases (Mandler

2004, Carey 2009) or specialized mechanisms (Tomasello 2003; Yang et al. 2017) that structure and optimize stimuli and cognitive resources during language learning, enhancing generalization capacity. Natural language is the product of a long evolutionary process: a remarkable communicative faculty that has developed within the cognitive and physiological constraints of our species. A better understanding of the linguistic and extralinguistic constraints that govern the learning and use of language, such as the ability to generalize from minimal data, or to integrate multimodal stimuli, can help us design computational models that are more efficient, less resource-intensive, and potentially even more capable, given the still-significant limitations of current systems (Mahowald et al. 2024; Shojaee et al. 2025).

The second contribution that linguistics can make to AI concerns the crucial issue of interpretability in Language Models, for which no clear explanatory theory yet exists. We have created machines with extraordinary linguistic abilities without fully understanding why they succeed when they do, or why they fail when they do not. This is due both to the immense architectural complexity of LLMs and to the implicit nature of the distributed representations they employ, which make them genuine black boxes. We do not know exactly what information is encoded in their billions of parameters, nor do we possess a causal theory linking a model's internal states to its observed behaviors.

Like other complex systems, LLMs exhibit non-locality: minimal linguistic variations in a prompt's phrasing can produce completely different responses or turn a correct answer into an error. The limited interpretability of LLMs is especially problematic when they are used as scientific models of human language. A scientific model of a process is expected to explain its causal structure, but LLMs still lack the level of interpretability required to provide a causal and theoretically grounded account even of their own linguistic behavior.

The same observable linguistic behavior can, in fact, result from fundamentally different internal strategies or representational formats (Warstadt, Bowman 2022). It is therefore essential to understand not only whether a model succeeds or fails, but also how and why it produces a given output. To address the challenge posed by the opacity of LLMs, a new research area has emerged: mechanistic interpretability (López-Otal et al. 2025). This focuses on analyzing the internal activations, representational structures, and mechanisms of Language Models.

In recent years, considerable effort has been devoted to interpreting the linguistic knowledge encoded in Language Models, particularly those based on the Transformer architecture. However, research has not yet yielded a comprehensive, coherent, and exhaustive account of how linguistic information is acquired, represented, and utilized by

LLMs. Nevertheless, recent advances in mechanistic interpretability are refining the tools available to researchers and promise to finally illuminate the black box of modern artificial systems. Following the distinction between formal and functional linguistic competence proposed by Mahowald et al. (2024), Hanna et al. (2025) conducted mechanistic interpretability analyses on several Large Language Models to determine whether a clear separation exists between functional and formal circuits, adapting the evaluation tasks to represent both types of competence. Hanna and Mueller (2025) began from the question of whether LLMs process sentences incrementally, similarly to humans, through an investigation of garden-path constructions, showing that such a parallel exists, though only in a limited form. Boguraev et al. (2025) examined the causal mechanisms underlying English filler-gap constructions, hypothesizing relevant similarities across instances (Sag 2010). Their results reveal that LLMs share partial causal structures among different filler-gap constructions, consistent with the relevant linguistic theory, while also identifying new factors that could help refine that theory, demonstrating how the interaction between linguistic theorization and mechanistic analysis can yield stronger hypotheses about human language.

Rodriguez et al. (2025) investigated the mechanisms governing the inheritance of properties in Language Models, providing mechanistic insights into how LLMs reason and connect concepts, while also suggesting new psycholinguistic experiments with human participants. Causal interpretability has also been used to explore phonological phenomena such as assimilation in neural speech-recognition models, allowing for comparisons between computational and human processing (Pouw et al. 2024). Other studies have focused on the geometry of embedding vector spaces to assess the internal conceptual organization performed by LLMs. These studies suggest that models partially encode concepts and relations as directional vectors within their representational spaces (Park et al. 2024; Proietti, Lenci 2025), although the organization of these spaces does not fully align with the semantic structures traditionally posited in linguistics.

