It Couldn't Help But Overhear: On the Limits of Modelling Meta-Communicative Grounding Acts with Supervised Learning

Brielen Madureira[1] **David Schlangen**[1,2]
[1]Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
{madureiralasota,david.schlangen}@uni-potsdam.de

Abstract

Active participation in a conversation is key to building common ground, since understanding is jointly tailored by producers and recipients. Overhearers are deprived of the privilege of performing grounding acts and can only conjecture about intended meanings. Still, data generation and annotation, modelling, training and evaluation of NLP dialogue models place reliance on the *overhearing paradigm*. How much of the underlying grounding processes are thereby forfeited? As we show, there is evidence pointing to the impossibility of properly modelling human meta-communicative acts with data-driven learning models. In this paper, we discuss this issue and provide a preliminary analysis on the variability of human decisions for requesting clarification. Most importantly, we wish to bring this topic back to the community's table, encouraging discussion on the consequences of having models designed to only "listen in".

1 Is Grounding "Supervisable"?

"What are you looking at?" asked Bob. "Magpies are building a nest outside!" Alice replied. If you were Bob, how would you continue that conversation? He could for instance say "Awesome!" or "I saw that". Whatever you say, it will probably differ from how he continued: "Building what?". The decision to request clarification depends on mutual understanding, which is contingent on *e.g.* the current situation, the familiarity between interlocutors and the previous utterances. Or, more formally, it depends on the clarification potential of these utterances (Ginzburg, 2012) and how they are assimilated into their *common ground* (Clark, 1996).

The one-to-many property of dialogue continuations is well-known in NLP (Zhao et al., 2017; Yeh et al., 2021; Towle and Zhou, 2022; Liu et al., 2023). There is a combinatorial explosion of possibilities for any interaction (Bates and Ayuso, 1991; Dingemanse and Enfield, 2023), and individual human behaviour may vary at each point. This variability is hard to measure, since arguably no two people will ever be in the exact same situation with the same conversation history to react to (Yeomans et al., 2023).

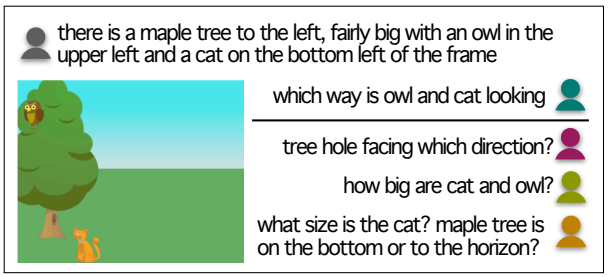

Figure 1: Variability of clarification requests produced by three overhearers in comparison to the original one, in an instance of the instruction-following CoDraw dialogue game (CC BY-NC 4.0), with cliparts from Zitnick and Parikh (2013).

Still, the prevailing end-to-end deep learning methods commonly rely on supervised learning (SL) from a sample of human behaviour instantiating the reaction of *a single human* at each observed context. Besides the issue of multiplicity of valid continuations, this paradigm faces another conceptual contention: dialogue models are trained to react upon a conversational history produced by someone else. In other words, they act as *overhearers*[1] of a dialogue in which they did not participate.

The suitability of data-driven methods and fixed corpora for modelling strategies and *conversational* grounding phenomena like Clarification Requests (CR) has been questioned (Schatzmann et al., 2005; Benotti and Blackburn, 2021b). Static datasets of human observations have empirically failed to provide enough information to define a human-like CR policy (Testoni and Fernández, 2024; Madureira and Schlangen, 2024). Moreover, chat-optimised

[1]We will use this term to also mean reading or seeing signs. Also called *observers* by Georgila et al. (2020).

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 149–158
September 18–20, 2024.

LLMs mostly do not engage in grounding acts and, when they do, it does not fully align with human behaviour (Kuhn et al., 2022; Deng et al., 2023; Shaikh et al., 2023). The latter is not necessarily a problem: one can use other effective methods when it comes to building applications. But the first is: grounding is essential for human communication, and lack of it can lead to undesired breakdowns (Benotti and Blackburn, 2021a).

