Conversational Grounding in the Era of LLMs
Workshop Booklet

ESSLLI Week one:
July 29 – August 2, 2024
Leuven, Belgium

Thanks for sponsorship to Inria Paris

Welcome!

We are proud to welcome you to the 2024 ESSLLI workshop on conversational grounding in the era of large language models!

Conversational grounding is the process through which collective knowledge and assumptions are built by interlocutors over the course of a conversation. This shared understanding involves continuous negotiation and resolution of uncertainties through providing additional context or clarification. Once shared in a manner satisfactory to the participants, the entities thus added to the common ground can be referred back to using reduced referring expressions such as pronouns. Effective grounding mechanisms are vital for dialogue systems, as they reduce ambiguity and facilitate effective and efficient communication, whether the system is the speaker or listener. As should be clear, this process encompasses more than the mere exchange of information through statements; it involves a complex interaction (often including visual cues if the conversation is taking place face-to-face or in shared space), inferential reasoning, and dynamic feedback mechanisms. Despite extensive research in this field, numerous challenges persist, especially when adapting to diverse scenarios and contexts. These can include varying numbers and types of conversational partners, different sensory and action capabilities, and distinct objectives.

The advent of large language models may introduce significant shifts in how grounding processes are understood, as well as new challenges to overcome, even as they offer novel solutions to existing problems. Now is therefore an opportune time to unite researchers and professionals who are exploring the various facets of conversational grounding. This workshop there aims to foster a deeper collective understanding (or common ground!) concerning this topic, and to share the latest best practices in the design, modeling, and application of conversational grounding in dialogue systems.

In this booklet you will find:

- A schedule for the 5 days of the workshop, listing when the different authors will be present, as well as the schedule for some great hands-on activities, and discussion topics suggested by our speakers.
- extended abstracts by the authors accepted for the workshop, as well as from our 2 invited speakers - Kristiina Jokinen and Massimo Poesio, both of whom are widely recognized for their groundbreaking work on the topic of conversational grounding in dialogue systems.
- A background to the topic of conversational grounding by David Traum, a co-organiser of the workshop, who published the first computational account of conversational grounding in 1992.
- For any questions, feel free to contact us at this email address: CG24_esslli@outlook.com

We look forward to welcoming you at 2pm on Monday July 29th in Leuven!

Table of Contents

SCHEDULE		
29TH JULY, MONDAY		
14:00 - 14:10	JUSTINE CASSELL	Audience design and workshop overview
14:10 - 14:20	DAVID TRAUM	Introduction to Conversational Grounding
14:20 - 14:35	ALL PARTICIPANTS	Participant Introductions
14:35 - 15:05	KRISTIINA JOKINEN	**Invited Talk** - From Words to the Real World: Conversational grounding and GenAI Models for Dialogues
15:05 - 15:30	GROUP DISCUSSION	Architectures for LLM-based dialog systems that ground.
30TH JULY, TUESDAY		
14:00 - 14:05	DAVID TRAUM	Review of Day 1
14:05 - 14:35	MASSIMO POESIO	**Invited Talk** - Conversational Agents and Reference in Minecraft
14:35 - 14:55	HARUHISA ISENO	Comparative Analysis of Common Ground Representations in Dialogue Systems: a Case Study using the OneCommon Task
14:55 - 15:30	GROUP DISCUSSION	Multimodal Grounding with LLMs
31ST JULY, WEDNESDAY		
14:00 - 14:05	JUSTINE CASSELL	Review of Day 2
14:05 - 14:25	YASUHIRO KATAGIRI	Grounding and Higher-Order Cooperation for Human-LLM Dialogues
14:25 - 14:45	BISWESH MOHAPATRA	Evaluating the Effectiveness of Large Language Models in Establishing Conversational Grounding
14:45 - 15:05	GROUP DISCUSSION	Modelling Common Ground
15:05 - 15:30	ALL PARTICIPANTS	Group Activity
1ST AUGUST, THURSDAY		
14:00 - 14:05	DAVID TRAUM	Review of Day 3
14:05 - 14:25	AMELIE ROBRECHT	Towards an Analysis of Discourse and Interactional Pragmatic Reasoning Capabilities of Large Language Models
14:25 - 14:45	GROUP DISCUSSION	Ethical issues arising from grounding with LLMs
14:45 - 15:30	ALL PARTICIPANTS	Group Activity
2ND AUGUST, FRIDAY		
14:00 - 14:05	JUSTINE CASSELL	Review of Day 4
14:05 - 14:25	MARIE BOSCARO	Cooperative Norms for Assertions and Conversational Grounding in X
14:25 - 14:55	GROUP DISCUSSION	What have we not yet covered?
14:55 - 15:30	DAVID TRAUM	Final Review of the workshop and future possible directions

Invited talk

From Words to the Real World: Conversational Grounding and GenAI Models for Dialogues

Kristiina Jokinen

AI Research Center, AIST Tokyo Waterfront JAPAN

Keywords: grounding, Generative AI, trustworthy dialogue systems, large language models, knowledge graphs, language-capable robots, conversation, cooperation

In recent years, research and development on interactive AI-agents have focused on new models for AI applications that aim to improve human-robot cooperation and communication in tasks like navigation, education, support for healthy living, and eldercare services. These applications must provide truthful and accurate information, besides being able to communicate in a natural and fluent manner.

Generative AI has radically changed dialogue modelling research and made fluent chatting a common model of interaction. It has also raised concerns due to the tendency of LLMs to fabricate facts and generate false information. As a solution, retrieval augmented generation (RAG) has been introduced to supply information from up-to-date and reliable documents, referred to as "Grounding the LLM in the documents". Another approach for trustworthy interaction is a graph-based RAG which uses knowledge graphs representing the structure of the world. This approach can be referred to as "Grounding the LLM in the real world", given that the knowledge graph is a suitable approximation of the real-world situation.

The notion of grounding has also been an important notion in the long tradition of non-generative dialogue modelling, where grounding refers to a collaborative mechanism for establishing mutual knowledge among participants. Such grounding is a necessary skill for language-capable AI-agents, and the recent progress in generative AI has brought the issues of “conversational grounding” into foreground.

In this talk I will discuss the di erent types of grounding from the point of view of trustworthy interaction in the context of LLMs. I will emphasise the importance of mutual knowledge and collaboration and focus on the integration of LLMs with knowledge graphs in order to develop reliable AI agents. I will give examples from our ongoing work and discuss implications, challenges, and opportunities to develop AI-agents that support friendly interaction in the era of LLMs.

Biography:

Kristiina Jokinen is Senior Researcher at AI Research Center (AIRC) at National Institute of Advanced Industrial Science and Technology (AIST) in Tokyo Waterfront, and Adjunct Professor at University of Helsinki. She is a member of the pan-European AI network of excellence ELLIS, Advisory Board for Japanese AIE (AI in Engineering) Programme, and Steering Committee for International dialogue workshop series IWSDS. She has led numerous national and international research projects, and most recently she led dialogue research in the EU-Japan collaboration project e-VITA.

She received her first degree at University of Helsinki, and her PhD from UMIST, Manchester. She was awarded a JSPS Fellowship for PostDoc research at NAIST (Nara Institute of Science and Technology), and was Invited Researcher at ATR Research Labs in Kyoto, and Visiting Professor at Doshisha University.

Her research concerns human-robot interaction, (Gen)AI-based dialogue modelling and multimodal communication, and she has published widely on these topics. She developed Constructive Dialogue Model as a general framework for interaction, and together with Graham Wilcock she developed the Wikipedia-based robot dialogue system WikiTalk, which won the Special Recognition for Best Robot Design (Software Category) at the International Conference of Social Robotics in 2017.

Invited Talk

Conversational Agents and Reference in Minecraft

Massimo Poesio, Chris Madge, and Juexi Shao

The domain of **virtual world games**–games in which conversational agents impersonating characters can learn to perform tasks, or improve their communicative ability, by interacting with human players in platforms such as Minecraft or Light–is a promising approach to study how conversational agents can carry out grounded interaction (Johnson et al., 2016; Urbanek et al., 2019; Narayan-Chen et al., 2019; Szlam et al., 2019; Kiseleva et al., 2022; Zhou et al., 2023). Virtual world games may approach the complexity of the real world; and virtual agents operating in such virtual worlds need to be able to develop a variety of interactional skills to be perceived as 'real' (Schlangen, 2023).

Among these virtual world games, Minecraft has become particularly popular in NLP as a domain to study several aspects of interpretation. The Minecraft Dialogue Corpus Narayan-Chen et al. (2019) has been used to investigate dialogue act interpretation, abstract meaning representation (AMR) interpretation, and discourse structure interpretation Thompson et al. (2024).

In the ARCIDUCA project Poesio et al. (2022), we are using Minecraft as an environment to study conversational agents. In this presentation, I will first of all discuss our work on using LLMs to act as Builder or Architect in the Minecraft Dialogue Corpus task Madge and Poesio (2024) and to carry out the two tasks in the IGLU version of Minecraft Kiseleva et al. (2022). Next, I will discuss our annotation of the Minecraft Dialogue Corpus for reference and coreference, soon to be released. Finally, I will discuss ongoing work on developing a conversational agent for the Minecraft domain able to carry out referring expression interpretation.

