Report from Dagstuhl Seminar 25301

Linguistics and Language Models: What Can They Learn from Each Other?

Anna Rogers*[1], Nathan Schneider*[2], Bonnie Webber*[3], A. Seza Doğruöz†[4], and Asad Sayeed†[5]

1 **IT University of Copenhagen, DK.** arog@itu.dk
2 **Georgetown University – Washington, DC, US.** nathan.schneider@georgetown.edu
3 **University of Edinburgh, GB.** bonnie.webber@ed.ac.uk
4 **Ghent University, BE.** as.dogruoz@ugent.be
5 **University of Gothenburg, SE.** asad.sayeed@gu.se

Abstract

An international group of 40 scholars in computational linguistics, natural language processing, and cognitive science was assembled to discuss the relationship between linguistics and contemporary language models. Over the course of the week, presentations and work sessions grappled with questions about how LMs can support linguistic research (either as a source of evidence, or as a tool); how linguistic knowledge can inform the design, interpretation, or application of LMs; and what framing is appropriate for the language functionality of LMs.

Seminar July 20–25, 2025 – https://www.dagstuhl.de/25301
2012 ACM Subject Classification Computing methodologies → Cognitive science; Computing methodologies → Natural language processing
Keywords and phrases cognitive modelling, language models, linguistic theory
Digital Object Identifier 10.4230/DagRep.15.7.187

1 Executive Summary

Nathan Schneider (Georgetown University – Washington, DC, US)
Anna Rogers (IT University of Copenhagen, DK)
Bonnie Webber (University of Edinburgh, GB)

Since the release of ChatGPT, language models (LMs) have stirred concerns in government, over the possibility that citizens will come to believe the textual and spoken output of such models. Similarly, they have caused panic in education, forcing a rethink of what students are learning and how to assess it. Of concern to us here, is whether LMs mean the end of computational and/or cognitive models of human language learning and language use. Does the practical success of LMs mean that computational linguistics (and perhaps even linguistics itself) is no longer relevant? Or are we missing problems with LMs that computational linguistics (and linguistics more generally) could help us both recognize and surmount?

* Editor / Organizer
† Editorial Assistant / Collector

Linguistics and Language Models: What Can They Learn from Each Other?, *Dagstuhl Reports*, Vol. 15, Issue 7, pp. 187–212
Editors: Anna Rogers, Nathan Schneider, Bonnie Webber, A. Seza Doğruöz, and Asad Sayeed
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

To have any hope of answering big questions about this technology, we need to foster interdisciplinary conversations and collaborations across the fields of machine learning, NLP, linguistics, and cognitive science. This Dagstuhl Seminar was organized to facilitate such conversations and collaborations among senior experts and rising stars. In particular, the five key questions were raised for discussion:

- What evidence, if any, do LMs provide about human language, world knowledge and/or cognition?
- How can LMs be used as tools for empirical research in linguistics?
- How can linguistics be brought to bear on interpreting the operation of LMs?
- How can linguistically-oriented perspectives enhance or complement LMs for greater reliability and robustness?
- What is the appropriate framing of LM-functionality, for scientists and the public?

An international group of 40 scholars in computational linguistics, natural language processing, and cognitive science was assembled for our week-long seminar. Commensurate with the broad questions raised in the seminar, participants were selected for their wide-ranging expertise on topics such as computational cognitive modeling and psycholinguistics; multilingual modeling and language variation; formal and functional aspects of language use; machine learning; LM interpretability; NLP for low-resource languages; applications and social impacts of language technologies; and philosophical underpinnings of modeling language.

The scientific program consisted of

- Eleven 20-minute talks raising perspectives and questions to inspire further discussion.
- Two rounds of working groups formed dynamically based on participant suggestions. The first set of groups held parallel meetings on Monday/Tuesday, each presenting a synopsis in a plenary session Tuesday evening. The second round of groups took place Wednesday morning and Thursday, reporting back in a Thursday evening plenary session.
- Friday morning was devoted to a plenary discussion of next steps, with about a dozen participants volunteering to organize follow-up initiatives to capitalize on some of the most fruitful conclusions of the working groups.

Abstracts from the talks as well as the working groups are reported below. In true Dagstuhl fashion, the formal scientific program was complemented by opportunities for socialization and recreation in and around the castle – the lively exchange of ideas and perspectives that began in the official sessions continued over meals, coffee breaks, nature hikes, and a sightseeing excursion to Trier.

Finally, a word of thanks from the organizers: We are grateful to all the attendees and the Dagstuhl staff who made the seminar an incredible experience. Special shoutouts go to Christina Schwarz for her administrative leadership; to participants A. Seza Doğruöz and Asad Sayeed, who agreed to serve as collectors for the final report; and to Asad also for his organizational assistance with the Wednesday social outing to Trier.

2 Table of Contents

3 Overview of Talks

3.1 Endangered Languages, Language Varieties, and LLMs

Antonios Anastasopoulos (George Mason University – Fairfax, US)

This talk explores the potential for interaction between documentary or descriptive linguistics on one side, and on NLP and LLM-focused researchers on the other. In this talk I'll argue the importance of documenting the wealth of linguistic diversity, and outline specific tasks amenable to technological intervention in the documentation process. Next, I will outline a proposal and show evidence from preliminary experiments for building language technologies in under-served languages, centered around careful small-scale data curation and on leveraging already-codified linguistic knowledge in the form of descriptive grammars.

3.2 Remarks on the Distributional Foundations of Language Models

Juan Luis Gastaldi (ETH Zürich, CH)

Linguistic distributionalism has been identified as a central theoretical principle explaining or justifying the success of neural LMs. However, there is a significant gap between the theoretical principles associated with distributionalism and the formal mechanisms governing current LMs. I proposed an interpretation of distributionalism based on current developments in category theory and type theory that could help bridge this gap and critically assess our knowledge about LMs. I introduced this topic with more general epistemological remarks concerning the kind of interpretability one can expect to achieve through this formal approach. Concretely, I defended the following claims:

1. LLMs have no a priori cognitive import
2. The empirical study of LLMs has no epistemological grounds
3. Distributionalism is the best theoretical candidate to study LLMs
4. Distributionalism is a corollary of structuralism
5. The general form of distributions is $\mathcal{M}\colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$
6. The general form of structures is $\mathcal{M}^*\colon \mathcal{V}^{\mathsf{C}^{\mathrm{op}}} \rightleftarrows (\mathcal{V}^{\mathsf{D}})^{\mathrm{op}}\colon \mathcal{M}_*$
7. This new structuralist formalism provides new representational tools for explainability and interpretability
8. Language models are culture models

3.3 Count me impressed

Adele Goldberg (Princeton University, US)

Constructions are learned pairings of form and function, at varying levels of complexity and abstraction. They are many, varied, context-dependent and interrelated to one another. The functions of constructions can involve subtle aspects of meaning, attitude, speech acts, and information structure. I will present two quite subtle and distinct aspects of language that are replicated by LLMs.