Since modelling human dialogue strategies and the use of meta-communicative acts remains an unsolved problem, we hereby wish to re-open the discussion on the consequences of overhearing, focusing on two grounding devices: backchannels and interactive repair (Fusaroli et al., 2017).

2 Overhearers in a Conversation

As Clark (1996) defined it, in addition to speakers and addressees,[2] a conversation can have *side-participants*, who are part of it but at a given moment are neither of the those two, and *overhearers*, who are spectators without any rights or responsibilities, e.g. a silent audience or a minute-taker who lacks the opportunity to interfere (Peters, 2010). They are further divided into *bystanders*, if one is aware of their presence, or *eavesdroppers*, who listen secretly (or at a later time). There is evidence that the very process of understanding differs between addressees and overhearers: while interlocutors actively construct mutual understanding with each other, overhearers only passively consume the product of that process (Schober and Clark, 1989).

Speakers can design their utterances while taking different attitudes towards overhearers when they are aware of their presence (Clark, 1992; Liu et al., 2016), but covert overhearers are not acknowledged at all in the conversation, and can only conjecture about the intended meanings (Clark, 1992). Although the grounding acts they witness, like backchannels, and the availability of multiple perspectives may indeed aid their comprehension (Tolins and Fox Tree, 2016; Tree and Mayer, 2008), the original interaction was opportunistically produced to be understood against the original participants' common ground (Schober and Clark, 1989).

In their corpus analysis of common ground in multi-party interactions, Eshghi and Healey (2007) showed evidence that overhearers reach lower levels of understanding than ratified side participants, who in their turn are not very different from direct addressees, in what they call *collective states of understanding*. Related to that, Georgila et al. (2020) showed that observers and participants perceive interactions differently and the experiments by Fox Tree (1999) provided evidence that overhearers can more easily comprehend instructions while listening to dialogues than to monologues. Clark (1992) even argued that most psycholinguistic subjects are actually overhearers, so theories of language processing may actually be theories of overhearing, due to their lack of interactivity.

Separating addressees from side participants and accommodating overhearers are salient problems in research on multi-party dialogue (Jovanovic and op den Akker, 2004; Ginzburg and Fernández, 2005; Traum et al., 2018; Parisse et al., 2022; Ganesh et al., 2023).

3 Are NLP Models Only Listening In?

More than a decade ago, Rieser and Lemon (2011) already discussed the limitations of using supervised approaches for learning dialogue strategies. They flagged up three concerns: textual data does not contain the underlying uncertainty measures, instances are treated as local point-wise estimates (instead of the sequences they really are) and exploration of novel strategies is not possible, since the model has access only to the outcomes of the chosen dialogue trajectory originally perpetrated by the humans. This reflects the (offline) *overhearering paradigm*: a person or agent interpreting a pre-existing conversation and deciding what to do if they were in the original participants' shoes.

In NLP, this paradigm is widely used in various modelling steps. Let us look closer at four main practices, which may have cascaded effects.

Data Collection Given the extra cost of coordinating the presence of more than one subject for generating dialogical data, especially in crowdsourcing campaigns, many strategies have been proposed to bypass that with overhearing. For instance, this happens when the data collection procedure is framed as a dialogue continuation task (Frommherz and Zarcone, 2021). To name a few related to grounding, we have Zhou et al. (2022) who extracted dialogue contexts from existing datasets and presented them in a two-stage approach for some workers to generate common ground inferences and, separately, others generated a continuation as a response. Variations of overhearing manifest in techniques to generate CRs or

[2] Or producers and recipients.

their responses (Aliannejadi et al., 2021; Gao et al., 2022; Addlesee and Eshghi, 2024) and are even embedded in data collection tools that allow dialogues to be constructed without persistent workers (Cascante-Bonilla et al., 2019).