References

Johnson, Matthew, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The malmo platform for artificial intelligence experimentation. In *Proc. of IJCAI*, pages 4246–4247.

Kiseleva, Julia, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Cˆot´e, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, Ahmed Awadallah, Seungeun Rho, Taehwan Kwon, Daniel Wontae Nam, Felipe Bivort Haiek, Edwin Zhang, Linar Abdrazakov, Guo Qingyam, Jason Zhang, and Zhibin Guo. 2022. Interactive grounded language understanding in a collaborative environment: Retrospective on Iglu 2022 competition. In *Proceedings of the NeurIPS 2022 Competitions Track, PMLR*, vol. 220, pages 204–216.

Madge, Chris and Massimo Poesio. 2024. Large language models as minecraft agents. In *Proceedings of Wordplay*.

Narayan-Chen, Anjali, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proc. of the 57th Annual Meeting of the ACL*, pages 5405–5415.

Poesio, Massimo, Richard Bartle, Jon Chamberlain, Julian Hough, Chris Madge, Diego Perez-Llebana, Matt Purver, and Juntao Yu. 2022. Arciduca: Annotating reference and coreference in dialogue using conversational agents in games. In *Proc. of the Workshop on the Semantics and Pragmatics of Dialogues (SEMDIAL)*.

Schlangen, David. 2023. Dialogue games for benchmarking language understanding: Motivation, taxonomy, strategy. arXiv:2304.07007 [cs.CL].

Szlam, Arthur, Jonathan Gray, Kavya Srinet, Yacine Jernite, Armand Joulin, Gabriel Synnaeve, Douwe Kiela, Haonan Yu, Zhuoyuan Chen, Siddharth Goyal, Demi Guo, Danielle Rothermel, C. Lawrence Zitnick, and Jason Weston. 2019. Why build an assistant in Minecraft? arXiv: 1907.09273.

Thompson, Kate, Julie Hunter, and Nicholas Asher. 2024. Discourse structure for the Minecraft corpus. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4957–4967. Torino, Italia: ELRA and ICCL.

Urbanek, Jack, Angela Fan, Siddharth Karamcheti, Saachi Jain, Emily Dinan, Tim Rockt¨aschel, Douwe Kiela, Arthur Szlam, Samuel Humeau, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. arXiv preprint arXiv:1903.03094.

Zhou, Pei, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 11136–11155. Association for Computational Linguistics, Toronto, CAN.

Biography:

Massimo Poesio is a professor at Queen Mary University, London, and at the University of Utrecht. His research interests include disagreement in language interpretation, anaphoric reference, semantic interpretation in dialogue, and games and NLP. His current ARCIDUCA project focuses on developing agents able to operate and interact with human players in virtual world games such as Minecraft or Dungeons and Dragon

Introduction to Conversational Grounding

David Traum, USC ICT

traum@ict.usc.edu

The term "Grounding" is used for many different concepts. Some are clearly unrelated to issues of establishing common ground between humans & machines, such as forcing a pilot not to fly, or attaching an electrical circuit to the earth. Others are somewhat related to current purposes, like “symbol grounding” (Harnad 1990) or “visual grounding” (Haber et al 2019, Shridhar et al 2020), which have more to do with associating language with non-linguistic concepts and the visual scene, respectively. Here, following Clark & Wilks-Gibbs (1986), we will use “grounding” to mean the processes of recognizing, establishing, maintaining, updating, and repairing common ground between participants in a conversation. We will often refer to this as “conversational grounding” to indicate that this is a process between individuals, not internal to a single individual incorporating perceptual inputs with mental states. It is also sometimes referred to as “social grounding”, e.g. (Dillenbourg and Schneider 1995). Common ground results from grounding processes, and can include various kinds of information, including facts and beliefs, but also assumptions, goals, and plans. Common ground as presumed background information is seen as a requirement for almost any communication (Stalnaker 2002). The study of common ground and grounding is interdisciplinary, studied in fields including philosophy, communication studies, linguistics, psychology, education, and computer science.

Here we briefly address some background literature that may be helpful in establishing common ground for the participants. We address the following questions: (1) What kind of information is in common ground? (2) In what domains has grounding been studied? (3) How is common ground represented in a formal or implementable model? (4) How is the grounding process modelled?

1. What kind of information is in common ground?

Common ground can include at least the following kinds of information: Facts about the current situation, Beliefs shared among individuals, Meanings of words, Aspects of dialogue context, such as most recent or most salient entities of various types. Dillenbourg and Traum (2006) identified the following types of information in their domain that were grounded differently (different rates of explicit feedback moves):

· Facts – information that was directly perceivable in the environment or part of explicit communication from another conversational participant

· Inferences – information that requires interpretation from what is explicitly observed or communicated (possibly involving other context and common ground)

· Management – discussion about how to proceed with a task

· Meta-communication – discussion about the communication process, including new communicative conventions

· Technical problems – discussion about how the user interface works and how to do things.

Clark (1996) discusses 4 levels of communication at which grounding takes place, including: behaviors, signals, meaning, and joint projects, mirroring the 4 feedback functions described by Allwood et al (1992): contact, perception, understanding, and attitudinal reaction.

2. In what communicative activities has grounding been studied?

Grounding has been studied in many different communicative settings, with differences including numbers of participants, relationship of participants, physical setting, communicative modalities available, topic of conversation, specific tasks the participants are trying to accomplish, and types of conversational and other actions undertaken by each participant role. We list here some different activities in which aspects of grounding have been studied:

· Casual conversation

· Direction giving

· Collaborative Planning & Decision-making

· Question-answering system

· Collaborative Learning

3. How is common ground represented in a formal or implementable model?

Schiffer (1972) proposed modelling common ground (or mutual knowledge as an infinite sequences of nested beliefs, of the form A and B mutually know X if each knows it and knows the other knows it and knows the other knows they know it,…
MK(A,B,X) = K(A,X), K(B,X), K(A,K(B,X), K(B,K(A,X)), K(A,K(B,K(A,x))),…

Cohen and Levesque (1991) introduced a one-sided version that they called BMB, that just concerns a single agent's beliefs about nested beliefs, and is thus more directly accessible by this agent.
Lewis (1969) proposed a model of common ground as a shared situation, in which everyone in a population has reason to believe a situation, the situation itself indicates that everyone has a reason to believe it, and the situation indicates to the population that some proposition X holds. This approach allows derivation of each of the nested beliefs of Schiffer, but using a more compact representation. This approach is also adopted by Clark and Marshall (1981) and Barwise (1987).

Another approach is to treat common ground as a primitive, not defined using other more basic components. From this, many of the same inferences about individual and nested beliefs could be derived.

Common Ground is sometimes modelled not as a binary concept (either present or absent), but as a matter of degree. This follows the general idea that there can be more or less support for or evidence of common ground or the elements it is built from. Paek and Horvitz (2000) model grounding as decision under uncertainty, and emerge with a probability-based concept of grounding for each of Clark (1996)'s four levels, using Baysian networks.

Roque and Traum (2008) has a model including degrees of grounding, where the amount of grounding is related to the type of evidence available.

4. How is the grounding process modelled?

Some common ground is assumed to already exist rather than established explicitly within a conversation. For instance, if A, a competent English Speaker, observes that B is also a competent English speaker, then A can assume that the meaning of most common English words are common ground between A and B. Secular speakers might assume the calendar date of Christmas is known (even if they do not celebrate Christian holidays), but may be less likely to assume knowledge of when Assumption day occurs or what it is meant to celebrate (while Catholic Priests might be expected to know both).

Some common ground is added explicitly by inference going from what is already known (or assumed) and new observations to conditions about a participants knowledge that are likely to hold for the observed behavior to be sensible. When participants are in a shared situation, these kinds of inference can lead to common ground between the participants. These inferences are generally defeasible and there may be cases where they are not correct leading one party to assume common ground where it does not actually exist (or sometimes to mutual assumptions about common ground but different views of what is actually in the common ground). Clark and Marshall (1981) present a set of copresence assumptions of different sorts that can license new additions to common ground depending on the type of shared situation. These include community co-membership, physical copresence, linguistic copresence, and indirect mixtures of these. For each of these, they lay out a set of relevant auxiliary assumptions that can be used to infer new common ground.

Much of adding to common ground involves not just inference from assumptions and passive observations of a common situation, but explicit communicative action by the participants. Many theories and computational systems assume that a single action performance is sufficient to add the communicative content to common ground, especially if copresence assumptions seem valid. The new common ground can contain both explicit content (e.g., what is said or pointed to), but also implicit information that can be inferred from what is said and the situation. For example, some utterances presuppose some information is already part of common ground rather than stating it as new. If it is not actually known already, then this information can be *accommodated* (Lewis, 1979, Stalnaker 1973, 2002). Von Fintel (2008) makes clear how accommodation can be used to add presupposed information to common ground, if it is not already present, and when it might be problematic to do so.

While in some cases, it may be sufficient to assert, implicate, or presuppose information with proper co-presence assumptions in order to get this content added to common ground. However, in other cases it clearly is not. In dialogue, there are many kinds of linguistic feedback (Allwood et al 1992) that are used to give evidence of how well listeners are able to receive, perceive, understand, and accept information from a speaker. This feedback can be positive or negative, or positive at some levels while being negative at others (Clark and Shaefer 1987, Clark 1996). Negative feedback can also lead to conversational repairs, that modify the content to be added to common ground or the evidence that it should be.

