

LUHME 2026

**The 3rd Workshop on Language Understanding in the
Human-Machine Era**

Proceedings of the Workshop

October 28, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-471-2

Introduction

The question whether machines understand language or merely simulate the appearance of understanding has moved from a philosophical curiosity to one of the central open problems in natural language processing. Large language models (LLMs) now seem to perform at – or close to – human level on a wide range of tasks, from summarization to parsing to dialogue (Manning, 2022; Subbanarasimhaiah, 2025), across many languages. In doing so, they have made the question harder to dismiss, rather than easier. If a system produces fluent, contextually appropriate language without anything resembling a human cognitive architecture behind it, what exactly is being tested when we ask whether that system “understands” language?

This is the question that has motivated the workshop Language Understanding in the Human-Machine Era (LUHME), and which becomes even more relevant in this third edition, collocated with EMNLP 2026, in Budapest. The disagreement on which LUHME sits is thus not new; however, as systems have improved, it has grown into something closer to a genuine controversy. Winograd’s (1972) work on natural language understanding fifty years ago treated comprehension as a problem of translating sentences into formal procedural representations inside a rule-based scenario – a tractable, even if narrow, concept of understanding related to what a system could do with an instruction. By the end of that decade, however, Winograd (1980) himself had turned that definition into a question rather than an answer, by arguing that “understanding” could not be reduced to symbolic manipulation without addressing the more profound question of what understanding a situation actually requires. That unresolved turn from understanding as procedure to understanding as an open cognitive and philosophical problem runs directly through the debate as it currently stands.

On one end of the current controversy are researchers who argue that understanding cannot be defined independently of causal, grounded cognition. Mitchell and Krakauer (2022) frame the dispute explicitly as a disagreement over whether LLMs construct genuine mental models that capture causal mechanisms, or whether their apparent competence is an artefact of scale and statistical coverage (in which case “understanding” is a kind of sophisticated interpolation rather than the real thing). McClelland et al. (2020) push this further, as they argue that language cannot be meaningfully separated from an “integrated understanding system”. For them, comprehension depends on tying linguistic symbols to multi-sensory perception, motor representation, and episodic memory, none of which text-only models have access to. On this account, a model trained purely on sequences of tokens is, by design, cut off from the kind of construal that ordinary language use presupposes (goal-oriented, culturally situated, addressed to a particular interlocutor).

On the other end of the controversy are researchers who treat this grounding requirement as either overstated or empirically beside the point. Manning (2022) argues that distributional models trained at sufficient scale end up encoding networks of linguistic and encyclopaedic relations that are sufficiently rich to support genuinely human-competitive reasoning and language use. This is not a simulation of understanding, but a different, statistically-derived route to essentially the same competence. Gastaldi (2020) goes further to treat the question “why can computers understand natural language?” as answerable on its own terms, without the need to first settle debates about consciousness or embodiment. For this faction, the operational tradition inherited from early computer science (whereby understanding is taken as demonstrated by what a system can correctly do with language, rather than by what internal architecture it possesses) is not a lesser standard, but simply a different, more manageable one (Wang et al., 2023; Kuiler, 2020).

In between these two poles stands a body of research that accepts evidence for both positions and treats the resulting picture as genuinely unresolved rather than as a matter of interpretation. Lenci’s critical analysis (2023) is the sharpest statement of this: according to the author, current NLU systems demonstrably capture ‘something’, but their well-documented brittleness under paraphrase, adversarial input, and distribution shift shows that whatever they capture diverges from the robust, revisable inference that characterises human understanding. Havlík (2023) frames the same tension as a question about meaning itself by inquiring whether “meaning” in a LLM can be anything more than a functional repla-

cement for the meaning that a speaker intends (this is perhaps a question that recent gains in benchmark performance have not settled and arguably cannot settle by benchmark performance alone). McShane and Nirenburg (2021), who write from a more applied linguistics perspective, take this as an argument to steer research in a different direction: building explicit, linguistically-informed knowledge structures into NLP systems rather than trusting distributional exposure to deliver understanding as a byproduct of scale.