Annotation and Analysis Corpus studies of interactive linguistic use can only be performed from an overhearer perspective, without full evidence of what participants intended and understood or the reasons for their decisions (Brennan, 2000; Brennan et al., 2005). This is particularly challenging for research on common ground. For instance, Rodríguez and Schlangen (2004) and Schlöder and Fernández (2015) were confronted with the limitations of overhearers having only indirect access to the intentions of interlocutors when annotating CRs, partly remediating that by making a long dialogue context available. Niekrasz and Moore (2010) annotated references to conversation participants, joint actions that also serve to build common ground, emphasising that annotators were overhearers instructed to judge the speaker's intended purpose. Other annotations of grounding acts and common ground states had to rely on overhearers (Markowska et al., 2023; Zhang et al., 2023; Mohapatra et al., 2024).

Modelling Prototypical data-driven models trained with supervised learning to *process* dialogue, and possibly continue it, can, by design, be regarded as overhearers. This fact was made clear, for instance, in the CR model by Schlangen (2004). Traum (2017) differentiated between the perspective of an observer in *dialogue modelling* and of a participant in *dialogue management*, stating that the main difference lies in the decision-making process of the latter, although some specific applications also exist for the first.

Evaluation In human evaluation, overhearer experiments (Whittaker and Walker, 2005) are very common, even though it limits the judgements and measurements to user's *perceptions* of the dialogue (rather than the actual behaviour) (Whittaker and Walker, 2005; Foster and White, 2005; Moore, 2011) and restricts assessment of metrics like effectiveness and efficiency (Paksima et al., 2009). It has historically been a ubiquitous approach due to advantages like having control on one aspect of the evaluation while avoiding navigational and timing aspects of real interactions (Villalba et al., 2017), avoiding interference from ASR and other technical problems (Buß et al., 2010), allowing the collection of feedback about alternative system responses (Walker et al., 2004) and avoiding natural language interpretation problems (Wärnestal et al., 2007). Demberg et al. (2011) contrasted text overhearers with speech overhearers, pointing out that reading dialogues is artificially simplified, since participants can go back to difficult portions and choose the pace, and the two setups may also impact how evaluators rate the system. The available context may also have to be adjusted (Spanger et al., 2010). Cercas Curry and Rieser (2019) explicitly addressed the limitations of evaluation by overhearing and advocated for interaction with users. For a recent overview of works that use similar forms of *static* evaluation, see (Finch and Choi, 2020).

As we have seen, the overhearing paradigm (fairly silently) permeates fundamental phases of dialogue modelling. The choice of this paradigm used to be a salient concept, with authors showing awareness of its limitations when it was employed. Kousidis and Schlangen (2015) even modelled a ratified side participant and had evaluators "overhear the overhearer". In recent publications, however, it is often taken for granted, as if it was the only natural way to go. What can be the consequences when it comes to cognitive models of conversational grounding?

4 Variability in Human Grounding Acts

As humans speak, they can provide positive and negative evidence of mutual understanding (Clark and Brennan, 1991; Roque and Traum, 2008), but modelling their timing and decision-making is challenging. Traum (2017) claimed that "it can be very difficult to efficiently capture regularities in behavioral patterns that lead to similar, but not identical structures". In connection to that, people may take various paths in similar conversational situations (Bates and Ayuso, 1991). It is thus an open question how far data-driven supervised learning can get given the inherent variability of explicit (not to mention the latent) collateral signs of grounding.

Backchannels, a positive evidence of grounding, were demonstrated to involve individual variability, and even idiosyncrasy, possibly due to personality, gender or randomness (Huang and Gratch, 2012; Blomsma et al., 2024). Although those works showed some regularity in their timing, the SotA for the backchannel prediction task is not very high (.66 weighted F1) (Liermann et al., 2023).