An influential early model of the grounding process that takes into account both initial utterances and subsequent feedback and repair from multiple utterances is the “Contributions” model by Clark and Schaefer (1989). This

assumes that common ground is added only at the successful creation of contributions, which consist of two parts, *a presentation phase* followed by an *acceptance phase:*

Presentation Phase: A presents utterance u for B to consider. He does so on the assumption that, if B gives evidence e or stronger, he can believe that B understands what A means by u.

Acceptance Phase: B accepts utterance u by giving evidence e′ that he believes he understands what A means by u. He does so on the assumption that, once A registers evidence e′ , he will also believe that B understands.

Both parts can be a sequence of utterances or a subdialogue rather than a single utterance. Each can also recursively contain sub-contributions that move some content into common ground as part of grounding the whole content. This leads to *contribution graphs* to represent how parts of a dialogue are related to establish grounding. Other factors influencing this model include a *grounding criterion* that influences how strong evidence is needed for a particular type of information and situation, and *strength of evidence*, that concerns what kinds of feedback actions provide which levels of evidence.

A general principle of *least collaborative effort* (Clark and Wilks-Gibbs 1986), accompanied with a cost of grounding for different kinds of communications (Clark and Brennan, 1991) enable decision-making about what kinds of grounding moves can establish the grounding criterion with lowest cost.

As Traum (1994,1999) points out, there are some difficulties with trying to use this model for dialogue agents engaged in grounding. Some of these are addressed by Cahn and Brennan (1999), who create contribution graphs for each participant, and have an implementation for a question answering dialogue system. The model was also extended by Novick et al 1996, to allow for more than two participants.

(Traum & Allen 1992, Traum 1994) present a computational account of the grounding process, in terms of individual *grounding acts* that can be performed and recognized in real time by conversational participants. A unit of dialogue structure (initially called Discourse Units (DUs), but later Common Ground Units (CGUs) (Traum & Nakatani, 1999), roughly similar to a top-level Contribution, in the Clark and Schaefer model, is constructed from individual grounding acts, and has both a content that would be added to common ground if the unit is grounded, and a state, representing what has been done and remains to be done to ground the content.

The original grounding acts model had 7 different grounding acts. Katagiri and Shimojima extended this with a Display act, that involves explicit feedback but could end up being an acknowledgement or repair-request or repair, depending on the certainty of the speaker and whether the responder is correct or not. Di Maro (2021) reviews work relating to each of the types of grounding act from Traum (1994).

Roque and Traum (2008, 2009), present a revised grounding model that accounts for different degrees of grounding coming from different sequences of action, which was present in the contribution model but missing in the grounding acts model.

Visser et al 2014 extend the Grounding acts model to cover incremental dialogue processing, where interpretations are made while a speaker is still speaking (and motivates grounding-related interruptions and overlaps).

Nakano et al., (2003) study grounding in a face-to-face setting, and point out (so to speak) that grounding acts need not occur in language. A confirmation grounding act may be a nod, while a shift in eye gaze towards a speaker may indicate the need for a repair.

Bavelas et al (2012, 2014) point out that in several domains, including therapy and education, a simple two-part model with a presentation by one party and feedback/acceptance from another is not sufficient for grounding. Instead, at least one more acknowledgment turn by the initial speaker is needed, to provide assessment as a kind of 3rd turn feedback whether the evidence provided by B is actually sufficient (and correct). While both the Contributions model and Grounding acts model do allow for this additional feedback, they do not require it to ground the original utterance (as opposed to material presented in the second utterance). This highlights the importance of taking the domain of conversation into account. A degrees of grounding model can distinguish levels of grounding where only a second turn acknowledgement is present from a case where third turn feedback is also present.

McRoy and Hirst (1995) focus on recognizing and repairing misunderstanding, including cases where participants have different strengths of belief about some facts, might be disposed to accept another's information as more likely, and in which interrogative utterances can be seen as ambiguous between an information-seeking question, and a "pre-tell" offer to tell if the other doesn't already know. They use an abductive model to represent different strengths of belief and interpretations of ambiguous utterances.

Paek and Horvitz (1996) Quartet architecture treats grounding as decision making under uncertainty, where Bayesian networks are used to model the grounding state and actions are selected to maximize mutual understanding. Clark's 4 levels are explicitly modeled as different parts of the network.

Udagawa and Aizawa (2020, 2021) use a neural net model to represent the grounding process and engage in dialogue. The model encodes spatial features of objects and performs target selection and reference resolution tasks, in addition to language response generation.

An important consideration for grounding is *information packaging.* Even once a decision has been made about what information needs to be added to common ground and what kinds of presentation or feedback should be done to accomplish this, there is still a question of how to package information into specific communicative actions. There are several sub-issues for information packaging, including (Bibyk et al 2001):

· How much information to put in each utterance? As Clark and Schaefer point out, a presentation could be all as one utterance, or installment by parts, such as individual or small groups of digits in a phone or social security number.
· Which aspects of information should be presented explicitly or implicitly (to be derived from the explicit parts in context)?
· What order should the information parts be presented in? Following issues of entrainment, familiar orders may be easier to interpret and more successful than other orders.

Some behavior that is considered part of grounding has been attributed to other processes. For example, it is a much-observed fact that as a dialogue proceeds, dialogue participants tend to produce communications that are

more similar at multiple levels (e.g. semantic, syntactic, prosodic) to those of the other participants. This process is sometimes called entrainment (Brennan 1996). This process has been explained as making grounding more efficient and a result of an audience design process (Clark and Wilks-Gibbs, 1986). But it has also been explained through a less conscious process of interactive alignment (Pickering & Garrod, 2004), or task-oriented synergy (Fusaroli et al 2014). While it seems that some aspects of entrainment can be explained by low level features, others do seem to require cognitive processes like conceptual pacts (Carbary & Tanenhaus, 2011). Moreover, Rothwell et al (2021) found that a complementarity model was better correlated with task success than an alignment model.

References

1. J. Allwood, Nivre, J. and Ahlsén, E. "On the Semantics and Pragmatics of Linguistic Feedback" in Journal of Semantics, 1992

2. Barwise, Jon. "Three views of common knowledge." Proceedings of the 2nd conference on Theoretical aspects of reasoning about knowledge. 1988.

3. Bavelas, Janet Beavin; De Jong, Peter; Smock Jordan, Sara; and Korman, Harry (2014) "The theoretical and research basis of co-constructing meaning in dialogue," Journal of Solution Focused Practices: Vol. 1 : Iss. 2 , Article 3.

4. Bavelas, J. B., De Jong, P., Korman, H., & Jordan, S. S. (2012, September). Beyond back-channels: A three-step model of grounding in face-to-face dialogue. In Proceedings of Interdisciplinary Workshop on Feedback Behaviors in Dialog(pp. 5-6).

5. Bibyk, Sarah A., Leslie M. Blaha, and Christopher W. Myers. "How Packaging of Information in Conversation Is Impacted by Communication Medium and Restrictions." Frontiers in Psychology 12 (2021): 1096.

6. Brennan, Susan E. "Lexical entrainment in spontaneous dialog." Proceedings of ISSD 96 (1996): 41-44.

7. Cahn, J. E., & Brennan, S. E. (1999). A psychological model of grounding and repair in dialog. Proceedings, AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems (pp. 25-33). North Falmouth, MA: American Association for Artificial Intelligence.

8. Carbary, K., & Tanenhaus, M. (2011). Conceptual pacts, syntactic priming, and referential form. In Proceedings of the CogSci Workshop on the Production of Referring Expressions: Bridging the Gap Between Computational, Empirical and Theoretical Approaches to Reference (PRE-CogSci 2011) (pp. 1-6).

9. Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press.

10. Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, & I. Sag (Eds.), Elements of discourse understanding (pp. 10-63). Cambridge: Cambridge University Press.

11. Clark, Herbert H., and Edward F. Schaefer. "Collaborating on contributions to conversations." Language and cognitive processes 2.1 (1987): 19-41.

12. Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. Cognitive Science , 13, 259-294.

13. Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition , 22, 1-39.

14. Cohen, P. R., & Levesque, H. J. (1991). Teamwork. *Nous 25*.4: 487-512.

15. Dillenbourg, P. & Schneider, D. (1995). Collaborative learning and the Internet. Proceedings of the International Conference on Computer Assisted Instruction (ICCAI) (pp. S-10-6 - S-10-13). Hsinchu: Taiwan, 7-10 March 1995.

16. Pierre Dillenbourg and David Traum, Sharing solutions: persistence and grounding in multi-modal collaborative problem solving. in Journal of the Learning Sciences, 15:121--151, 2006.

17. Di Maro, Maria. "Computational Grounding: An Overview of Common Ground Applications in Conversational Agents." IJCoL. Italian Journal of Computational Linguistics 7.7-1, 2 (2021): 133-156.

18. Fusaroli, Riccardo, Joanna Rączaszek-Leonardi, and Kristian Tylén. "Dialog as interpersonal synergy." New Ideas in Psychology 32 (2014): 147-157.

19. Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., & Fernández, R. (2019, July). The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1895-1910).

20. Harnad, Stevan. "The symbol grounding problem." Physica D: Nonlinear Phenomena 42.1-3 (1990): 335-346.

21. Y. Katagiri and A. Shimojima. Display acts in grounding negotiations. In Proceedings of Gotalog 2000, the 4th Workshop on the Semantics and Pragmatics of Dialogue, pages 195–198, 2000.