First, is an historical shift in word order amongst a cluster of conjunctions: English speakers used to say, uncles and aunts, nephews and nieces, and pa and ma, but the preference for these cases reversed to today's female-first order today, over the course of many decades (Goldberg & Lee, 2021). By training a GPT2LMHeadModel from scratch on 111GB of text, with closely related conjunctions filtering out, we demonstrate that the gradual shift is replicated by the model at 3 iterations.

Second, English speakers' judgments about the information structure of canonical sentences predicts independently collected acceptability ratings on corresponding "long distance dependency" (LDD) constructions, across a wide array of base constructions and multiple types of LDDs (Cuneo & Goldberg, 2023). To determine whether any LM captures this relationship, we probe GPT-4 on the same tasks used with humans and new extensions. Results reveal reliable metalinguistic skill on the information structure and acceptability tasks, replicating a striking interaction between the two, despite the 0-shot, explicit nature of the tasks, and little to no chance of contamination. A second study manipulates the information structure of base sentences and confirms a causal relationship: increasing the prominence of a constituent in a context sentence increases the subsequent acceptability ratings on an LDD construction (Cuneo et al., 2025). The findings suggest a remarkable relationship between natural and LM-generated English that make LMs a rich resource for testing theories of language.

3.4 A model for language models

Aurelie Herbelot (Denotation – Pritzwalk, DE)

In spite of their excellent performance on a range of benchmarks, LLMs (and other computational models) have been found to struggle with phenomena that require precise extensional knowledge. This talk highlights a range of quantification phenomena which, arguably, require some kind of representation of individual entities and events to be processed. I will suggest that quantification is a case where linguistic theory can be usefully integrated into computational systems and I will show what such integration can look like when expressing fundamental parts of model theory in vector spaces. I will also briefly demonstrate that the integration has benefits for standard lexical tasks when learning over small corpora. The second part of the talk asks to what extent this kind of representational integration can take place in LLMs, where there is much less control over the spaces that emerge in the course of training. I will present a real-world case study where a LM must be trained over limited data and could benefit from some form of limited exensional knowledge.

3.5 The Changing Roles of (Linguistic) Structure in Computational Linguistics

Mark Johnson (Macquarie University – Sydney, AU)

This talk describes the various roles that linguistic theory and structure have played in computational linguistics, and speculates about the role that they may play in the future. The closest relationship between linguistics and computational linguistics was probably with the Unification Grammars introduced in the 1980s, where the goal was to develop a computational model that implemented the linguistic theory. This close relationship proved impractical for scientific and sociological reasons that I'll describe, and since then the relationship has steadily weakened, as reflected by Jelinek's "when I fire a linguist . . . " quip and Sutton's "Bitter Lesson". I argue that the huge training data and long context windows of Deep Learning models makes it unnecessary to incorporate any specific linguistically-inspired parsing architecture into such models. While there are deep scientific questions about how LLMs "understand" human languages, their linguistic ability is sufficiently good for most practical tasks. Quite reasonably most current research focuses on the information content of the language LLMs generate, such as reducing hallucinations and improving instruction-following. Thus it seems the main opportunities for linguistics to contribute to modern computational linguistics are in model evaluation and explainability.

3.6 What kind of learning is in-context-learning? Evidence from psycholinguistics

Tom McCoy (Yale University, US) and Robert Frank (Yale University, US)

The field of psycholinguistics has developed fine-grained methods for using behavior to understand the mechanisms that underlie human language processing. In this talk, we will argue that such methods can also be used to analyze the language-processing mechanisms employed by large language models (LLMs). We will illustrate this point through a case study about in-context learning, the poorly-understood ability of LLMs to "learn" a task via examples presented to them in context, without explicit parameter updates. One line of research has claimed that in-context learning is functionally equivalent to gradient descent, a type of error-driven learning mechanism, but this claim remains controversial. As a new source of evidence in this debate, we draw a connection between in-context learning and structural priming, the psycholinguistic phenomenon in which people tend to produce sentence structures they have encountered recently. Structural priming literature has argued that human structural priming involves error-driven learning, a conclusion based on evidence from the inverse frequency effect (IFE) – a phenomenon in which an agent's behavior is influenced to a greater degree when presented with improbable examples as compared to more likely ones – which is a signature of error-driven learning. We show that the IFE is also present in LLMs performing in-context learning, which is evidence that in-context learning does indeed behave as an implicit type of error-driven learning. This work provides an example of how psycholinguistic methods can illuminate not only the aspects of grammar that LLMs have or have not captured (a direction that is common in existing work) but also the types of processing mechanisms that drive the language-processing abilities of LLMs.

3.7 Causal abstraction as a toolkit for developing linguistic theories

Christopher Potts (Stanford University, US)

Large language models (LLMs) are incredible new investigative tools for linguistic research; modern linguistics is based in distributional evidence, and LLMs are exceptionally powerful distributional learners. In this talk, I'll illustrate this potential using causal abstraction, which can characterize the abstract representations that LLMs have learned to use. Causal abstraction analyses reveal that LLMs have induced many of the core constructs posited by syntactic theories, while also suggesting new factors that may be relevant to such theories. I'll close with a discussion of the poverty of the stimulus. LLMs reveal that the distributional evidence is richer than previously assumed, which should lead us to reassess arguments that specific phenomena cannot be learned from the available data.

3.8 Linguistics from First Principles and LLMs

Siva Reddy (MILA – Montreal, CA & McGill University – Montreal, CA)

LLMs present an opportunity for linguistics. Instead of verifying whether linguist-annotated representations are present in these models, we can reverse the question: If we were to rely on foundational principles of linguistics, can we extract linguistic representations from the internals of LLMs? In this talk, we will use foundational principles for syntax and semantics to directly extract dependency structures and model-theory-based semantics from LLMs.

In the latter part of the talk, we will focus on how to use LLMs to gain insights into their linguistic representations, directly using prompts. We will contrast different settings: probabilities versus meta-linguistics, and instruction-following versus base models. Finally, using recent reasoning models like DeepSeek-R1, which are trained to discover chains of thought that could lead to the correct answer, we will analyze their reasoning processes on linguistic stimuli.

3.9 What are Large Language Models Models Of?

Philip Resnik (University of Maryland – College Park, US)

Should we be thinking of LLMs as cognitive models for human language processing? This talk went beyond the "argument from amazingness" -- that LLMs are so good at human language that they must be doing something similar to what we do -- to a more careful and critical assessment of what it means to model human language processing, and why it might or might not make sense to view LLMs as models in that sense.

The discussion is organized in terms of Marr's levels of explanation. At the implementation level, we argue that, to whatever extent neo-connectionist models like transformers remain "biologically inspired", that inspiration comes from a version of neurobiology that is at least

a half-century out of date. As such, LLMs are a poor candidate for implementation-level models of human language processing, especially in contrast to the current and actively growing new generation of implementation-level models of processing that actually are in line with current neuroscientific knowledge.