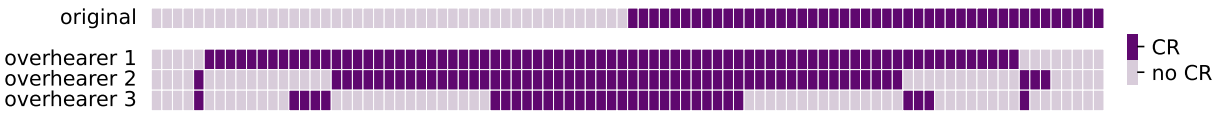

Figure 2: Variability in the decision of when to request clarification, comparing the decision of the original player with those of three overhearer annotators over 90 instances (horizontal axis) of the CoDraw game. Each cell is a data point and columns correspond to decisions on the same instance.

Findings on the variability of human decisions to initiate a CR, a negative sign of grounding, are still sparse. Stoyanchev et al. (2013) measured an absolute agreement of 39% among three annotators for *scripted* dialogues with missing ASR information. As another reference, Shaikh et al. (2023) reported a Cohen's κ of 48.45 for clarification in emotional support conversations, which, they claimed, may even be inflated. The task of deciding when to request clarification in collaborative instruction following is under active investigation, but models' performance is still suboptimal (Shi et al., 2022; Li et al., 2023; Madureira and Schlangen, 2023; Mohanty et al., 2023). Recent works on the multimodal CoDraw dialogue game (Kim et al., 2019) argued that this may be due to the variability in human decisions and the limitations of using supervised learning (Testoni and Fernández, 2024; Madureira and Schlangen, 2024).

5 A Brief Analysis of Regularity in CRs

In CoDraw, an instruction follower receives instructions to reconstruct a scene using cliparts (as in Figure 1). Only the instruction giver sees the target scene. Madureira and Schlangen (2023) identified all CRs (around 11% of the instruction follower's utterances) and defined the task of deciding when to request clarification, where models reached only up to .41 binary F1. What is missing as evidence for the claim that data-driven models cannot fully succeed in learning a "when policy" from human data is the actual human performance on this NLP task, i.e. what *overhearers* predict.

For an initial analysis, we collected a convenience sample with three annotators performing a similar task as the trained models: given a dialogue history and the current state of the reconstructed scene, decide which actions to take and, if needed, request clarification (details in Appendix). We randomly selected a sample with 90 instances; in half of them, the original player had produced a CR.

The average binary F1 of overhearers with respect to the original decision was .51, not much above what SotA models achieve. But the proportion of CRs widely ranged from 36 to 85%. Among the three annotators, the Krippendorff's α was 0.10 and the mean pairwise Cohen's κ was 0.18. That is already low, but if we consider the original decision as a fourth annotator, measures are even lower: α was 0.02 and κ was 0.06. This indicates that there was slightly more agreement among overhearers than among addresses and overhearers, but in general there was little agreement on deciding when a CR should be realised. Figure 2 presents the main binary decision (whether to request clarification or not) for each of the 90 annotation instances, serving to provide a visual overview of such variability.

In terms of surface forms, the average BLEU score was 0.11 (std= 0.10) using the original CR as a source and the produced utterances as a reference. The mean cosine similarity between the embedding of the produced and the original CRs was 0.38, 0.29 and 0.36 for the three overhearers. Figure 1 shows an example of how diverse the produced clarifications can be, both in form and in content, even when all subjects made the same decision to clarify at a given point.

These are preliminary insights from a pilot study. Further standardised experiments with a larger sample must be conducted. Still, the results are already useful to strengthen the argument that, like backchannelling, human CR decisions lack regularity and overhearers have a much harder task trying to interpret and act upon someone else's grounding acts. Decisions depend on how interlocutors distribute grounding costs, as per the principle of least collaborative effort (Clark and Brennan, 1991). Besides, there might be adaptive behaviours that models are not capturing (Dideriksen et al., 2023).