22. David K. Lewis. Convention: A Philosophical Study. Harvard University Press, 1969.

23. David Lewis(1979). Scorekeeping in a language game. In Semantics from different points of view (pp. 172-187). Springer, Berlin, Heidelberg.

24. Susan W. McRoy and Graeme Hirst. The repair of speech act misunderstandings by abductive inference. Computational Linguistics, 21(4):5–478, 1995.

25. Nakano, Y., Reinstein, G., Stocky, T., Cassell, J. (2003) "Towards a Model of Face-to-Face Grounding" *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* July 7-12, Sapporo, Japan.

26. Novick, D., Walton, L., and Ward, K. (1996). Contribution graphs in multiparty conversations, Proceedings of the International Symposium on Spoken Dialogue (ISSD-96), Philadelphia, PA, October, 1996, 53-56.

27. T. Paek and E. Horvitz. “Conversation as Action Under Uncertainty” Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000), Stanford, CA, June 2000.

28. Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and brain sciences, 27(2), 169-190.

29. Antonio Roque and David Traum, Degrees of Grounding Based on Evidence of Understanding In proceedings of The 9th SIGdial Workshop on Discourse and Dialogue (SIGdial 2008), June, 2008.

30. Antonio Roque and David Traum, Improving a Virtual Human Using a Model of Degrees of Grounding, in proceedings of International Joint Conerence on Artificial Intelligence IJCAI-09, Pasadena, CA.

31. Rothwell, Clayton D., Valerie L. Shalin, and Griffin D. Romigh.(2021) "Comparison of Common Ground Models for Human--Computer Dialogue: Evidence for Audience Design." ACM Transactions on Computer-Human Interaction (TOCHI) 28.2 (2021): 1-35.

32. Stephen R. Schiffer. Meaning. Oxford University Press, 1972.

33. Shridhar, Mohit, Dixant Mittal, and David Hsu. "INGRESS: Interactive visual grounding of referring expressions." The International Journal of Robotics Research 39.2-3 (2020): 217-232.

34. Stalnaker, Robert. 1973. Presuppositions. Journal of Philosophical Logic 2:447-457.

35. Robert Stalnaker. 2002. “Common Ground.” Linguistics and Philosophy 25 (5/6): 701–21.

36. David Traum, A Computational Theory of Grounding in Natural Language Conversation, TR 545 and Ph.D. Thesis, Computer Science Dept., U. Rochester, December 1994.

37. David R. Traum Computational Models of Grounding in Collaborative Systems, in working notes of AAAI Fall Symposium on Psychological Models of Communication, p. 124-131, November, 1999.

38. David R. Traum and James F. Allen, A Speech Acts Approach to Grounding in Conversation, In Proceedings 2nd International Conference on Spoken Language Processing (ICSLP-92), pages 137-40, October 1992.

39. David R. Traum and Christine H. Nakatani, A Two-level Approach to Coding Dialogue for Discourse Structure: Activities of the 1998 Working Group on Higher-level Structures, in proceedings of the ACL'99 Workshop Towards Standards and Tools for Discourse Tagging, pp 101--108, June 1999.

40. Udagawa, Takuma, and Akiko Aizawa. "An annotated corpus of reference resolution for interpreting common grounding." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05. 2020.

41. Udagawa, Takuma, and Akiko Aizawa. "Maintaining Common Ground in Dynamic Environments." Transactions of the Association for Computational Linguistics 9 (2021): 995-1011.

42. Visser, T., Traum, D., & DeVault, D. (2014). A model for incremental grounding in spoken dialogue systems. Journal on Multimodal User Interfaces, 8(1), 61-73.

43. Von Fintel, Kai. "What is presupposition accommodation, again?." Philosophical perspectives 22 (2008): 137-170.

Cooperative Norms for Assertions and Conversational Grounding in X

Marie Boscaro[1], Anastasia Giannakidou[2], Alda Mari[1], and Valentin Tinarrage[1]

[1]IJN CNRS/ENS-PSL/EHESS [2]University of Chicago

1 Question and Scope

Assertions are acknowledged to have one aim : to offer to add p (i.e., the asserted propositional content) to the common ground shared by the participants of the conversation (see a.o. Stalnaker, (1978), (2002); Clark & Brennan, (1991), Traum (1994), Beyssade and Marandin, (2009), Farkas & Bruce (2010), Krifka (2015), Geurts (2019)).

The grounding of p in *common knowledge* is the result of mutual acceptance (Farkas & Bruce (2010), Clark & Brennan (1991), Ginzburg, J. (1996)). This acceptance relies on fulfilling norms of cooperative assertions - as summed up in the *Veridicality Principle for cooperative assertions* : one must assert p if and only if one believes in p or knows p to be true (see a.o. Searle (1975), Grice (1975), Bach and Harnish, (1984), Davidson (1985), Vanderveken (1990), Harnish (1994), Williamson (1996), Williams (2010), Portner (2018), Giannakidou and Mari (2021a), Lauer (2013),) (see for commitment as act Farkas and Bruce, (2010), Krifka (2015), Guerts (2019), a.o.). Additionally, it relies on indicating some particular evidence for the asserted claim. It has been recognized that one's evidence for a propositional content can facilitate its grounding in common knowledge and might influence its degree of groundedness. The groundedness level of a propositional content depends on the type of speech act studied (see a.o. Traum (1994)) but also on the evidence type used to justify it and on the strength and trustworthiness associated with the evidence. For example, reported evidence has been analyzed as a weak evidence and is recognized to be unpreferred for grounding information (Faller (2002), Krifka (2023)). Faller (2002) observed that Assertions with Relayed evidence did not aim at adding the propositional content asserted to the common knowledge fo the hearer but only at presenting it, thereby achieving weak grounding.

Additionally, it has been observed in the literature that the norms governing grounding might evolve depending on the conversational medium studied (Brennan (2014), Clark & Brennan (1991)). Clark and Brennan (1991) identified 8 factors governing evolving grounding: *co-presence, visibility and audibility, simultaneity, spontaneity, sequentiality, reviewability, revisability* and claim that when a medium lacks one of these characteristics, it can modify the grounding. This modification incurs different costs (for example understanding or production costs). Emails for example have understanding costs because this medium is neither cotemporal nor sequential making understanding harder because the addressee has to reconstruct the utterance context (see Clark & Brennan (1991)).

We propose to analyze the evolution of conversational grounding and of the norms and constraints that govern grounding on the social media platform X (formerly known as Twitter). We claim that the factors governing grounding are different on X because they are governed by new constraints, including a specific algorithm (which discriminates the information disseminated on the platform), a specific utterance situation (based on delocalization), and the use of extralingusitic tools (such as hyperlinks, #, mention @).

To conduct this survey, we did an empirical study on a corpus of French tweets on ecological crisis (Kozlowski et al (2020)) (fires, hurricanes, storms, flooding etc). We observed a significant markedness of information sources (and in particular with hyperlinks) and that Reported information sources constitute the default evidence for bare assertions in X.

We propose in our study that the norms for the production of Assertions are redefined in X, and suggest three new grounding principles: (1) the *Sustain a topic function*, (2) *Veridicality* (following the Traditional Picture see a.o. (Grice (1975), Giannakidou & Mari (2021)), (3) *Affiliation Function*– the latter two necessitating a new of definition of cooperativity as not just adding to the common ground but also headlining a topic and signaling affiliation to a social group (for a further view on cooperativity see Grice (1975), Allwood et al (2000)).

We suggest that one information is grounded in the common knowledge of the speakers (in X) when the speakers fulfill these three norms. We also claim that the information added to the common ground can be polymorph : the propositional content itself, but also additional information on the speaker (his social identity) and the relevance of the tweet itself.

2 Data

Our study relies on a French ecological crises corpus of 13,378 tweets gathered in 2019 (Kozlowski et al. (2020), Laurenti et al. (2022)) and already annotated for speech acts categories: *Assertives, Subjectives, Interrogatives, Jussives* following Laurenti et al. (2022)'s framework. We will primarily focus on *Assertives* statements which are bare declarative sentences with no mark of subjectivity (no hedges, epistemic modals, or perspectival elements such as 'I believe, in my opinion', predicates of personal taste) (see examples in 1). Assertives convey a stronger veridical commitment (Giannakidou & Mari 2021a,b) and offer to add the propositional content to the common ground.

Relying on several categorizations for evidentiality (Aikhenvald (2004), Willett (1988)), including those discussed in the NLP literature (a.o. Castillo et al. (2011), Zahra et al. (2020)), we identified four main evidential type categories [1]: *Direct*, *Relayed*, *Loose sources* and *Lack of testimony*. We will study more in depth the two *Reported* evidentials : *Relayed* (1) and *Loose Source* (2). *Relayed* evidentiality in social media is conveyed by extralinguistic markers of information source, most notably hyperlink, mention (@) and #sourcename. *Loose sources* are marked with a mere # where related information can be found, without a precise link leading to the source of the information. The source of information is not directly specified in the tweet content contrary to *Relayed* statement.