At the algorithmic/representation level, it's worth acknowledging that aspects of deep learning are genuinely a game-changer with respect to theories of human language processing, notably in the success of representation learning as an alternative to manually constructed representations. However, the feed-forward architecture of LLMs is misaligned with the increasing body of evidence that relevant aspects of human perception and cognition are predictive in nature – cf. the Bayesian Brain hypothesis and its realization in predictive coding models. Moreover, there is no basis for the widely held belief that two systems solving the same problems must necessarily have similar solution mechanisms: nature provides numerous counter-examples where different organisms solve the same problems in vastly different ways. And if you look at how model organisms are chosen in medical research, a good model is not one that just manifests a similar behavior or phenomenon, there are also independent reasons for believing what we learn from a model organism will transfer over to humans. For example, mice and rats are used in cancer and cardiovascular disease research respectively, not the reverse, because we know mice have similar oncogene pathways to humans and rats have similar cardiovascular physiology.

At Marr's computational theory level, even LLMs' amazing capabilities in language input/output behavior don't constitute a convincing argument for their treatment as cognitive models. Two systems with identical input/output can still differ drastically in the way they work – e.g., exactly the same input and output pairs are generated by iterative and recursive implementations of the factorial function. We care about computation-level theories in large part because they are a step toward underlying mechanistic understanding. Even if LLMs could achieve the right input/output behavior across a wide range of human language phenomena, there's no reason to believe that there will then be a way to get from the one-size-fits-all, homogenous computation-level framework to anything resembling the evolved solution inside human heads. In that regard, faith in the cognitive relevance of LLMs is ironically very similar to the Chomskian approach in theoretical linguistics, both fixated on achieving an elegant and uniform solution to problems that nature solves messily – nature being a scavenger building on existing capabilities, not a designer refactoring a system to maintain its elegance and uniformity.

Ultimately, then, we find that LLMs don't offer a convincing case as cognitive models, at any level of explanation. What they do offer – in a truly game changing way – is a new kind of support tool for developing actual cognitive models: an ability to provide missing jigsaw puzzle pieces in models that require distributed representations, proxies for world knowledge, stand-ins for plausible inference, and much more. One can lean into the amazing capabilities of LLMs for developing cognitive models without having to buy into the idea that they constitute models in and of themselves.

3.10 Combining Large Language Models and Symbolic Systems for Logical Inference

Mark Steedman (University of Edinburgh, GB)

The traditional view of Natural Language Processing (NLP) distinguishes two components: The Grammar, which is recursive and logic-related, and supports syntactic analysis and semantic interpretation; and The Model, which is probabilistic, and assigns measures of similarity of association or context to symbols and sequences.

This distinction is parallel to a widespread view of psychologists that human understanding proceeds via "Dual Processes", involving both systems of "Type 2", consisting of high-precision but slow symbolic systems, and "Type 1" systems, consisting of fast, sloppy, imprecise but high-recall systems such as Finite-State Transducers (FST) or Language Models (LM). Type 1 systems are usually "good enough", but can be monitored or augmented by Type 2 systems via channels of very limited bandwidth.

The primary purpose of the Grammar/Type 2 system of NLP is to support sound logical inference, while the purpose of the Model/Type 1 system is to limit ambiguity in mapping strings to meanings and vice versa, via the context of utterance.

Large Language Models (LLM) consisting of up to a trillion parameters and trained on trillions of words of human generated text have recently proved extremely useful in a number of NLP tasks such as question answering and summarization. However, such tasks involve inference from the form of statements in documents to possible statements in the output. The talk will review capabilities of LLMs for Natural Language Inference (NLI) in such applications.

We can think of LLMs as a hypersphere of vector embeddings with only a few hundred dimensions of associative similarity, algorithmically compressed by dimension reduction from the vaster dimensionality and empty sparsity of the raw association space obtained from text.

Crucially, there is known to be a gradient of generality on each axis from very general terms at the center of this hypersphere to more specific terms at the periphery. It will be important to what follows to recall that terms that stand in the relation of a generalization ore known to also form a partial ordering of frequency in text.

The talk will argue that, once certain biases in the human-constructed NLI datasets and in the LLM themselves are controlled for, LLMs perform quite badly on pure NLI tasks, despite fair recall, showing poor precision or false-positive conclusions ("hallucinations"). In particular, LLM are prone to two biases that are inherent in the text training data, namely an Attestation Bias stemming from the fact that those data concern facts, and a Frequency Bias, stemming from the fact that some things are more written about than others.

The paper will argue instead for the unsupervised extraction of "Entailment Graphs" (EG) or probabilistic NLI networks via detection using machine reading of probabilistic typed entailment relations such as that buying events imply ownership states events. EG are higher-precision, and our methods scale. However, Zipf's Law means that they are subject to an intrinsic curse of sparsity.

The paper presents results from the Edinburgh group developing hybrid systems that combine the high precision of EG with the high recall of LLM by exploiting the biases themselves of the latter. We show that the Frequency gradients in the latter can be leveraged to "smooth" entailment premises that are missing from EG via nearest LLM neighbors that

are present in the EG, with the entailment graph doing the rest of the work. We also exploit the Attestation Bias of LLM to find entailments de novo by eliciting attested analogs of the Premise, then querying the attestation of parallel Hypotheses.

Our conclusion is that the future of NLI and related applications lies with such cautious hybridized combination of the Precision of symbolic Type 2 inference systems with the Recall of LLM Type 1 systems.

3.11 Can (and should) an AI do science?

Adina Williams (Meta Platforms – New York, US)

LLMs are now good enough to generate text that can plausibly pass as a scientific paper. Papers have been used as evidence of scientific advancement, both for hiring/recruiting and for progressing in our field's scientific goals. The incentives of our field make it likely that people will use them for generating papers. Since an LLM is not a part of our community and we cannot trust that the texts it outputs is good evidence of science having been done, LLMs may contribute to undermining our scientific community especially our peer review system which we all rely on to scale scientific trust beyond tight friend and or collaborator networks.

4 Working groups

4.1 Multilingualism and low-resource languages

A. Seza Doğruöz (Ghent University, BE), Verena Blaschke (Ludwig-Maximilians-Universität München, DE), David Adelani (MILA – Montreal, CA), Lori Levin (Carnegie Mellon University – Pittsburgh, US), Xixian Liao (Barcelona Supercomputing Center, ES), Joakim Nivre (Uppsala University, SE), Alexis M. Palmer (University of Colorado – Boulder, US), Gözde Gül Şahin (Koç University – Istanbul, TR), and Amir Zeldes (Georgetown University – Washington, DC, US)

Although there are ∼7,000 languages in the world, most research in computational linguistics and natural language processing (NLP) focuses on English [11, 5, 19]. To study multi- and cross-lingual aspects of language models, multilingual datasets and/or datasets in low-resource languages are key. Nevertheless, many datasets cover English and multilingual datasets often contain content specific to some cultures more than others. Many multilingual datasets are based on translations, requiring caution with respect to translationese [17]. Some datasets might also disregard linguistic variation, especially within non-standardized language varieties. A lack of standardization of formats, tools, and APIs makes it harder to compare datasets (or to analyze models across datasets). Additionally, quality control and reachability of a responsible human maintainer are important for the use of datasets by others.