To continue this investigation, we propose distinguishing between the clarification potential (Ginzburg, 2012; Benotti, 2009) and the clarification need. The first is a larger set of possibilities for clarification of a given utterance, while the latter refers to the decision of whether and what to clarify taken by a given individual operating with that ut-

terance and identifying something worth clarifying. Or, in other words, the clarification need, which is *in the agent*, refers to what was asked among all that could be asked. It is challenging to design experiments that can capture the clarification need among individuals, in particular due to the difficulty in replicating a given dialogue context for different subjects if they are not acting as overhearers. A possible next step is to turn the CR decision into an acceptability task, regarding it as a contrast. For each instance, the annotator would see a set of CRs. The actual CR observed in the data should ideally be accepted, but possibly others too. If the original CR falls into the empirical potential, there should be a plausible need for it at that point. Such experiment could also aim to measure uncertainty at each turn.

6 Discussion

Mutual understanding is crafted by "interacting minds" (Dingemanse et al., 2023). In dialogue, "interlocutors share or synchronise aspects of their private mental states and act together in the world" (Brennan et al., 2010). On the other hand, we have shown that the current NLP methodology mostly limits us to learning how *overhearers* predict discourse representations without the actual joint decision making facet, due to the way that data is produced and annotated, the assumptions behind training mechanisms and the evaluation protocols, each adding a layer of overhearing.

What can be a better setup to learn human dialogue behaviour, realising it as a truly interactive process? One needs to move on from one-off supervised learning to sequential models that not only *understand* dialogues but also *participate* in them.[3] Reinforcement learning provides that framing with a fully accessible and explorable environment (Rieser and Lemon, 2011), but somewhat circularly requires a good simulation of an user or interlocutor (Schatzmann et al., 2005; Georgila et al., 2006; Li et al., 2020). Although LLMs can serve as speaker simulators, so far they cannot fully model all dialogue phenomena. Another possibility are hybrid combinations of supervised and reinforcement learning (Henderson et al., 2008), as well as further improvements in techniques like RLHF, PPO and DPO. But independently of the learning regime, data-driven approaches, which rely on extracting latent patterns and regularities in a corpus, stumble upon the individual variability of some dialogue phenomena, so that tasks may be ill-defined in datasets. Besides, although transcribed dialogue contain clues about the decision making during a conversation, they provide only limited evidence of what participants understood or intended, or their internal states (Brennan et al., 2005), which are pertinent for modelling some dialogue decisions and meta-communicative acts.

Indeed, interfaces do not necessarily have to conform to human behaviour, as long as they can sustain *graceful interaction* (Hayes, 1980). But from a cognitive perspective, the current NLP resources do not seem to satisfactorily meet our needs for modelling grounding mechanisms. To study the human mind, do we want cognitive models of how meaning and common ground are constructed or only of how they can be reverse engineered from someone else's interactions?

To conclude With this argumentative paper, we wish to encourage more studies on the variability of human grounding acts and its impact in modelling human dialogue strategies. Besides, we advocate making the overhearing paradigm explicit whenever it is used in future publications and discussing how it can have influenced reported findings.

[3] See (Min et al., 2022) for a related discussion on the limitations of imitation learning and behaviour cloning for embodied agents. See also (Ortega et al., 2021) for a discussion on supervised learning and the sequential aspect of an interaction.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. We are also thankful to the three annotators who took part on the pilot study.

References

Angus Addlesee and Arash Eshghi. 2024. You have interrupted me again!: making voice assistants more dementia-friendly with incremental clarification. *Frontiers in Dementia*, 3:1343052.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Madeleine Bates and Damaris Ayuso. 1991. A proposal for incremental dialogue evaluation. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Luciana Benotti. 2009. Clarification potential of instructions. In *Proceedings of the SIGDIAL 2009 Conference*, pages 196–205, London, UK. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2021a. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2021b. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4065–4077, Online. Association for Computational Linguistics.

Peter Blomsma, Julija Vaitonyté, Gabriel Skantze, and Marc Swerts. 2024. Backchannel behavior is idiosyncratic. *Language and Cognition*, page 1–24.

Susan E. Brennan. 2000. Invited talk: Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 1–11, Hong Kong. Association for Computational Linguistics.