(1) *Relayed : Reported evidence*: des rafales de vent jusqu'à 110 km/h attendues dans l'Yonne http://ift.tt/2EAdBaJ
Wind gust up to 110km/h expected in Yonnes

(2) *Loose sources : Reported evidence*:
#VentViolent cette nuit : forte migration de Normands vers l'Alsace ... #ViolentWind tonight: strong migration from Normandy to Alsace

[1] Note that we are not dealing here with grammatical evidentials

The annotation of evidentiality was performed by two annotators with a kappa of 0.7. We found that the most frequent category is *Relayed* which accounts for 62.64%. We also studied the correlations between speech acts categories (and especially Assertives) and Evidential categories.

In the table 1 we highlight the significant positive correlations and the slightly positive correlations between type of evidentiality and Speech act categories. We observed that Assertives statement are highly correlated to Relayed evidence and that Loose Sources are dispreffered for Assertive Statements. Furthermore, we found a high markedness of information source with 65, 37% of the tweets (on a sample of 1000 tweets) which contain at least one hyperlink.

Evidentiality	Assertive	Subjective	Interrogative	Jussive	Total
Direct	123 (3.92%)	75 (2.39%)	6 (0.19%)	17 (0.54%)	221 (7.04 %)
Relayed	1442 (45.97%)	161 (5.13%)	33 (1.05%)	326 (10.39%)	1962 (62.64%)
Loose Sources	150 (4.78%)	217 (6.92%)	26 (0.83%)	22 (0.70%)	415 (13.23%)
No Testimony	177 (5.64%)	235 (7.49%)	31 (0.99%)	96 (3.06%)	539 (17.18%)
Total	1892 (60.31%)	688 (21.93%)	96 (3.06%)	461 (14.70%)	3137 (100%)

Table 1: Evidentiality vs Speech Acts

3 Analysis and Discussion

We propose that the discourse constraints in social media are redesigned and the traditional norms governing the grounding of information are too limited in this context.

We claim that the conversational constraints identified by Clark and Brennan (1991) for face-toface conversation but also for other conversational mediums (e.g. emails, letters, video teleconference, telephone etc) are different for discourse on produced X. Discourses disseminated online have new features : speakers make expensive use of new extralinguistic tools (hyperlinks, #, mention @), the utterance situation is based on a delocalization and , more strikingly , discourses are governed by a specific algorithm that discriminates the information relayed online as more or less relevant. The discrimination is based on the choice of topic but also on the use of extralinguistic tools (hyperlinks, #, mention @, pictures etc). The more relevant information is classified as “Twitter Trend” whereas the less relevant is invisibilized on X users’ feeds.

In our empirical study, we observed that *Relayed* evidence is the strongest candidate for grounding the speaker’s veridical commitment (ie the strong correlation between *Relayed* and *Assertives*).

Relayed evidence in this scenario might fulfill what we call the Traditional Picture, the Gricean conversational norm “Be truthful” (Quality maxim) and its sub-maxim (“Don’t say what you don’t have sufficient reason to believe is true.”). Speakers choose to mark their evidence within their discourse in order to justify their veridical commitment but also to encourage the other conversational participants to incorporate the propositional content into the common ground. However, marking one’s statement with a Relayed evidence has not been recognized in the traditional literature as facilitating the grounding of propositional content (Faller (2002), Krifka (2003)). In X, therefore, the threshold of evidence for assertions becomes more lax, as reportative assertions become the norm (corroborating the idea of X as a platform for information dissemination see a.o. Kwak et al (2010)) and the choice of evidence depends on the speaker’s evaluation of the trustworthiness of the evidence.

However, we believe that this predominance of Relayed evidence for Assertions can be analyzed as a response to some new rules governing information grounding.

On social networks we observe a high markedness (almost obligatory) of information sources, contrary to what is traditionally observed in nonevidential languages. We argue that this high markedness, along with the predominance of Relayed Assertives, serves purposes beyond merely adding truthful information to the common knowledge of the conversational participants. These two new purposes are intertwined with the new conversational constraints observed on X.

The first purpose is to *Sustain a topic*: we propose that all hyperlinks, regardless of their function or content, aim to sustain or establish relevance and popularity. Integrating one hyperlink is a way increase the visibility of a tweet and enhance its likelihood of being shared or retweeted. In this scenario, conversational participants ground in their common knowledge the information that the tweet is relevant and highly significant, especially if it is part of the Twitter Trend.

The second purpose is the *Affiliation Function*, serving as means to highlight one's affiliation to a political, ideological or social group. Selecting one specific information source instead of another is a way to emphasize ones belonging to a particular community. In this case, conversational participants ground their common knowledge on information about the social identity of the speaker, such as their political or ideological beliefs.

The cooperativeness of speakers on social media,their information grounding and the marking of their information sources are governed by three norms : *Sustain a topic, Affiliation and Veridicality*. Grounding in this type of discourse is polymorphic, encompassing the grounding of propositional content, political or ideological affiliation, or the relevance of one's information.

References

Aikhenvald, Alexandra Y. Evidentiality. OUP Oxford, 2004.

Allwood, J., Traum, D., Jokinen, K. (2000). Cooperation, dialogue and ethics. International Journal of Human-Computer Studies, 53(6), 871914.

Bach, K., & Harnish, R. M. (1979). Communication and speech acts. Cambridge, Mass.: Harvard.

Bach, K. and R. M. Harnish (1984). Linguistic communication and speech acts. 2nd edition, Cambridge [Mass.] ; London : MIT Press.

Beyssade, C. and J.-M. Marandin (2009). Commitment: une attitude dialogique. Langue francaise (2), 89–107.

Brennan, S. E. (2014). The grounding problem in conversations with and through computers. In Social and cognitive approaches to interpersonal communication (pp. 201-225). Psychology Press.

Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), Perspectives on socially shared cognition (pp. 127–149). American Psychological Association.

Davidson, D. (1985). Communication and convention. Dialogue: An interdisciplinary approach, 11–26.

Faller, M. T. (2002). Semantics and pragmatics of evidentials in Cuzco Quechua. Stanford university.

Farkas, D. F. and K. B. Bruce (2010). On reacting to assertions and polar questions. Journal of semantics 27 (1), 81–118.

Geurts, B. (2019). Communication as commitment sharing: speech acts, implicatures, common ground. Theoretical linguistics 45 (1-2), 1–30.

Giannakidou, A., & Mari, A. (2021a). Truth and veridicality in grammar and thought: Mood, modality, and propositional attitudes. University of Chicago Press.

Giannakidou, A., & Mari, A. (2021b). A Linguistic Framework for Knowledge, Belief, and Veridicality Judgment. KNOW: A Journal on the Formation of Knowledge, 5(2), 255-293.

Ginzburg, J. (1996). Dynamics and the semantics of dialogue. Logic, language and computation, 1, 221-237.

Goldberg, S. C. (Ed.). (2020). The Oxford handbook of assertion. Oxford University Press. Grice, H. P. (1975). Logic and conversation. In Speech acts (pp. 41-58).

Harnish, R. M. (1994). Mood, meaning and speech acts. In S. L. Tsohatzidis (Ed.), Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives, pp. 407–459. Routledge.

Kozlowski, D., E. Lannelongue, F. Saudemont, F. Benamara, A. Mari, V. Moriceau, and B. A. (2020). A three-level classification of french tweets in ecological crises. Information Processing & Management. 57 (5), 1–46.

Krifka, M. (2015). Bias in commitment space semantics: Declarative questions, negated questions, and question tags. In Semantics and linguistic theory, Volume 25, pp. 328–345.

Krifka, M., Hartmann, J. M., & Wollstein, A. (2023). Layers of assertive clauses: Propositions, judgements, commitments, acts. Propositionale Argumente im Sprachvergleich: Theorie und Empirie/Propositional Arguments in Cross-Linguistic Research: Theoretical and Empirical Issues (= Studien zur deutschen Sprache 84). Tübingen: Narr, 115-182.

Kwak, H. and Lee, C. and Park, H. and Moon, S.(2010), What is Twitter, a social network or a news media?, in Proceedings of the 19th international conference on World wide web, 591–600.

Lauer, S. (2013). Towards a Dynamic Pragmatics. Ph. D. thesis, Stanford University.

Laurenti, E., N. Bourgon, F. Benamara, A. Mari, V. Moriceau, and C. Courgeon (2022). Give me your intentions, i'll predict our actions: A two-level classification of speech acts for crisis management in social media.

In N. Calzolari, F. B´echet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022, pp. 4333–4343. European Language Resources Association.

Laurenti, E., Bourgon, N., Benamara, F., Mari, A., Moriceau, V., & Courgeon, C. (2022, July). Speech acts and communicative intentions for urgency detection. In 11th Joint Conference on Lexical and Computational Semantics (* SEM 2022). ACL: Association for Computational Linguistics.

Matthewson, L. (2020). Evidence type, evidence location, evidence strength. In Evidentials and modals (pp. 82-120).

Pagin, Peter and Neri Marsili, "Assertion", The Stanford Encyclopedia of Philosophy (Winter 2021 Edition)

Portner, P. (2018). Mood. Kettering, Northamptonshire, UK: Oxford University Press.

Speas, M. (2010). Evidentials as generalized functional heads. Edges, heads, and projections: Interface properties, 127-150.

Stalnaker, R. C. (1978). Assertion. In Pragmatics (pp. 315-332)

Stalnaker, R. (2002). Common ground. Linguistics and philosophy, 25(5/6), 701-721.