It is also difficult to evaluate multilingual models meaningfully. Massively multilingual benchmarks can be biased (as mentioned above). Intrinsic, task-independent evaluation that is comparable across languages remains a challenge, and intrinsic evaluation results are not necessarily correlated with real-world performance. Automatic metrics can be biased against certain language types and they can be brittle towards variation [20, 1]. If they are learned from data they may simply not cover low-resource languages. Possible steps forward include taking inspiration from the Multidimensional Quality Metrics framework [13], categorizing evaluation metrics for different languages, monitoring existing evaluation sets for content and language issues, and prioritizing speaker community involvement in the creation of new datasets.

Insights from linguistic typology could also be used for interpreting multilingual language models. Methods from interpretability research (e.g., probing, training sparse auto-encoders, or designing causal interventions; see [3], and [16]) could be used to investigate whether the internal representations of multilingual models organize linguistic information in a way that it mirrors typological patterns or shows systematic differences between language-specific vs. cross-lingual representations (e.g., [6], and [21]).

Although many assumptions in NLP research may not necessarily hold for low-resource languages, and standard parameter and/or tokenization settings might be suboptimal for many of them, these challenges may still be disregarded for processing and studying these languages. Language models that are of sufficient quality for some downstream applications might not meet the quality standards needed for archival-quality documentation that can support linguistics research and language revitalization [10]. Language use can also be different than frequently assumed. For example, some speaker communities are multilingual and utilize more than more language/dialect on a daily basis [9]. Furthermore, textual data is also prioritized over informal, spoken data [7] although there are many languages that are only spoken (not written) and/or they may not have standardized orthographies.

Lastly, it is important to support and involve speaker communities and field linguists when creating NLP tools for low-resource languages [8]. Community members could be consulted before building NLP products for their languages to ensure that their needs and preferences are met [12, 14, 4]. Furthermore, there are many opportunities (especially in low resource languages) for building NLP tools that can help field linguists (e.g., for transcribing speech data, or OCR tools for extracting text from images) if they are built with their needs in mind [15, 10]. Additionally, NLP tools could be used to aid with analyzing language data. A task like inferring linguistic structures from data (meta-linguistic reasoning [18]) might be of interest to both linguists and computer scientists.

References

1 Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore, December 2023. Association for Computational Linguistics.

2 Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online, November 2020. Association for Computational Linguistics.

3 Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

4 Steven Bird and Dean Yibarbuk. Centering the speech community. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

5 Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland, May 2022. Association for Computational Linguistics.

6 Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

7 Grzegorz Chrupała. Putting natural in natural language processing. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7820–7827, Toronto, Canada, July 2023. Association for Computational Linguistics.

8 A. Seza Doğruöz and Sunayana Sitaram. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In Maite Melero, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France, June 2022. European Language Resources Association.

9 A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. A survey of code-switching: Linguistic and social perspectives for language technologies. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online, August 2021. Association for Computational Linguistics.

10 Luke Gessler, Alexis Palmer, and Katharina Von Der Wense. Understanding the gap: an analysis of research collaborations in NLP and language documentation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 867–877, Vienna, Austria, July 2025. Association for Computational Linguistics.

11 Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.

12 Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. Not always about you: Prioritizing community needs when developing endangered language technology. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland, May 2022. Association for Computational Linguistics.

13 Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib.

14 Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada, July 2023. Association for Computational Linguistics.

15 Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaum. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12, 2018.

16 Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for nlp through the lens of causal mediation analysis. *Computational Linguistics*, pages 1–48, 09 2025.

17 Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Oh. Culture is everywhere: A call for intentionally cultural evaluation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19156–19168, Suzhou, China, November 2025. Association for Computational Linguistics.

18 Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. PuzzLing Machines: A Challenge on Learning From Small Data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online, July 2020. Association for Computational Linguistics.

19 Anders Søgaard. Should we ban English NLP for a year? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

20 Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. Dialect-robust evaluation of generated text. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada, July 2023. Association for Computational Linguistics.

21 Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

4.2 Accommodation and Social Aspects of LLMs

A. Seza Doğruöz (Ghent University, BE), David Adelani (MILA – Montreal, CA), Antonios Anastasopoulos (George Mason University – Fairfax, US), Verena Blaschke (Ludwig-Maximilians-Universität München, DE), Aurelie Herbelot (Denotation – Pritzwalk, DE), Yu-Yin Hsu (Hong Kong Polytechnic Univ., CN), Xixian Liao (Barcelona Supercomputing Center, ES), Siva Reddy (MILA – Montreal, CA & McGill University – Montreal, CA), Philip Resnik (University of Maryland – College Park, US), Anna Rogers (IT University of Copenhagen, DK), Gözde Gül Şahin (Koç University – Istanbul, TR), Asad Sayeed (University of Gothenburg, SE), Noah A. Smith (University of Washington – Seattle, US), and Bonnie Webber (University of Edinburgh, GB)

LLMs are widely used across domains but there are relatively less insights into the variation and change within languages in terms of context and speakers/users involved in communication. Socio-demographic characteristics of the speakers and context in human-human communication influence the language use ([4]). Humans share common ground and accommodate each other to achieve common communication goals ([5, 2]). Since there is not always a common ground between LLMs and a human ([3]), it is difficult to measure and evaluate accommodation.

However, not all humans approve of LLM's accommodation in language use (e.g., attitudes of dialect speakers toward LLMs by [1, 6]). It is also the case that for certain domains (e.g., mental health), it is favorable if the LLMs do not always accommodate the users.

Possible applications of this research direction include personalization of LLMs for specific users. However, there is a need for more research to understand the goals and social aspects of human-human interaction across contexts and among users with varying socio-demographic characteristics.

References

1 Blaschke, V., Purschke, C., Schuetze, H., Plank,B. (2024). What Do Dialect Speakers Want? A Survey of Attitudes Towards Language Technology for German Dialects. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 823-841, Bangkok, Thailand. Association for Computational Linguistics.

2 Herbert H Clark. 1996. Using language. Cambridge university press.

3 Doğruöz, A.S. & Skantze, G. (2021). How "open" are the conversations with open-domain chatbots? A proposal for Speech Event based evaluation. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 392–402, Singapore and Online. Association for Computational Linguistics.

4 Nguyen, D., Doğruöz, A.S., Rosé, C.P., de Jong, F. (2016). Computational Sociolinguistics: A Survey. Computational Linguistics, 42(3):537–593.

5 Pickering MJ, Garrod S. (2021) Understanding Dialogue: Language Use and Social Interaction. Cambridge University Press.

6 Sandoval, S.C, Acquaye, C., Cobbina, K.A., Teli, M.N., and Daumé, H. (2025). My LLM might Mimic AAE – But When Should It?. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5277–5302, Albuquerque, New Mexico. Association for Computational Linguistics.