Grounded, distributional, and critically agnostic approaches contribute to building a three-way disagreement that maps relatively well (even if not perfectly) onto disciplinary lines. Theoretical linguistics tends toward denotational semantics grounded in hierarchical, rule-governed competence (Sciullo, 2009); cognitive science, instead, leans towards causal mental models verified against human processing data (Wang et al., 2023); early computer science privileges formal procedural translation (Winograd, 1972); and contemporary NLP favours parametrised distributional representations enhanced for task performance (Manning, 2022; McShane and Nirenburg, 2021). Yet, the more interesting friction in the current literature is not so much between them as it is within them: linguists are divided over whether grounding is a precondition or a red herring, and NLP researchers are divided over whether brittleness is a solvable engineering problem or a sign that something structurally different from human understanding is going on (Lenci, 2023; Mitchell and Krakauer, 2022).

It is this live, unresolved disagreement, rather than a settled consensus to be reached, that gives the LUHME workshop its focus, especially in this edition. Contributors were invited to address how meaning is represented, inferred, and negotiated in human-machine interaction, with particular attention to logic-aware, pragmatics-sensitive, and socially grounded accounts of understanding, on the premise that the grounding debate is not merely academic once these systems are deployed in high-stakes domains such as law, health, finance, and governance, where the practical difference between genuine understanding and plausible output has direct consequences (Navigli, 2018). The resulting workshop programme and respective proceedings span a wide range of subjects directly related to natural language understanding, including: evaluating language understanding in LLMs; pragmatics, context and meaning beyond the literal; human-machine comparisons of language understanding; conversational understanding, repair and interaction; instruction following and context-sensitive behaviour; multilingual, cross-lingual and low-resource language understanding; reasoning, memory and internal representations; human-centred, social and safety dimensions of language understanding; and the interpretability and accountability of systems that produce linguistically fluent but potentially misleading output.

Held for the third time, LUHME continues to serve as a locus where these positions can be argued, rather than assumed. The papers that follow do not certainly resolve the dispute between grounded and distributional accounts of understanding, but they advance it, from a range of disciplinary points, toward a more precise account of what current systems do and do not yet do.

The LUHME Workshop Organisers

Rui Sousa-Silva, Henrique Lopes Cardoso, Maarit Koponen, Antonio Pareja-Lora

Acknowledgments

This research was financially supported by UID/00022/2025 – Centre for Linguistics of the University of Porto (CLUP) with DOI <https://doi.org/10.54499/UID/00022/2025> and UID/00027/2025 – Artificial Intelligence and Computer Science Laboratory (LIACC) with DOI <https://doi.org/10.54499/UID/00027/2025>, funded by Fundação para a Ciência e a Tecnologia, I.P./MECI Portugal through national funds, and by project NORTE IA+ | Transparent and Safe Artificial Intelligence for a More Humane and Competitive Future (2021-2027 Programme, supported by FEDER via NORTE2030-FEDER-02724200).

References

- Juan Luis Gastaldi. 2020. [Why can computers understand natural language?](#) *Philosophy & Technology*, 34:149–214.
- Vladimír Havlík. 2023. [Meaning and understanding in large language models](#). *Synthese*, 205.
- Erik W. Kuiler. 2020. [Natural Language Processing \(NLP\)](#), pages 1–3. Springer International Publishing, Cham.
- Alessandro Lenci. 2023. [Understanding natural language understanding systems. a critical analysis](#). *ArXiv*, abs/2303.04229.
- Christopher D. Manning. 2022. [Human language understanding & reasoning](#). *Daedalus*, 151:127–138.
- James L. McClelland, Felix Hill, Maja R. Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. [Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models](#). *Proceedings of the National Academy of Sciences*, 117:25966 – 25974.
- Marjorie McShane and Sergei Nirenburg. 2021. [Linguistics for the Age of AI](#). The MIT Press.
- Melanie Mitchell and D. Krakauer. 2022. [The debate over understanding in ai’s large language models](#). *Proceedings of the National Academy of Sciences of the United States of America*, 120.
- Roberto Navigli. 2018. [Natural language understanding: Instructions for \(present and future\) use](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5697–5702. International Joint Conferences on Artificial Intelligence Organization.
- A. D. Di Sciullo. 2009. *New Trends in Software Methodologies, Tools and Techniques*, chapter Natural Language Understanding. IOS Press.
- Narendra Subbanarasimhaiah. 2025. [How do machines understand language? a beginner’s guide to natural language processing](#). *World Journal of Advanced Research and Reviews*.
- Shaonan Wang, Nai Ding, Nan Lin, Jiajun Zhang, and Chengqing Zong. 2023. [Language cognition and language computation - human and machine language understanding](#). *ArXiv*, abs/2301.04788.
- Terry Winograd. 1972. [Understanding natural language](#). *Cognitive Psychology*, 3(1):1–191.
- Terry Winograd. 1980. [What does it mean to understand language?](#) *Cogn. Sci.*, 4:209–241.