Our understanding of the internal mechanisms within the black box of Language Models is still in its infancy and is hindered by the immense complexity of the largest models. Yet success in this endeavor cannot occur without the contribution of the language sciences, whose task is to formulate hypotheses about the structures and processes we seek to identify in these models. In this sense, linguistics remains the indispensable light needed to illuminate the black box and to guide our exploration within it.

5 Conclusions

The rise of generative AI and LLMs has been interpreted by many as either the long-awaited or the much-feared end of existing theories of language. This perceived 'end' affects not only Minimalist Generative Grammar, although it is often the primary target, but theoretical linguistics as a whole.

The 'Copernican revolution' brought about by LLMs has profoundly transformed the role of the language sciences within AI, since machines exhibiting linguistic abilities nearly indistinguishable from those of humans have been built without directly implementing the structures, constraints, and processes theorized by linguistics. This has led some to conclude that we no longer need linguistic theory at all, and that LLMs themselves are the best theory of language currently available (Piantadosi 2024), given their exceptional performance acquired purely through inductive learning from training data via probabilistic prediction mechanisms.

The arguments presented here point in precisely the opposite direction (for a similar position cf. Chesi 2025; Futrell, Mahowald 2025). The success of LLMs makes a theory of language more necessary than ever. The Wright brothers' first flight also occurred before the science of aerodynamics was fully established; likewise, we now need new theories to explain these 'talking machines.' We do not yet know what shape such theories will take, but that does not mean they cannot be formulated. We are just at the beginning, as 1903 marked the beginning of the history of aeronautics. In the age of generative AI, linguistic theory plays a crucial role:

i. to help design artificial systems that rely less on data quantity, by incorporating the constraints and biases that enable humans to learn language with limited data and cognitive resources;
ii. to understand whether and how neural models represent the structures of natural language, equipping themselves with new theoretical and methodological tools capable of grappling with their complexity.

At the same time, if LLMs cannot dispense with linguistics, it is equally true that linguistics itself cannot remain unchanged in the face of LLMs. These models can become virtual laboratories for testing and falsifying theories and hypotheses about the nature of language itself. However, linguistics will have to relinquish many of its long-held certainties to adapt to the new universe of Language Models. We need theories of language capable of interfacing with complex models that operate on sub-symbolic and continuous structures, where information is represented through overlapping states and entanglement (Huang et al. 2024) reminiscent of quantum systems. The convergence of

neural Language Models and linguistic theory may, in fact, enable linguistics to take a further step toward the full realization of the Galilean scientific method, whereby the hypotheses of linguistic theory can be experimentally tested through computational models simulating the processes of language acquisition and processing, analogous to what occurs in the physics of subatomic states.

This integration between Language Models and linguistic theory is also necessary on practical and technological grounds. It cannot be abandoned if we truly wish to understand AI systems and move beyond mere text generation toward authentic forms of communication. This transition is essential for defining the role of LLMs in our future. In the case of flight, the goal was not to imitate birds but to make aerial mobility possible; for language, by contrast, we seek machines capable not only of producing imitations of human texts, but of engaging in social interaction and communication with us, sharing knowledge, reasoning, and values. This is crucial, given the increasing presence of AI in our daily lives.

Bibliography

Baggio, G. (2025). "Could Machine Learning Help to Build a Unified Theory of Cognition?". *Nature*, 644, 881-3.

Binz, M. et al. (2025). "A Foundation Model to Predict and Capture Human Cognition". *Nature,* 644, 1002-9.

Boguraev, S.; Potts, C.; Mahowald, K. (2025). "Causal Interventions Reveal Shared Structure Across English Filler-Gap Constructions". *arXiv*, 2505.16002. https://arxiv.org/abs/2505.16002.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Chesi, C. (2025). "Is It the End of (Generative) Linguistics as We Know It?". *Italian Journal of Linguistics*, 37(1), 3-44.

Frank, M.C. (2023). "Bridging the Data Gap Between Children and Large Language Models". *Trends in Cognitive Sciences*, 27(11), 990-2.

Futrell, R.; Mahowald, K. (2025). "How Linguistics Learned to Stop Worrying and Love the Language Models". *Behavioral and Brain Sciences*, 49, 1-80.