4.3 Goals of Linguistic Theory

Robert Frank (Yale University, US), Katherine Demuth (Macquarie University - Sydney, AU), Juan Luis Gastaldi (ETH Zürich, CH), Mark Johnson (Macquarie University - Sydney, AU), Najoung Kim (Boston University, US), Roger Levy (MIT - Cambridge, US), Siva Reddy (MILA - Montreal, CA & McGill University - Montreal, CA), Philip Resnik (University of Maryland - College Park, US), Anna Rogers (IT University of Copenhagen, DK), Rachel Rudinger (University of Maryland - College Park, US), Nathan Schneider (Georgetown University - Washington, DC, US), Mark Steedman (University of Edinburgh, GB), Tiago Torrent (Federal University of Juiz de Fora, BR), and Adina Williams (Meta Platforms - New York, US)

This group focused on questions at the foundation of the science of language. Our first focus for discussion sought to identify the goals of linguistic theory and its objects of study. There was broad agreement that the focus of the field should be on "explaining patterns of language," but there was broad ranging discussion about what constitutes such a pattern. On the one hand, participants agreed that patterns of sound (phonology), word structure (morphology) and sentence structure (syntax) are central, as are the ways in which meaning is encoded in the words of language (lexical semantics) and the ways such meanings can be put together to produce interpretations of larger units (compositional and discourse semantics). Patterns of language use (pragmatics) and the fit of linguistic form to social and political context were also mentioned as important areas of study, but there was some belief that social and political factors governing language use may lie outside the core of language behavior and would be better understood through the interaction of the language faculty with other cognitive capacities.

This latter point led to a discussion about the ways in which the field has partitioned linguistic and non-linguistic cognition. We reviewed Fodor's ([4]) notion of modularity, according to which language knowledge, as a cognitive module, should be be cleanly separated from non-modular systems of thought involved in the use of world knowledge. LLMs don't appear to make such a division between linguistic and non-linguistic cognition, and this led to a discussion about whether this might be taken as a reason to reconsider the modularity assumption. One empirical reason to move in this direction might come from the line of work in the psycholinguistics in the 1980s and 90s (e.g., [7]) which showed that it was difficult to distinguish the processing impact of linguistic and non-linguistic information in sentence processing (though there are ways of reconciling these results with modularity – cf., [1]). Recent work on LLMs as models of sentence processing (e.g., [8]) seems to point to the conclusion that LLMs transcend human performance, in that they do not show sensitivity to the same limitations that humans do. Consequently, the lack of modularity in LLMs might not be reason to doubt its presence in human processing.

Our discussion then moved to the means by which linguistic theories should be judged. General scientific desiderata were discussed, including falsifiability and testability of predictions and insight and explanatory value. Other factors that came up were more specific to linguistics: the explanation of the capacity for systematic generalization (e.g., to novel forms) and the ability to account for typological patterns (including entailments from one type of grammatical property to another and the non-existence of certain types of grammars). Finally, there was broad agreement that theories of language should engage with patterns of behavior in comprehension and production, as well as with the characterization of developmental trajectories during language acquisition.

With these foundations in mind, we returned to the question of assessing the viability of LLMs as theories of human language. Group members agreed that LLMs are clearly impressive in their ability to carry out certain tasks. However, there was broad agreement that it is difficult indeed to know how to evaluate their success scientifically. A number of open questions were raised about what could in principle be learned from LLMs: what is the nature of their inductive bias that contributes to their success in language learning? what can LLMs tell us about the relevant data structures for language? Much of the work on LLM interpretability in the linguistic vein has focused on finding analogs of known linguistic structures in the learned representations of LLMs (e.g., [6]). An especially interesting example of this arose in the talk at the workshop by Chris Potts ([2]) pointing to the convergence in representation across the wh-movement constructions (C[3]). If this is the kind of finding we repeatedly find, it points to the validity of the abstractions of current linguistic theory, but also says that the model structure, training data and gradient descent-driven learning process are sufficient to yield relevant abstractions without innate guidance.

Our final day of discussion focused on the problem of evaluation metrics for LLMs as models of language structure. The problem stems from the fact that traditional ways of thinking about grammar learning assume discrete languages, yet LLMs are probabilistic models, which do not simply rule strings in and out of a language, bur rather provide a distribution over such strings. To judge the fit between and LLM and a linguistic pattern, a variety of approaches have been explored. Most simply, one can compare the probability assigned to sentences in a minimal pair, with the expectation that the grammatical sentence should receive higher probability than the ungrammatical one. A number of problems were discussed for this approach: how do we determine what constitutes a minimal pair? How do we deal with other facts that can impact assigned probability (e.g., frequency effects, semantic plausibility, etc.), but which are plausibly orthogonal to language structure itself? More complex approaches attempt to provide a fixed score for sentence, by normalizing the assigned probability for sentence length and word frequency ([5]). We began to explore a third possibility based on the idea that learning requires demonstrating use of relevant latent variables or structures. We discussed a number of ways that such latent structures could be identified. One direct way would be via interpretability methods, but the participants expressed varying degrees of skepticism as to whether we can have confidence in current methods. An alternative approach began to emerge by examining the impact the relevant abstractions would have on the distributions generated by a model containing them.

References

1 Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. Cognition, 30(3), 191–238.

2 Boguraev, S., C. Potts & K. Mahowald. (2025). Causal interventions reveal shared structure across English filler-gap constructions. Proceedings of EMNLP.

3 Chomsky, N. (1977). On Wh-Movement. In P. Cullicover, T. Wasow, & A. Akmajian (Eds.), Formal Syntax (pp. 71-132). New York Academic Press.

4 Fodor, J. A. (1983). Modularity of Mind: An Essay on Faculty Psychology. Cambridge, Massachusetts: MIT Press.

5 Lau, J.H., Clark, A. and Lappin, S. (2017), Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. Cogn Sci, 41: 1202-1241.

6 Manning, C.D., K. Clark, J. Hewitt, U. Khandelwal, & O. Levy. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision, Proc. Natl. Acad. Sci. U.S.A. 117 (48) 30046-30054.

7 Tanenhaus, M. K., Dell, G. S., & Carlson, G. (1987). Context effects in lexical processing: A connectionist approach to modularity. In J. L. Garfield (Ed.), Modularity in knowledge representation and natural-language understanding (pp. 83–108). The MIT Press.
8 Van Schijndel, M. & T. Linzen (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. Cognitive Science, 45(6), e12988.

4.4 Tools for exploring linguistic theories

Lori Levin (Carnegie Mellon University – Pittsburgh, US), Adele Goldberg (Princeton University, US), Yu-Yin Hsu (Hong Kong Polytechnic Univ., CN), Najoung Kim (Boston University, US), Tom McCoy (Yale University, US), Joakim Nivre (Uppsala University, SE), Alexis M. Palmer (University of Colorado – Boulder, US), Jakob Prange (Universität Augsburg, DE), Rachel Rudinger (University of Maryland – College Park, US), Nathan Schneider (Georgetown University – Washington, DC, US), and Tiago Torrent (Federal University of Juiz de Fora, BR)

The research question

This group investigated what kind of tools could allow a linguist to visualize the internal representations of Large Language Models (LLMs) so that the linguists could see whether significant linguistic phenomena show up in the internal representations of LLMs and study their representation in LLMs.

(Chris Potts and Mark Johnson mentioned some such tools in their plenary talks.)