Organizing Committee

Workshop Organisers

Rui Sousa-Silva, University of Porto, Portugal
Henrique Lopes Cardoso, University of Porto, Portugal
Maarit Koponen, University of Eastern Finland, Finland
Antonio Pareja-Lora, Universidad de Alcalá, Spain

Proceedings Chair

Gil Rocha, University of Porto, Portugal

Website Chair

David Freitas, University of Porto, Portugal

Program Committee

Program Committee

Aida Kostikova, Bielefeld University
António Branco, University of Lisbon
Barbara Lewandowska-Tomaszczyk, University of Applied Sciences in Konin
Belinda Maia, University of Porto
Chaya Liebeskind, Jerusalem College of Technology
Diana Santos, Universitetet i Oslo
Efsthios Stamatatos, University of the Aegean
Ekaterina Lapshinova-Koltunski, University of Hildesheim
Elena Lloret, University of Alicante
Eliot Bytyçi, Universiteti i Prishtinës “Hasan Prishtina”
Federico Ruggeri, University of Bologna
Gil Rocha, University of Porto
Haim Dubossarsky, Queen Mary University of London
Lynne Bowker, Université Laval
Marcos Zampieri, George Mason University
Pablo Gamallo, Universidade de Santiago de Compostela
Paula Carvalho, University of Aveiro
Paolo Rosso, Universitat Politècnica de València
Tharindu Ranasinghe, Lancaster University

Keynote
**Show Me the Work: Human-Centric Automated
Fact-Checking**

Isabelle Augenstein
University of Copenhagen

Abstract: As AI systems increasingly participate in how information is evaluated online, understanding why a system reached a conclusion may be as important as the conclusion itself. In this talk, I will present our work on human-centric automated fact-checking. Starting from studies of professional fact-checkers' explanation needs, I will show how these requirements motivate research on evidence-centred explanations and uncertainty communication for fact-checking systems. Across user studies and model evaluations, our findings suggest that people rely heavily on evidence when assessing AI-generated judgements and benefit from explanations that make reasoning paths, information gaps, and conflicts between sources explicit. I will argue that automated fact-checking provides a particularly rich testbed for studying language understanding in the human-machine era, where successful collaboration depends on making evidence, reasoning, and uncertainty transparent to users.