Traum, D. (1994). A computational theory of grounding in natural language conversation, Thesis, University of Rochester. Dept. of Computer Science.

Vanderveken, D. (1990). Meaning and speech acts. Principles of Language Use. Cambridge: Cambridge University Press. Vol.1.
Willett, T. (1988). A cross-linguistic survey of the grammaticization of evidentiality. Studies in Language. International Journal sponsored by the Foundation "Foundations of Language", 12(1), 5197.

Williamson, T. (1996). Knowing and asserting. The Philosophical Review, 105(4), 489-523.

Williams, B. (2010). Truth and truthfulness: An essay in genealogy. Princeton University Press.

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. IP&M, 57(1)

Comparative Analysis of Common Ground Representations in Dialogue Systems: a Case Study using the OneCommon Task

Haruhisa Iseno ***Ryuichiro Higashinaka***
Graduate School of Informatics, Nagoya University
iseno.haruhisa.h4@s.mail.nagoya-u.ac.jp
higashinaka@i.nagoya-u.ac.jp

Abstract

For realizing sophisticated dialogue systems, modeling common ground is crucial. Recently, dialogue systems have been developed where the common ground between interlocutors is represented in some form, and using this representation, large language models (LLMs) generate utterances. However, the optimal way to represent common ground and input it into an LLM remains unclear. As a preliminary step in identifying the optimal representation of common ground for dialogue systems utilizing LLMs, this study constructs two types of LLM agent using deterministic and probabilistic representations of common ground, respectively, and quantitatively evaluates their effectiveness in the OneCommon task.

1 Introduction

Common ground refers to the shared knowledge and beliefs among interlocutors (Clark, 1996). For dialogue systems to be able to engage in more sophisticated interactions, it is essential to model dialogues that incorporate this common ground.

In recent years, many dialogue systems have been developed by using large language models (LLMs) controlled by prompts for utterance generation. By representing common ground in some form and providing it as a prompt to the LLMs, it is anticipated that more sophisticated dialogues that consider common ground can be conducted. However, the optimal form for representing common ground is not yet clear.

This study investigates the optimal representation of common ground for dialogue systems using LLMs. Specifically, we consider two types of representation: a deterministic representation, which includes only belief with certainty, and a probabilistic representation, which includes belief with its probabilities. The impact of these representations on dialogue is examined by dialogue simulations.

2 Related work

In recent years, there have been attempts to generate utterances using LLMs on the basis of common ground in some form, as in (Qiu et al., 2023). Dialogue tasks focusing on common ground have also been proposed (Kim et al., 2019). The OneCommon task (Udagawa and Aizawa, 2019), among others, involves two participants, each given a different visual arrangement of dots. Through dialogue, they attempt to identify dots visible in both viewpoints. This task necessitates that the participants effectively establish a common ground regarding their respective viewpoint. With the availability of the dataset and the simplicity of the task, this study focuses on the OneCommon task.

3 Experiment

3.1 Setup

The experiment explored two types of representation for common ground: a deterministic representation and a probabilistic representation. Using these representations, agents based on GPT-4 were constructed to represent common ground on prompts and generate utterances to complete tasks.

For evaluation, dialogue simulations were conducted among the agents to measure dialogue performance. The simulations continued until the agents completed the task or exceeded a predetermined number of turns, set at 20 for this study. The effectiveness of each representation was quantitatively evaluated on the basis of two metrics: the success rate of the agents in completing the task and the average number of turns in a dialogue.

3.2 Systems for comparison

Five different agents were constructed for comparative analysis as follows.

Baseline This agent generated utterances based on the task description and dialogue history.

	No. dialogues	Task success rate↑	Average turns↓
Baseline	30	40.0	6.0
Deterministic	30	60.0	9.4
w/o belief of self from viewpoint of partner	30	50.0	10.4
Probabilistic	30	30.0	9.75
w/o belief of self from viewpoint of partner	30	43.3	9.5

Table 1: Dialogue simulation results

Deterministic This agent enumerated dots (dotIDs with their coordinates) about which there was certainty regarding “belief of partner” and “belief of self from viewpoint of partner” using GPT-4, and it generated utterances based on this information. Additionally, a variant of this agent was constructed without including “belief of self from viewpoint of partner” to evaluate the effectiveness of having the recursive structure of common ground.

Probabilistic This agent listed dots regarding “belief of partner” and “belief of self from viewpoint partner” along with the probabilities that these considerations would be established as common ground, using GPT-4 to generate utterances. The same variant as Deterministic was also prepared for comparison.

3.3 Result

Table 1 presents the task success rates and average dialogue turns for the simulations in which each agent chatted with an agent with the same configuration. Focusing on the task success rates, agents utilizing deterministic representations showed up to a 20-point increase compared with the Baseline, whereas those employing probabilistic representations did not demonstrate improvements in task success rates. This suggests that leveraging probabilistic information may be challenging for LLMs. Probabilistic values are often overlooked, causing all information to be seen as certain. This could have resulted in misunderstandings. Furthermore, all agents, compared with the Baseline, exhibited an increase in the number of dialogue turns.

Examining the content of dialogues conducted by each agent, significant differences were observed in the granularity of information grounded during the conversations. The Baseline agent tended to switch topics easily, moving on to new subjects without sufficiently grounding information, often resulting in less detailed discussions. In contrast, the agent achieving the highest task success rate (Deterministic) displayed a tendency to thoroughly ground information before proceeding to the next topic. Specifically, there were more utterances asking for details about a single piece of information. The result demonstrates that representing common ground in an appropriate format can enhance the grounding capabilities of LLMs, leading to more successful task completions.

4 Conclusion

In this study, two types of agents were constructed using deterministic and probabilistic representations of common ground. By simulating dialogues between these constructed agents in the OneCommon task, their conversational abilities were evaluated. The results indicated that agents employing deterministic representations of common ground exhibited higher dialogue performance compared with those using probabilistic representations.

Future research should extend evaluations to other dialogue tasks, such as those in (Kim et al., 2019), to further assess the effectiveness of various common ground representations in addition to those used in this study.

References

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative drawing as a testbed for grounded goaldriven communication. In *Proc. ACL*, pages 6495– 6513.

Shuwen Qiu, Song-Chun Zhu, and Zilong Zheng. 2023. MindDial: Belief dynamics tracking with theory-ofmind modeling for neural dialogue generation. In
Proc. the First Workshop on Theory of Mind in Communicating Agents.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proc. AAAI*, pages 7120–7127.

Grounding and Higher-Order Cooperation for Human-LLM Dialogues

Yasuhiro Katagiri

AIST Artificial Intelligence Center

Joint activities, including dialogues, executed by a group of independent minds always have a potential to go awry. Grounding in dialogue is a mechanism to avoid or at least reduce the possibility of the missteps that could happen in a dialogue interaction. Dialogue failures occur in multiple types at multiple levels: misperception of a word, misidentification of a sentence, or misunderstanding of what is said or intended by the speaker. Speaker and Hearer are both responsible to work to remove the errors to make the interaction go back on track. But, repair operations are themselves impediment to dialogue objectives. Usually, the longer it takes before the error gets noticed and attended, the more cumbersome it becomes. Since speech is a slow medium, people invariably follow a dialogue version of the least effort principle to omit facts, sentiments, or rules they believe shared between them and their interlocutors. And when their beliefs happen to be wrong, they eventually resort to repair operations. Dialogue is a balancing act between the achieving of an adequate level of sharing and efficient proceedings of interaction.

Grice observed that a general principle of cooperation plays an underpinning role to support human conversational interactions. At the base level of cooperation, following of conversational maxims makes it possible for a speaker to convey and for a hearer to recover what is implicated beyond what an utterance literally means. On top of that, the speaker and hearer can share further implicatures, e.g., ironies, by flouting these maxims, relying on the assumption that they both follow the cooperative principle . Their shared presupposition that they both observe the cooperative principle enables them to engage in higher-order cooperation.

Similarly in conversational grounding, dialogue participants not only perform grounding acts when they are relevant, they also share a presupposition that the speaker and hearer are both committed to work together to preserve common ground. This shared presupposition works to motivate them to attend to the grounding status to intervene in case of an error, as well as to produce an expectation of the same behavior from their partner. This higher-order cooperation makes it possible for dialogue participants to jointly realize a dynamic balance of sharing and efficiency. Let's look at the following dialogue snippet:

(1) a. A: Did you see Sato-san in the meeting yesterday?
b. B: Yes, he seemed to be happy because of his recent promotion.
c. A: He got promoted?
d. B: Didn't you mean Yo Sato?
e. A: Um, I meant Toshi Sato.
f. B: Oh, sorry. No, I didn't see him yesterday.

A's initial presentation (1a), a question, was followed by B's continuous presentation (1b), an answer to the A's question, which in this case also functions as an acceptance to ground A's initial presentation (1a). The status of grounding is not complete but presupposed at this stage, and B actually misidentified who A was actually talking about. Higher-order cooperation in grounding prescribes that this presupposition holds unless an error is discovered and presented by either party. If no error is discovered, this prescription can support quick and efficient flow of dialogue. In this example, B eventually raised a possibility of an error and initiated a repair in the fourth turn (1d). However, it could well have been A who has to initiate the repair if B had kept his presupposition by saying in place of (1d): "Yes, he was promoted last month to a senior position." The point is both A and B are responsible in maintaining grounding status, and higherorder cooperation based on their joint commitment to this responsibility is indispensable in human dialogue grounding behaviors.