The linguistic phenomenon

The group considered three linguistic phenomena:

1. The distinction between arguments and adjuncts
2. The distinction between new and presupposed components of a sentence
3. The distinction between metaphorical and non-metaphorical components of a sentence

The group decided to focus on the distinction between arguments and adjuncts. In the sentence *Alex ate the cake hungrily in Dagstuhl at 3:30pm*, *Alex* and *cake* are arguments of the verb *eat*, whereas *Dagstuhl* and *3:30pm* are adjuncts.

The argument-adjunct distinction plays a central role in many different linguistic theories. In Cognitive Grammar and Frame Semantics arguments may be conceived of as core participants in a scene, which can be brought to mind by either the valency properties of a predicator or by a (non-lexical) construction. In X-bar Theory arguments are closer to the head in a non-recursive projection; adjuncts are farther from the head in a recursive projection. In Case and Theta theory a head can assign case and semantic role to an argument. In Tree Adjoining Grammar arguments are attached to projections of heads via a substitution operation in contrast to adjuncts, which are attached by an adjunction operation. In Categorial Grammar heads are functions that apply to arguments, while adjuncts are functions that apply to the phrases they attach to. We would like to know whether arguments and adjuncts feature as significantly in the internal representations of LLMs as they do in linguistic theory.

A plan

Step 1: Create a dataset of behavioral diagnostics. There are many behaviors that distinguish arguments from adjuncts, including the following:
Omissability (adjuncts are more omissible)
I gave books to students on Thursday.
I gave books to students. (better)
I gave books on Thursday. (worse)
Repeatability (adjuncts are repeatable)
I gave books to people to students on Thursday. (worse)
I gave books to students on Thursday at 3:00. (better)
Selectional Restrictions (predicates impose selectional restrictions on their arguments)
I ate cake (normal)
I ate ideas (metaphorical)
Ideas eat cake (metaphorical)
Order in noun phrases (arguments before adjuncts)
A student of linguistics with long hair (Radford textbook) (better)
A student with long hair of linguistics (worse)
Order in sentences arguments before adjuncts
I gave books to students on Thursday. (better)
I gave books on Thursday to students. (worse)

Step 2: Test whether the LLM replicates human behavior, for clear-cut and non-clear-cut cases. Replicating human behavior might be reflected in perplexity, surprisal, or some other metric in an LLM. We would need to see whether the sentences that are judged as better and worse by humans in the examples above are also measurable as better and worse by some metrics in an LLM.

The argument-adjunct distinction, in spite of its central role in linguistic theory, is notoriously fuzzy around the edges. There are many diagnostics that give ambiguous results. For example, optionality is a defining characteristic of adjuncts, but some adjunct-like things are required in some constructions and many arguments are optional. In addition, humans cannot always make clear judgements about the acceptability of sentences. It is important to know whether LLMs behave like humans in these non-clear cut cases.

What if LLMs do not replicate human behavior? We will experiment with modifications of LLMs to see what kinds of adjustments affect their behavior with respect to the argument-adjunct distinction.

Step 3: Interpretability. We will examine components of LLMs such as circuits and subspaces of hidden states to see which components are implicated in their behavior toward arguments and adjuncts. The major research questions will be where to look, how to avoid confounds, and whether we will need to develop new interpretation techniques if we see nothing at first.

Step 4: See what those internal components do in non-clear-cut cases.

Step 5: Bring the lessons we learn from LLMs' representation of arguments and adjuncts back to linguistic theory.

Inspiring new human experiments: If our methodology for studying interpretability of the argument-adjunct distinction is unsupervised, it could lead us to discover new features or mechanisms that linguists were previously unaware of. We could then develop experiments that would verify whether similar things can be found in humans

Gaining new insights into characterizing the argument/adjunct distinction: If LLMs capture human behavior we might conclude that what they are doing provides one possible

account of how the behavior can be captured. That is, we might use LLMs to discover mechanisms for representing the argument-adjunct distinction that were not found by traditional methods such as corpus studies and human psycholinguistic experiments.

We would especially want to know whether the non-clear-cut cases exhibit a blend of the argument and adjunct signatures, or combinations of discrete sub-pieces of these, or something totally different? If it is a blend, how should this be reflected in updated linguistic theories?

4.5 The (Im)possible Languages Group

Christopher Potts (Stanford University, US), Marie-Catherine de Marneffe (UC Louvain-la-Neuve, BE), Katherine Demuth (Macquarie University – Sydney, AU), Robert Frank (Yale University, US), Juan Luis Gastaldi (ETH Zürich, CH), Hagen Blix, Coleman Haley (University of Edinburgh, GB), Mark Johnson (Macquarie University – Sydney, AU), Roger Levy (MIT – Cambridge, US), Kyle Mahowald (University of Texas – Austin, US), Mark Steedman (University of Edinburgh, GB), and Adina Williams (Meta Platforms – New York, US)

The (Im)possible Languages Group sought to understand the nature of the distinction between possible and impossible human languages, and to explore ways in which Large Language Models (LLMs) might help us explore the issue. Like many past researchers in this area, we were partly inspired by Jorge Luis Borges' short story "The library of Babel" (see [3]), but we infused this with (perhaps less erudite) references to the Mission Impossible movie franchise ([2]) and the Cole Porter musical Anything Goes ([6]).

We centered our discussion around the following claim: There are limits on the set of languages humans can acquire and use as full-fledged natural languages.

What are the nature of these limits? We identified five classes: (1) formal learnability and complexity, (2) expressivity, (3) efficiency, (4) historical/social, and (5) cognitive. We ventured that everyone accepts that (1)-(4) contain robust constraints. The central question is whether there is anything left for the cognitive class (5) that has, often implicitly, been the focus of debates in the field. (See also [4], chapter 3.)

We also agreed that talking about a "set of languages" in this context is misleading. Most or all of the limiting factors (1)-(5) will be a matter of degree, and acquisition itself might be partial and fluid. Thus, the relevant notions are actually quite fuzzy. Nonetheless, there will be clear enough cases (even if there is a large gray area), and so we feel we can talk informally about sets of languages.

What does the phrase "acquire and use as full-fledged natural languages" mean? Here, we did not achieve complete consensus. We agreed that, for a language to count as possible, it must be learnable and usable by ordinary humans without auxiliary tools, and it must be expressive enough to serve as a medium of human thought and communication. Other potential criteria include whether a language could have a lasting and stable community of fluent speakers, or whether it is processed in the brain with the same neural correlates we associate with other natural languages.

The criterion of neural correlates led us to pose a thought experiment. Identical twins Avery and Blake are separated at birth. Avery learns a regular old natural language, while Blake grows up in a community that speaks a typologically unusual system with counting-based rules that linguists would consider "impossible". Now suppose we scan their brains during language use and find that Avery's processing patterns are the expected ones for language whereas Blake's are ones we associate with different cognitive processes. Do you conclude that Blake's language isn't a natural language, or do you conclude that the neuroscience was wrong? Members of our group gave different answers.

We then turned our attention to the role that LLMs might play in these investigations. Using the above framework, we determined that the Shuffle language experiments of [2] are focused on contraints that are not in the cognitive group (5) above, but rather can be attributed to issues of formal complexity or expressivity. However, their Hop languages seem clearly to be oriented around a potential cognitive constraint.