Goldberg, A.E. (2024). "Usage-Based Constructionist Approaches and Large Language Models". *Constructions and Frames*, 16(2), 220-54.

Hanna, M.; Belinkov, Y.; Pezzelle, S. (2025). "Are Formal and Functional Linguistic Mechanisms Dissociated in Language Models?". *Computational Linguistics*, 52(1), 43-83.

Hanna, M.; Mueller, A. (2025). "Incremental Sentence Processing Mechanisms in Autoregressive Transformer Language Models". *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3181-203.

Hart, B.; Risley, T.R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Paul H. Brookes.

Huang, J.; Wu, Z.; Potts, C.; Geva, M.; Geiger, A. (2024). "RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations".

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 8669-87.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Katzir, R. (2023). "Why Large Language Models Are Poor Theories of Human Linguistic Cognition: A Reply to Piantadosi". *Biolinguistics*, 17.

Lan, N.; Chemla, E.; Katzir, R. (2024). "Large Language Models and the Argument from the Poverty of the Stimulus". *Linguistic Inquiry*, 1-28. Lingbuzz 006829.

Lenci, A. (2023). "Understanding Natural Language Understanding Systems". *Sistemi Intelligenti*, 2, 277-302.

Lenci, A.; Sahlgren, M. (2023). *Distributional Semantics*. Cambridge: Cambridge University Press.

Lenci, A.; Vestrucci, A. (2024). "Artificial Intelligence and Language". Boobalan, R.; Shrivastava, N.; Kiruthigadevi, N.; Ravichandran, P. (eds), *International Encyclopedia of Language and Linguistics*. 3rd ed. Amsterdam: Elsevier.

Lenci, A.; Auriemma, S.; Miliani, M. (2025). *Linguistica computazionale. Natural language processing e intelligenza artificiale*. Milano: Hoepli.

Liu, Y.; Cao, J.; Liu, C.; Ding, K.; Jin, L. (2024). "Datasets for Large Language Models: A Comprehensive Survey". *arXiv*, 2402.18041. https://arxiv.org/abs/2402.18041.

López-Otal, M.; Gracia, J.; Bernad, J.; Bobed, C.; Pitarch-Ballesteros, L.; Anglés-Herrero, E. (2025). "Linguistic Interpretability of Transformer-Based Language Models: A Systematic Review". *arXiv*, 2504.08001. https://arxiv.org/abs/2504.08001.

Mahowald, K.; Ivanova, A.A.; Blank, I.A.; Kanwisher, N.; Tenenbaum, J.B.; Fedorenko, E. (2024). "Dissociating Language and Thought in Large Language Models". *Trends in Cognitive Sciences*, 28(6), 517-40.

Mandler, J.M. (2004). *Foundations of Mind*. Oxford: Oxford University Press.

McCoy, R.T.; Yao, S.; Friedman, D.; Hardy, M.; Griffiths, T.L. (2023). "Embers of Autoregression: Understanding Large Language Models Through the Problem They Are Trained to Solve". *arXiv*, 2309.13638. https://arxiv.org/abs/2309.13638.

McCoy, R.T.; Yao, S.; Friedman, D.; Hardy, M.D.; Griffiths, T.L. (2024). "When a Language Model Is Optimized for Reasoning, Does It Still Show Embers of Autoregression? An Analysis of OpenAI o1". *arXiv*, 2410.01792. https://arxiv.org/abs/2410.01792.

Mitchell, M.; Krakauer, D.C. (2023). "The Debate over Understanding in AI's Large Language Models". *Proceedings of the National Academy of Sciences*, 120(13), e2215907120.

Park, K.; Choe, Y.J.; Veitch, V. (2024). "The Linear Representation Hypothesis and the Geometry of Large Language Models". *arXiv*, 2311.03658. https://arxiv.org/abs/2311.03658.

Piantadosi, S.T. (2024). "Modern Language Models Refute Chomsky's Approach to Language". Gibson, E.; Poliak, M. (eds), *From Fieldwork to Linguistic Theory. A Tribute to Dan Everett*. Berlin: Language Science Press, 353-414.