Susan E Brennan, Alexia Galati, and Anna K Kuhlen. 2010. Two minds, one dialog: Coordinating speaking and understanding. In *Psychology of learning and motivation*, volume 53, pages 301–344. Elsevier.

Susan E Brennan et al. 2005. How conversation is shaped by visual and spoken evidence. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, pages 95–129.

Okko Buß, Timo Baumann, and David Schlangen. 2010. Collaborating on utterances with a spoken dialogue system using an ISU-based approach to incremental dialogue management. In *Proceedings of the SIGDIAL 2010 Conference*, pages 233–236, Tokyo, Japan. Association for Computational Linguistics.

Paola Cascante-Bonilla, Xuwang Yin, Vicente Ordonez, and Song Feng. 2019. Chat-crowd: A dialog-based platform for visual layout composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 138–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Amanda Cercas Curry and Verena Rieser. 2019. A crowd-based evaluation of abuse response strategies in conversational agents. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 361–366, Stockholm, Sweden. Association for Computational Linguistics.

Herbert H Clark. 1992. *Arenas of language use*. University of Chicago Press.

Herbert H. Clark. 1996. *Common ground*, page 92–122. "Using" Linguistic Books. Cambridge University Press.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition.*, pages 127–149. American Psychological Association.

Vera Demberg, Andi Winterboer, and Johanna D. Moore. 2011. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Christina Dideriksen, Morten H Christiansen, Kristian Tylén, Mark Dingemanse, and Riccardo Fusaroli. 2023. Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General*, 152(3):864.

Mark Dingemanse and NJ Enfield. 2023. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*.

Mark Dingemanse, Andreas Liesenfeld, Marlou Rasenberg, Saul Albert, Felix K Ameka, Abeba Birhane, Dimitris Bolis, Justine Cassell, Rebecca Clift, Elena Cuffari, et al. 2023. Beyond single-mindedness: A figure-ground reversal for the cognitive sciences. *Cognitive science*, 47(1):e13230.

Arash Eshghi and Patrick G.T. Healey. 2007. Collective states of understanding. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 2–9, Antwerp, Belgium. Association for Computational Linguistics.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Mary Ellen Foster and Michael White. 2005. Assessing the impact of adaptive generation in the comic multimodal dialogue system. In *Proceedings of the IJCAI 2005 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 24–31.

Jean E Fox Tree. 1999. Listening in on monologues and dialogues. *Discourse processes*, 27(1):35–53.

Yannick Frommherz and Alessandra Zarcone. 2021. Crowdsourcing ecologically-valid dialogue data for german. *Frontiers in computer science*, 3:686050.

Riccardo Fusaroli, Kristian Tylén, Katrine Garly, Jakob Steensig, Morten H Christiansen, and Mark Dingemanse. 2017. Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In *the 39th Annual Conference of the Cognitive Science Society (CogSci 2017)*, pages 2055–2060. Cognitive Science Society.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2023. A survey of challenges and methods in the computational modeling of multi-party dialog. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 140–154, Toronto, Canada. Association for Computational Linguistics.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056.

Kallirroi Georgila, Carla Gordon, Volodymyr Yanov, and David Traum. 2020. Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 726–734, Marseille, France. European Language Resources Association.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: learning and evaluation. In *Ninth International Conference on Spoken Language Processing*.

Jonathan Ginzburg. 2012. *The interactive stance: Meaning for conversation*. Oxford University Press.

Jonathan Ginzburg and Raquel Fernández. 2005. Scaling up from dialogue to multilogue: Some principles and benchmarks. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 231–238, Ann Arbor, Michigan. Association for Computational Linguistics.

Phil Hayes. 1980. Expanding the horizons of natural language interfaces. In *18th Annual Meeting of the Association for Computational Linguistics*, pages 71–74, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2008. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4):487–511.

Lixing Huang and Jonathan Gratch. 2012. Crowdsourcing backchannel feedback: understanding the individual variability from the crowds. In *Feedback behaviors in dialog*.

Natasa Jovanovic and Rieks op den Akker. 2004. Towards automatic addressee identification in multiparty dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 89–92, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

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 goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy. Association for Computational Linguistics.