Table of Contents

<i>PCP: A Controlled Benchmark for Pragmatic Intention Recognition in Clinical Communication</i> Kun Zhang	1
<i>Where Does Conversational Repair Happen? Meaning Revision Inside Language Models</i> Vedanth Nanesha	25
<i>Reading Between the Turns: User Repair as a Context-Dependent Signal of Breakdown and Recovery in Multi-Turn LLM Assistants</i> Leah Michel, Nuno Miguel Martins Pacheco, Choro Ulan uulu, Mikhail Kulyabin, Jan Joosten, Filippos Petridis, Michael A. Hedderich and Barbara Plank	38
<i>Beyond Fluent Responses: Finding User Dissatisfaction in Human–AI Interactions</i> Yuka Saito, Ikumi Numaya, Jun Suzuki and Reina Akama	54
<i>Representational Similarity Alignment of Figurative Language in Humans and Large Language Models</i> Samhita Bollepally, Aurora Sloman-Moll, Takashi Yamauchi and Neeka Motamedrasa	70
<i>False Memories in LLMs: Gist-Based Recognition Despite Verbatim Context</i> Robin Orthey, Guillaume Pelat and Yuri Nakao	89
<i>Human and Machine Word Associations in Slovenian: A Comparison of Associative Behaviour and Its Explanations</i> Klara Vrabl, Mojca Brglez and Špela Vintar	98
<i>Task Competence Is Not Instruction Following: Evaluating Instruction-Conflicting Behavior in Small Language Models</i> Mahdiyeh Farajidizaji and Vatsal Raina	109
<i>Probing Language Understanding in Colombian Indigenous Languages: Zero-Shot Translation, Visual Grounding, and Cultural Fidelity</i> Santiago Arenas and Ruben Manrique	121
<i>How AI Assistants Respond to Repeated Abuse</i> William Guey, Wei Zhang, Pierrick Bougault, Yi Wang, Agoston Bodo, Vitor D de Moura and José O Gomes	134
<i>PlotTwist: A Creative Plot Generation Framework with Small Language Models</i> Abhinav Thorat, Ravi Kumar Kolla, Jyotin Goel, Madhav Kataria and Niranjana Pedanekar ..	144
<i>Assessing the Comprehension of Regional Colombian Idioms by Language Models</i> Santiago Bobadilla Suarez, Juan Diego Osorio, Maria Alejandra Pinzon Roncancio, Ignacio Charro and Ruben Manrique	173
<i>Representing Human Values in Argumentation: A Linguistically Grounded Comparison of Lexicons, Embeddings, and LLMs</i> Chaya Liebeskind, Barbara Lewandowska-Tomaszczyk and Shmuel Liebeskind	183
<i>Calibrating LLM-as-a-Judge for Argument Quality Assessment through Prompting and Scoring Scale Design</i> Guilherme Aren Marchetti, Gil Rocha and Henrique Lopes Cardoso	197
<i>Lexical Redefinition Reveals Context-Sensitive Instruction Following</i> Roksana Goworek and Haim Dubossarsky	215

<i>To “Ba” or Not to “Ba”? LLM Surprisal as a Diagnostic of Context-Sensitive Language Understanding in L2 Chinese</i>	
Shengjie Zhang and Yan Cong	227
<i>Same Problem, Different Answer: Accuracy Hides Cross-Lingual Inconsistency in Multilingual Inference</i>	
Ahsan Habib Shihab, Assel Ospan, Aman Mussa and Madina Mansurova	246
<i>FrugalPrompt: Reducing Contextual Overhead in Large Language Models via Token Attribution</i>	
Syed Rifat Raiyan, Md Farhan Ishmam, Abdullah Al Imran and Mohammad Ali Moni	256
<i>Sweet Talkers: How Query Formulation Shapes Sycophancy in Romantic Relationship Advice</i>	
Helena Choi, Edric Castel C. Hao, Karl Bautista, Francis Gabriel A. Magleo, Renzo Panti and Danielle Beatrice Ledesma Olalia	272
<i>Beyond Warmth: Risk-Tiered Auditing of Credibility in LLM Companion Responses</i>	
Danping Yan	286
<i>Evaluation of Motivational Interviewing Counsellors with Task-Aware Multi-Stage LLM-Based Simulated Clients</i>	
Jiading Zhu, Xinyu Cindy Wang, Thomas Nguyen, Yan Qing Lee, Osnat C. Melamed, Peter Selby and Jonathan Rose	304
<i>Who Benchmarks the Benchmarks? Towards Comprehensive Evaluation of Commonsense Reasoning Benchmarks</i>	
Pavel Chizhov, Anton Changalidis, Vishnu Prasad Vijaya Kumar, Yannick Detrois, Mattia Nee, Pierre-Carl Langlais and Ivan P. Yamshchikov	360
<i>Inverse Turing Bench: Evaluating Language Models as Judges of Human vs. AI Dialogue</i>	
William Hager, Ishika Rathi, Masum Hasan and Cameron Robert Jones	378
<i>Sedimented Use and Borrowed Competence: A Relational Framework for Language Models</i>	
Yanan Long	389