[1] is an insightful critical review of [2]. We focused on Hunter's concern that Kallini et al.'s counting-based rules should be compared against something of similar complexity that crucially makes reference to constituency. To this end, Hunter sketches a "Sister Hop" language in which an agreement marker is placed after the constituent that is the sister of the verb. This is a realistic pattern that, he argues, is a fairer comparison to the Word Hop rule of Kallini et al., which places the agreement marker 4 words after the verb regardless of consituency. The group agreed that Hunter's language is hard to operationalize because of a lack consensus on constituency and/or a lack of parsers that could perfectly implement the correct theory at scale. However, we hit upon an alternative: use a language where Sister Hop is the actual pattern, and then derive the unnatural language from pattern. English verb–particle cases provide one such pattern; in phrases like "Pick the book on the table up", the particle appears immediately to the right of the sister of the verb. The (by hypothesis) impossible variant would place the particle N words from the associated verb, with N likely set to 4. This seems like a very promising basis for an informative experiment.

To close, we asked how we explain the result, now reproduced in a few papers ([5], [6]; but see [7]), that LLMs do distinguish at least some possible languages from impossible variants. One idea is that the AI community has evolved towards architectures that favor possible languages. This evolution could even extend to hyperparameter choices people have inherited over time from prior work and now use without much reflection. The precise factors remain to be discovered. Another idea is that complexity notions might organize the languages considered thus far experimentally. This would mean that we haven't yet evaluated any of the "cognitive" constraints that are the most meaningful to the debate.

References

1 Hunter, Tim. 2025. Kallini et al. (2024) do not compare impossible languages with constituency-based ones. https://direct.mit.edu/coli/article/doi/10.1162/coli_a_00554/128121

2 Kallini, Julie; Isabel Papadimitriou; Richard Futrell; Kyle Mahowald; and Christopher Potts. 2024. Mission: Impossible Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 14691-14714. Bangkok, Thailand: Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.787/

3 Moro, Andrea. 2015. The Boundaries of Babel: The Brain and the Enigma of Impossible Languages, 2d edition. MIT Press.

4 Nefdt, Ryan M. 2024. The Philosophy of Theoretical Linguistics. Cambridge University Press.

5 Xu, Tianyang; Tatsuki Kuribayashi; Yohei Oseki; Ryan Cotterell; and Alex Warstadt. 2025. Can Language Models Learn Typologically Implausible Languages? https://www.arxiv.org/abs/2502.12317
6 Yang, Xiulin; Tatsuya Aoyama; Yuekun Yao; and Ethan Wilcox. 2025. Anything Goes? A Crosslinguistic Study of (Im)possible Language Learning in LMs. https://arxiv.org/abs/2502.18795
7 Ziv, Imry Ziv; Nur Lan; Emmanuel Chemla; and Roni Katzir. 2025. Large Language Models as Proxies for Theories of Human Linguistic Cognition. https://arxiv.org/abs/2502.07687

4.6 "Post-cephalopod" natural language understanding

Asad Sayeed (University of Gothenburg, SE), Ryan Cotterell (ETH Zürich, CH), Marie-Catherine de Marneffe (UC Louvain-la-Neuve, BE), Aurelie Herbelot (Denotation – Pritzwalk, DE), Najoung Kim (Boston University, US), Christopher Potts (Stanford University, US), Siva Reddy (MILA – Montreal, CA & McGill University – Montreal, CA), Rachel Rudinger (University of Maryland – College Park, US), Bonnie Webber (University of Edinburgh, GB), and Ethan Wilcox (Georgetown University – Washington, DC, US)

After the publication of Emily Bender and Alexander Koller's ACL award-winning 2020 article "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data", we decided it was time to discuss whether the debate it instigated is still current in the face of recent developments in LLM technology, or whether there are cogent objections to their layout of debate. This is relevant, because the "Octopus Paper" still to a large extent sets the tone for claims of machine understanding in the face of chatbots that produce output that is usually understandable in conversational context and that humans widely consider plausible that a human could have written. That is, it rejects the idea that meaning can be acquired simply through the manipulation of form (textual representations) alone, which is a key characteristic of the text-only chatbots that were prevalent when the paper was published.

Discussion proceeded through an examination of a common claim: that the Octopus Paper was written in a manner that resembled classical arguments against the possibility of considering machines to be human-like intelligences, most particularly the famous Searle "Chinese Room" Gedankenexperiment. The group concluded that the Octopus Paper makes assumptions about its core allegorical scenario, that of an octopus clandestinely intervening in the conversations of two island-dwellers communicating by wire, that are actually incompatible with the Searlean scenario. Bender and Koller, most notably, do not believe that symbol streams are sufficient for behavioural perfection (unlike Searle), but they do allow that symbol manipulation could actually represent understanding under the right conditions (also unlike Searle).

The group noted that Bender and Koller would allow that the machine would have "understood" if there were evidence of grounding, and that that grounding could be as thin as multimodal content other than text (e.g., images). What has happened since the Octopus Paper is a vast expansion of multimodal models that do just that. However, they often accomplish this also using a Transformer architecture that also depends crucially on conversion of multimodal content into abstract symbols. If symbol-manipulation (in the sense of text) is not in itself meaning, then how is "multimodal" symbol-manipulation a more authentic way to characterize meaning?

The group discussed a potential resolution to this problem that consisted of the following:

- Taking an internalistic view of meaning representation: understanding as the correspondence between language and internal representations.
- Accepting that "understanding" is a continuum, with some entities understanding little, and some understanding much more.

In this way, we can say that a simple household cleaning robot could "understand" in some sense, but it is a very simple form of understanding.

If this concept of understanding is unsatisfactory to some (due to questions of qualia, grounding, embodiment, and so on), the problem is that there is essentially an infinite regression of potential objections, with a new objection surfacing every time an old one is satisfied. E.g.: How many aspects of human experience need to be represented? Is simulating biology enough? And so on.

The group resolved to write an article assessing how to incorporate the current state of technology into the on-going debate.

4.7 LLMs and memory

Amir Zeldes (Georgetown University – Washington, DC, US), Ryan Cotterell (ETH Zürich, CH), Yu-Yin Hsu (Hong Kong Polytechnic Univ., CN), Mark Johnson (Macquarie University – Sydney, AU), Siva Reddy (MILA – Montreal, CA & McGill University – Montreal, CA), Philip Resnik (University of Maryland – College Park, US), Anna Rogers (IT University of Copenhagen, DK), Gözde Gül Şahin (Koç University – Istanbul, TR), and Ethan Wilcox (Georgetown University – Washington, DC, US)