Compare this with a human-LLM dialogue about a travel planning advice.

(2) a. User: Can you suggest a travel plan to Taiwan for three days?

b. LLM: Absolutely! Here's a three-day travel plan for Taiwan:

⋯ <Travel itinerary suggestion> ⋯

c. User:What about the effects of the earthquake?

d. LLM: Taiwan is located in a seismically active region,

... <General description of effects of earthquake on travel in Taiwan> ⋯

e. User:What itinerary are you talking about?

f. LLM: My apologies for the confusion.

⋯ <Excuse and additional advice information> ⋯

g. User:I am concerned about the earthquake that occurred in Hualien.

h. LLM: I understand your concern. In February 2018, a magnitude 6.4 earthquake struck near Hualien City in eastern Taiwan.

⋯ <Explanation of an old earthquake and general advice on travel to

Hualien after an earthquake > ⋯

i. User:I'm talking about the very recent one. Do you have any information on the damage it caused?

j. LLM: I'm sorry for the oversight.

⋯ <Excuse and further advice information> ⋯

All the advices are admittedly helpful in their contents, but LLM's grounding behavior is obviously different from human's. Higher-order cooperation appears to be completely missing. LLM's responses to User's queries always take the form of continuous presentations. As was the second turn utterance of B in the previous example (1b), they indicate presupposition of successful grounding of what the User said. However, rather than to address the potential problem in grounding, it orients itself to cram more information into an advice ignoring the User's orientation to get back on track in establishing some amount of common ground. User is single-handedly required to take care of grounding. This asymmetry has (at least) two problems: (1) the lack of higher-order cooperation is noticed and it gives uncooperative or domineering impressions to users, (2) the style of always signaling presupposed grounding by returning continuous presentations provides no clues to users how to adjust the level of information they should disclose to maintain the balance between sharing and efficiency. A mechanism of higher-order cooperation, even if restricted to information on people or events introduced into the discourse, would go a long way toward building a human-friendly LLM dialogue agent.

Evaluating the Effectiveness of Large Language Models in Establishing Conversational Grounding

Biswesh Mohapatra[1], Manav Kapadnis[2], Laurent Romary[1], Justine Cassell[1,3]
1 - Inria, 2 - IIT Kharagpur, 3 - Carnegie Mellon University

1 Introduction

The concept of "common ground" in linguistics refers to the collective knowledge and assumptions built by interlocutors over the course of a conversation. This shared understanding involves continuous negotiation and resolving uncertainties through additional context or clarification and, once clarified, can be referred back to using reduced referring expresions such as pronouns. Effective grounding mechanisms are vital for dialogue systems to reduce ambiguity and facilitate effective and efficient communication, whether the system is the speaker or listener. Research in both linguistics and conversational agents has addressed grounding challenges, especially in rule-based modular dialog systems. Recent studies have shown the potential of LLMs in interactive settings, however the application of grounding concepts to contemporary LLMs is limited. And those studies that have addressed grounding in pre-trained LLMs, lacked a comprehensive evaluation framework. This paper therefore takes on the challenging of evaluating the performance of various LLMs with respect to different aspects of conversational grounding, analyzes performance differences, and proposes methods to enhance the capabilities of underperforming models.

To this end, we developed a series of tests to evaluate LLMs in the role of both speaker and listene, iin various grounding scenarios. They cover abilities such as handling repaired or canceled information, asking clarification questions, and providing disambiguating information. The evaluation method involves analyzing the model perplexity of correct and incorrect responses to given contexts that were developed based on the Meetup dataset (Ilinykh et al. 2019). Meetup was chosen as it involves two participants navigating a 2D grid with the objective of converging in the same room, despite only being able to see their own rooms, and only one room at a time. Participants must articulate room descriptions, formulate strategies for finding one another, remember shared descriptions, and mentally model the other participant's room configurations. The task requires continuous negotiation in the use of referring expressions, making it an ideal candidate for studying conversational grounding. Additionally, its structure, where both participants can interchangeably assume initiator or responder roles, mirrors the dynamic nature of real conversations, further enhancing its suitability for this research. Furthermore, the dataset involves images of rooms which, can be deployed to test grounding in the real world. With 430 dialogues and 5131 utterances, the dataset provides a rich resource for testing language models.

In order to assess whether the size of the training set impacts performance, the study tests LLMs of varying sizes, including T5-Large, Godel-Large, Llama (7 & 13 Billion), GPT 3.5, and GPT 4. The appropriate and inappropriate responses were designed to look similar to one another, but to require a good pragmatic understanding of the context to reply correctly. We also finetuned the open-sourced models on the Meetup

dataset using next-token prediction to test whether fine-tuning helps these models or not. For closed-sourced models like GPT 3.5 and GPT 4, we conducted proompt testing rather than perplexity due to the lack of available model weights. Key findings from the perplexity testing were that:

1. Larger models like Llama-13B, GPT-4 performed significantly better across categories.
2. Smaller models like T5 and Godel struggled to differentiate between contextually appropriate and inappropriate responses, particularly for repair and request repair.
3. Fine-tuning smaller and medium-sized models increased their likelihood of generating dataset-like utterances but did not improve understanding.
4. Llama-13B demonstrated superior performance over GPT-3.5, suggesting that the size of the pre-training dataset influenced grounding capabilities.

In order to explicate these results, we examined the models'embedding vectors. We found that the smaller models tended to rely on lexical over pragmatic content. However, The consistent superior performance of the fine-tuned Llama model over its original version suggested the potential benefits of finetuning to enhance performance. Based on the previous analysis, additional training data were created using GPT-4, and smaller and medium-sized models were fine-tuned using Positive and Negative Reward Training. This method improved the performance of medium-sized models like Llama 7B & 13B models, but smaller models like T5-Large and Godel-Large showed limited improvement.

The study suggests that pragmatic abilities such as generating and understanding grounding acts may be emergent properties in LLMs trained on extensive data. It also highlights the potential for targeted training to improve the performance of medium-sized models, offering a balance between grounding capabilities and computational efficiency. We note that models trained on larger amounts of data might perform better than larger models in some cases. However, we note that the Dataset offered limited numbers of examples of some of the grounding categories, therefore limiting our ability to compared the different grounding acts. We addressed this by conducting additional tests on specific grounding acts with larger occurances in the dataset, which confirmed our initial findings. Future work will explore techniques like Direct Preference Optimization and Reinforcement Learning from Human Feedback. We have also begun to design approaches to including data from the speech signal and other modalities, as they have shown their importance in conversational grounding (Nakano et al. 2003).

References

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet up! a corpus of joint activity dialogues in a visual environment. In Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers, London, United Kingdom. SEMDIAL.

Yukiko Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 553–561, Sapporo, Japan. Association for Computational Linguistics.

Towards an Analysis of Discourse and Interactional Pragmatic Reasoning Capabilities of Large Language Models

Amelie Robrecht
Social Cognitive Systems
Bielefeld University

Judith Sieker
Computational Linguistics
Bielefeld University

Clara Lachenmaier
Computational Linguistics
Bielefeld University

Sina Zarieß
Computational Linguistics
Bielefeld University

Stefan Kopp
Social Cognitive Systems
Bielefeld University

1 Introduction

Within the landscape of linguistic capabilities that have been studied and analyzed in Large Language Models (LLMs), a considerable amount of research has focused on phenomena on the level of morphology and syntax (Marvin and Linzen, 2018; Hu et al., 2020). Here, the community seems to have agreed on benchmarks and phenomena that an LLM should be capable of (e.g. agreement phenomena (Warstadt et al., 2020)). Various studies show that LLMs can handle a rich and diverse set of such phenomena (Chang and Bergen, 2023). Recent inquiries have expanded to investigate the proficiency of LLMs in pragmatic discourse processing (Ruis et al., 2022; Hu et al., 2023a; Sieker et al., 2023).

Pragmatic phenomena are often utilized when arguing for or against the reasoning capabilities of LLMs, which are a requirement for grounding in dialog. However, research on pragmatic abilities in LLMs remains more scarce and less systematic. We argue that studying the pragmatic competencies of LLMs is particularly interesting as it bridges aspects of 'core-linguistic' knowledge with the communicative, functional, and contextual aspects of grounding and is still actively discussed in current research (Mahowald et al., 2024). What does it mean when models can infer mental states while struggling with implicit meaning (Chang and Bergen, 2023)? Why do Language Models tussle, especially with phenomena that break language rules, such as humor, irony, and conversational maxims (Hu et al., 2023a)?

To address these questions and categorize findings effectively, capabilities related to pragmatics and grounding must be mapped out clearly and defined in relation to one another. In this work, we want to give an overview on which pragmatic abilities have been tested in LLMs so far and how these tests have been carried out. To do this, we first discuss the scope of the field of pragmatics and suggest a subdivision into *discourse pragmatics* and *interactional pragmatics*. We give a nonexhaustive overview of the phenomena of those two subdomains and the methods traditionally used to analyze them. We subsequently consider the resulting heterogeneous set of phenomena and methods as a starting point for our survey of work on discourse pragmatics and interactional pragmatics in the context of LLMs.