Portelance, E.; Jasbi, M. (2024). "The Roles of Neural Networks in Language Acquisition". *Language and Linguistics Compass*, 18(6), e70001.

Pouw, C.; Kloots, M.D.H.; Alishahi, A.; Zuidema, W. (2024). "Perception of Phonological Assimilation by Neural Speech Recognition Models". *Computational Linguistics*, 50(4), 1557-85.

Proietti, M.; Lenci, A. (2025). "The Quasi-Semantic Competence of LLMs: A Case Study on the Part-Whole Relation". *arXiv*, 2504.02395. https://arxiv.org/abs/2504.02395.

Ren, S.; Tomlinson, B.; Black, R.W.; Torrance, A.W. (2024). "Reconciling the Contrasting Narratives on the Environmental Impact of Large Language Models". *Scientific Reports*, 14(1).

Resnik, P. (2025). "Large Language Models Are Biased Because They Are Large Language Models". *Computational Linguistics*, 51(3), 885-906

Rodriguez, J.D.; Mueller, A.; Misra, K. (2025). "Characterizing the Role of Similarity in the Property Inferences of Language Models". *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 11515-33.

Sag, I.A. (2010). "English Filler-Gap Constructions". *Language*, 86(3), 486-545.

Shanahan, M. (2024). "Talking About Large Language Models". *Communications of the ACM*, 67(2), 68-79.

Shojaee, P.; Mirzadeh, I.; Alizadeh, K.; Horton, M.; Bengio, S.; Farajtabar, M. (2025). "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity". *arXiv*, 2506.06941. https://arxiv.org/abs/2506.06941.

Suozzi, A.; Capone, L.; Lebani, G.E.; Lenci, A. (2025). "BAMBI: Developing Baby Language Models for Italian". *Lingue e Linguaggio*, 24(1), 83-102.

Tie, G.; Zhao, Z.; Song, D.; Wei, F.; Zhou, R.; Dai, Y.; Yin, W.; Yang, Z.; Yan, J.; Su, Y.; Dai, Z.; Xie, Y.; Cao, Y.; Sun, L.; Zhou, P.; He, L.; Chen, H.; Zhang, Y.; Wen, Q.; Liu, T.; Gong, N.Z.; Tang, J.; Xiong, C.; Ji, H.; Yu, P.S.; Gao, J. (2025). "Large Language Models Post-Training: Surveying Techniques from Alignment to Reasoning". *arXiv*, 2503.06072. https://arxiv.org/abs/2503.06072.

Tomasello, M. (2003). *Constructing a Language. A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. (2017). "Attention Is All You Need". *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA.

Villalobos, P.; Ho, A.; Sevilla, J.; Besiroglu, T.; Heim, L.; Hobbhahn, M. (2024). "Position: Will We Run out of Data? Limits of LLM Scaling Based on Human-Generated Data". *Proceedings of the 41st International Conference on Machine Learning*. Vienna.

Warstadt, A.; Bowman, S.R. (2022). "What Artificial Neural Networks Can Tell Us About Human Language Acquisition". *arXiv*, 2208.07998. https://arxiv.org/abs/2208.07998.

Warstadt, A.; Mueller, A.; Choshen, L.; Wilcox, E.; Zhuang, C.; Ciro, J.; Mosquera, R.; Paranjabe, B.; Williams, A.; Linzen, T.; Cotterell, R. (2023). "Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora". *Proceedings of the Baby Language Model Challenge at the 27th Conference on Computational Natural Language Learning*, 1-34.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E.H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; Fedus, W. (2022). "Emergent Abilities of Large Language Models". *Transactions on Machine Learning Research*.

Yang, C.; Crain, S.; Berwick, R.C.; Chomsky, N.; Bolhuis, J.J. (2017). "The Growth of Language: Universal Grammar, Experience, and Principles of Computation". *Neuroscience & Biobehavioral Reviews*, 81, 103-19.

Zamparelli, R. (2025). "Doing Linguistics in the GPT Era". *Italian Journal of Linguistics*, 37(1), 177-84.