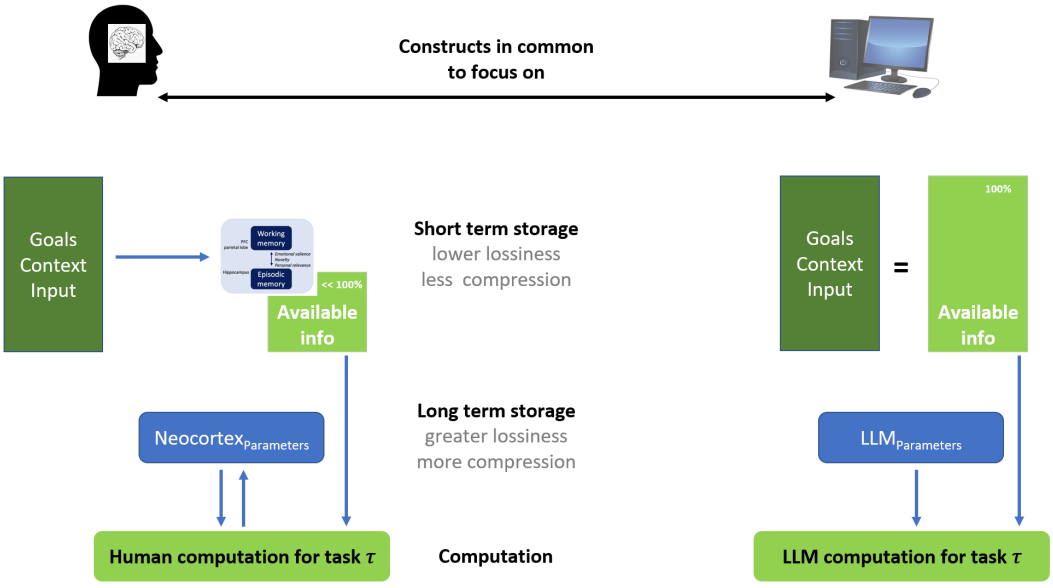
Our group discussed differences and similarities between human linguistic memory and possible LLM analogues. We initially set out to explore a number of themes, including

- Different kinds of memory (working/procedural/semantic/meta-memory)
- Tasks providing insights on memory: summarization, long-form QA, coreference resolution etc.
- Lossiness and information compressed in memory for humans vs. LLMs
- The nature of reasoning and inference, the role of analogy and scratchpads in Chain of Thought generation
- Structured units in representations such as entities/events/common ground and their prioritization/salience
- Implications for Cognitive Science and linguistic theories such of information and discourse structure
- Memory in dialog, including user and conversation memory, dialog states and context representations

We identified a number of problems in the analogy between humans and LLMs, including issues with current RAG-based approaches to long-term memory, which is not well-updated (RAG may retrieve unordered facts, some of which are no longer true or internally contradictory). Some analogies between LLM and human memory concepts are illustrated in the following table:

Brain	Concept	LLM
Experience -> Short term storage	Short Term Memory (STM) - Low compression	Context
Salience bias (emotional, semantic, personal…)	Prioritization at perception time	?
Sleep		(retraining)
Long Term Memory	Long Term Memory (LTM) - High compression	Parameters
Can't know tasks in advance due to linearity of time; must derive all-purpose representations	Task	Can decide at inference time what to attend to based on lossless storage (not lossless *access*, pace Ryan)

After some discussion about how and why humans remember some things (such as how to get back to a weapon store in Trier to catch a shuttle back to Dagstuhl) but not others (what was in the store's window), the group eventually focused on the imperfect human recall of prior linguistic interactions versus LLM access to a potentially perfect transcript in long-context prompts (chat history). Discussion clarified that LLM access to context as memory was mediated in practice by attention bottlenecks at generation time for the model, but this is still not equivalent to human memory, which must prioritize content at perception time, and not at recall time. This disparity is illustrated in the following chart:



The group concluded that some of the most pressing questions are how more human representation types could be incorporated into language technology, though we agreed that evaluation would be a major challenge in establishing both whether certain representations

are in fact more human-like, and whether/how this might be helpful in practical terms. A key direction that arose from discussions in terms of approaching this challenge was the need for constraints on LLM access to prior context, and attention to the characteristics of a system as a coherent agent, for example possessing underlying assumptions and knowledge about oneself. Whereas humans have a notion of "who they are" and therefore of what content they attend to more or less, LLMs are more "generic" and internally inconsistent by nature.

The group continued to communicate after the seminar, including at informal meetings during the following ACL conference in Vienna, and we have set up a mailing list (`lingmem@googlegroups.com`) to keep in touch about the topic. Some members have proposed to collaborate on a potential paper with experiments targeting memory comparisons between humans and LLMs. One proposal was to benchmark humans on a closed and open book summarization task (with and without the text remaining available for consultation after being read once), and comparing LLM behavior with full or partial/attenuated access to the preceding context to simulate a type of compression of information during a left-to-right pass on the text in the closed book scenario. It was suggested that this could be achieved by using existing spoken and written English data with salience scores for sub-spans of input text representing mentioned entities (Ling & Zeldes 2025). A second proposal suggested similar experiments applied to emotional salience, personal relevance or novelty, using data from human dialog annotated for common-ground problems (Sarkar et al. 2025).

Participants

- David Adelani
MILA – Montreal, CA
- Antonios Anastasopoulos
George Mason University – Fairfax, US
- Gašper Beguš
University of California – Berkeley, US
- Verena Blaschke
Ludwig-Maximilians-Universität München, DE
- Ryan Cotterell
ETH Zürich, CH
- Marie-Catherine de Marneffe
UC Louvain-la-Neuve, BE
- Katherine Demuth
Macquarie University – Sydney, AU
- A. Seza Doğruöz
Ghent University, BE
- Robert Frank
Yale University, US
- Juan Luis Gastaldi
ETH Zürich, CH
- Adele Goldberg
Princeton University, US
- Coleman Haley
University of Edinburgh, GB
- Aurelie Herbelot
Denotation – Pritzwalk, DE
- Yu-Yin Hsu
Hong Kong Polytechnic University, CN
- Mark Johnson
Macquarie University – Sydney, AU
- Najoung Kim
Boston University, US
- Alessandro Lenci
University of Pisa, IT
- Lori Levin
Carnegie Mellon University – Pittsburgh, US
- Roger Levy
MIT – Cambridge, US
- Xixian Liao
Barcelona Supercomputing Center, ES
- Kyle Mahowald
University of Texas – Austin, US
- Tom McCoy
Yale University, US
- Joakim Nivre
Uppsala University, SE
- Alexis M. Palmer
University of Colorado – Boulder, US
- Christopher Potts
Stanford University, US
- Jakob Prange
Universität Augsburg, DE
- Siva Reddy
MILA – Montreal, CA & McGill University – Montreal, CA
- Philip Resnik
University of Maryland – College Park, US
- Anna Rogers
IT University of Copenhagen, DK
- Rachel Rudinger
University of Maryland – College Park, US
- Gözde Gül Şahin
Koç University – Istanbul, TR
- Asad Sayeed
University of Gothenburg, SE
- Nathan Schneider
Georgetown University – Washington, DC, US
- Noah A. Smith
University of Washington – Seattle, US
- Mark Steedman
University of Edinburgh, GB
- Tiago Torrent
Federal University of Juiz de Fora, BR
- Bonnie Webber
University of Edinburgh, GB
- Ethan Wilcox
Georgetown University – Washington, DC, US
- Adina Williams
Meta Platforms – New York, US
- Amir Zeldes
Georgetown University – Washington, DC, US