2 Pragmatics in Linguistics

Unlike other linguistic fields, such as syntax or phonetics, which focus on more structured and formal aspects of language, pragmatics encompasses a more heterogeneous set of phenomena that are often less systematic

and more context-dependent (Ariel, 2010). Negative definitions like the investigation of meaning distinct from pure semantics (Cummings, 2013) are fuzzy, and therefore, pragmatics is sometimes even referred to as the garbage can of linguistics (Bar-Hillel, 1971). Cummings (2013) contends that defining pragmatics as the study of how context affects meaning or as language usage analysis is overly broad. Instead, she proposes to define pragmatics as all intentionally expressed meanings that go beyond what is literally said. However, numerous endeavours have been made to establish clearer definitions or categorizations within the field. The Stanford Encyclopedia of Philosophy article on pragmatics, for example, distinguishes between 'classical' and 'contemporary' pragmatics, with classical pragmatics further divided into 'near-side' and 'far-side' (Korta and Perry, 2020). Near-side pragmatics focuses on explicit content, while far-side pragmatics explores implications beyond literal meanings. Contemporary pragmatics, on the other hand, includes works like Sperber & Wilson's relevance theory. Within these categories, Korta and Perry (2020) cover several pragmatic phenomena like ambiguity and implicatures. Yet, notably, grounding-relevant phenomena such as turn-taking or repair are overlooked despite being clearly pragmatic in nature.

Nevertheless, or precisely because of the diversity in the set of pragmatic phenomena, subcategorization is needed. We propose to cluster them into two main categories: *discourse pragmatics* and *interactional pragmatics*. While the *discourse pragmatics* describe formal reasoning processes, including phenomena such as presupposition, implicatures and figurative speech (i.e., aspects of pragmatics that were considered in Korta and Perry's article and could be described as near-side pragmatics), the *interactional pragmatics* address conversational reasoning phenomena, such as politeness, turn taking or repair (which could be designated as far-side pragmatics). Discourse pragmatics is often addressed in classical pragmatics and Natural Language Processing. The phenomena are mostly connected to text coherence. They can be found in a dialog but do not require direct interaction. These phenomena have been in the center of attention for decades. Often, testing instruments – drawing from the field of psychology or psycholinguistics (Ettinger, 2020; Sieker and Zarrieß, 2023) – are established. Additionally, theories from discourse pragmatics provide frameworks to describe these phenomena (Frank and Goodman, 2012; Degen, 2023).

Besides, there is a field of pragmatics that we refer to as interactional pragmatics. Here, the focus is rather on the interlocutors' interplay. A lot of research has been done on conversation analysis (Sacks et al., 1978; Atkinson and Heritage, 1984) or politeness theory (Brown and Levinson, 1987; Goffman, 1955; Leech, 2014). Conversation analysis utilises a strictly qualitative methodology borrowed from sociology and addresses the issue of "how we use language" at its core. The investigations on natural data focus on the organising principles that underlie human communication (Sacks et al., 1978; Atkinson and Heritage, 1984). In politeness theory, nuances of spoken language are emphasized(Brown and Levinson, 1987). Research in computer science and computational linguistics often addresses similar questions from the perspective of human-robot interaction (HRI). Kumar et al. (2022) reveal the positive impact politeness has on the enjoyment, satisfaction and trust participants perceive in an interaction with a robot. And Skantze (2021) give an overview of research on turn-taking behavior in HRI. Also, further interactional phenomena such as adaptation (Robrecht et al., 2023; Axelsson and Skantze, 2023; Stange, 2022) or grounding (Jung, 2017) have been subject to manifold approaches and studies in the field.

3 Approaches to pragmatics in LLMs

There are various examples of research that tests discourse pragmatic reasoning capabilities in language models. Ruis et al. (2022) investigate the extent to which LLMs such as OPT, T5 or GPT-4 may understand conversational implicatures. Inscale and between-scale scalar inferences in BERT are tested by comparing the

model's abilities to the human performance by Hu et al. (2023b). Carenini et al. (2023) take a look at the understanding GPT2 has of metaphors, explaining their results using the Rational Speech Act theory. Hu et al. (2023a) test seven discourse pragmatic phenomena (including maxims, metaphor, and coherence) in different versions of GPT-2, GPT-3 and T5. Moreover, the outcomes appear less promising when examining the study of interactional pragmatics in LLMs, the pragmatic category which covers most of the grounding-related phenomena. As this field of pragmatics is not as settled and the phenomena are harder to analyze due to their close connection to interaction, spoken language and spontaneous adaptation, there is a lack of instruments and measurements. Milicka et al.˘ (2024) show that GPT-3 and GPT-4 are able to decrease their cognitive abilities to simulate other personas. Also Wilf et al. (2023) test the perspective-taking abilities of GPT-3, GPT4, and Llama2, using chain-of-thought prompting. Nevertheless, most research connected to interactional pragmatics focuses on Theory of Mind or related theories (Gandhi et al., 2023; Wilf et al., 2023). It remains questionable whether these phenomena should be considered part of interactional pragmatics or not.

4 Contribution

We argue that there is a need for a more precise definition of pragmatic capabilities in research that studies the communicative behavior of LLMs. As a first step, we propose to distinguish discourse and interactional pragmatic abilities, for which we will discuss classification criteria and borderline cases. Further, we summarize which pragmatic phenomena have been tested in LLMs, how they are related to grounding, which methodology has been used, and which models have been considered.

5 Acknowledgments

Amelie Robrecht's and Stefan Kopp's research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824. Judith Sieker, Clara Lachenmaier and Sina Zarrieß received financial support from the project SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems: NW21-059A, funded by the program "Netzwerke 2021" of the Ministry of Culture and Science of the State of North Rhine-Westphalia, Germany.

References

Mira Ariel. 2010. *Defining Pragmatics*, 1 edition. Cambridge University Press.

J Maxwell Atkinson and John Heritage. 1984. *Structures of social action*. Cambridge University Press.

Agnes Axelsson and Gabriel Skantze. 2023. Do You Follow?: A Fully Automated System for Adaptive Robot Presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 102–111, Stockholm Sweden. ACM.

Yehoshua Bar-Hillel. 1971. Out of the pragmatic wastebasket. *Linguistic Inquiry*, 2(3):401–407.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Gaia Carenini, Louis Bodot, Walter Schaeken, Luca Bischetti, and Valentina Bambini. 2023. Large Language Models Behave (Almost) As Rational Speech Actors: Insights From Metaphor Understanding.

Tyler A. Chang and Benjamin K. Bergen. 2023. Language Model Behavior: A Comprehensive Survey. *arXiv preprint*. ArXiv:2303.11504 [cs].

Louise Cummings. 2013. *Pragmatics: A multidisciplinary perspective*. Routledge.

Judith Degen. 2023. The Rational Speech Act Framework. *Annual Review of Linguistics*, 9(1):519–540.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding Social Reasoning in Language Models with Language Models. *arXiv preprint*. ArXiv:2306.15448 [cs].

Erving Goffman. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3):213–231.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023a. A finegrained comparison of pragmatic language understanding in humans and language models.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023b. Expectations over Unspoken Alternatives Predict Pragmatic Inferences. *arXiv preprint*. ArXiv:2304.04758 [cs].

Malte F. Jung. 2017. Affective Grounding in Human-Robot Interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on HumanRobot Interaction*, pages 263–273, Vienna Austria. ACM.

Kepa Korta and John Perry. 2020. Pragmatics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2020 edition. Metaphysics Research Lab, Stanford University.

Shikhar Kumar, Eliran Itzhak, Yael Edan, Galit Nimrod, Vardit Sarne-Fleischmann, and Noam Tractinsky. 2022. Politeness in Human–Robot Interaction: A Multi-Experiment Study with Non-Humanoid Robots. *International Journal of Social Robotics*, 14(8):1805–1820.

Geoffrey N Leech. 2014. *The pragmatics of politeness*. Oxford University Press, USA.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, page S1364661324000275.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Ji˘rí Milicka, Anna Marklová, Klára VanSlambrouck,˘ Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ond˘rej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLOS ONE*, 19(3):e0298522.

Amelie Sophie Robrecht, Markus Rothgänger, and Stefan Kopp. 2023. A Study on the Benefits and Drawbacks of Adaptivity in AI-generated Explanations. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. The Goldilocks of Pragmatic Understanding: FineTuning Strategy Matters for Implicature Resolution by LLMs. Publisher: [object Object] Version Number: 2.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7– 55. Elsevier.

Judith Sieker, Oliver Bott, Torgrim Solstad, and Sina Zarrieß. 2023. Beyond the Bias: Unveiling the Quality of Implicit Causality Prompt Continuations in Language Models. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 206–220, Prague, Czechia. Association for Computational Linguistics.

Judith Sieker and Sina Zarrieß. 2023. When Your Language Model Cannot Even Do Determiners Right: Probing for Anti-Presuppositions and the Maximize
Presupposition! Principle. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 180–198, Singapore. Association for Computational Linguistics.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

Sonja Stange. 2022. *Tell Me Why (and What)! SelfExplanations for Autonomous Social Robot Behavior*. Ph.D. thesis, Universität Bielefeld. Artwork Size: 9237985 bytes Medium: application/pdf Publisher: Universität Bielefeld.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377– 392.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. *arXiv preprint*. ArXiv:2311.10227 [cs]