Spyridon Kousidis and David Schlangen. 2015. The power of a glance: Evaluating embodiment and turn-tracking strategies of an active robotic overhearer. In *Proceedings of AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Haau-Sing (Xiaocheng) Li, Mohsen Mesgar, André Martins, and Iryna Gurevych. 2023. Python code generation by asking clarification questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14287–14306, Toronto, Canada. Association for Computational Linguistics.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2020. Rethinking supervised learning and reinforcement learning in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3537–3546, Online. Association for Computational Linguistics.

Wencke Liermann, Yo-Han Park, Yong-Seok Choi, and Kong Lee. 2023. Dialogue act-aided backchannel prediction using multi-task learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15073–15079, Singapore. Association for Computational Linguistics.

Kris Liu, Jean Fox Tree, and Marilyn Walker. 2016. Coordinating communication in the wild: The artwalk dialogue corpus of pedestrian navigation and mobile referential communication. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3159–3166, Portorož, Slovenia. European Language Resources Association (ELRA).

Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023. PVGRU: Generating diverse and relevant dialogue responses via pseudo-variational mechanism. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 3295–3310, Toronto, Canada. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2023. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the CoDraw dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2303–2319, Dubrovnik, Croatia. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2024. Taking action towards graceful interaction: The effects of performing actions on modelling policies for instruction clarification requests. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 1–21, Malta. Association for Computational Linguistics.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. 2023. Finding common ground: Annotating and predicting common ground in spoken conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore. Association for Computational Linguistics.

So Yeon Min, Hao Zhu, Ruslan Salakhutdinov, and Yonatan Bisk. 2022. Don't copy the teacher: Data and model challenges in embodied dialogue. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9361–9368, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shrestha Mohanty, Negar Arabzadeh, Julia Kiseleva, Artem Zholus, Milagro Teruel, Ahmed Awadallah, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. 2023. Transforming human-centered ai collaboration: Redefining embodied agents capabilities through interactive grounded language instructions. *arXiv preprint arXiv:2305.10783*.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In *Proceedings of LREC-COLING 2024*.

Johanna D. Moore. 2011. Language generation for spoken dialogue systems [invited talk]. In *Proceedings of the 13th European Workshop on Natural Language Generation*, page 132, Nancy, France. Association for Computational Linguistics.

John Niekrasz and Johanna D. Moore. 2010. Annotating participant reference in English spoken conversation. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 256–264, Uppsala, Sweden. Association for Computational Linguistics.

Pedro A Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, et al. 2021. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*.

Taghi Paksima, Kallirroi Georgila, and Johanna Moore. 2009. Evaluating the effectiveness of information presentation in a full end-to-end dialogue system. In *Proceedings of the SIGDIAL 2009 Conference*, pages 1–10, London, UK. Association for Computational Linguistics.

Christophe Parisse, Marion Blondel, Stéphanie Caët, Claire Danet, Coralie Vincent, and Aliyah Morgenstern. 2022. Multidimensional coding of multimodal languaging in multi-party settings. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2781–2787, Marseille, France. European Language Resources Association.

Stanley Peters. 2010. Listening in. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 31–31, Tohoku University, Sendai, Japan. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media.

Kepa Joseba Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*.

Antonio Roque and David Traum. 2008. Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54–63, Columbus, Ohio. Association for Computational Linguistics.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 45–54, Lisbon, Portugal. Special Interest Group on Discourse and Dialogue (SIGdial).

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 136–143, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Julian J. Schlöder and Raquel Fernández. 2015. Clarifying intentions in dialogue: A corpus study. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 46–51, London, UK. Association for Computational Linguistics.

Michael F Schober and Herbert H Clark. 1989. Understanding by addressees and overhearers. *Cognitive psychology*, 21(2):211–232.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2023. Grounding or guesswork? large language models are presumptive grounders. *arXiv preprint arXiv:2311.09144*.

Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to execute actions or ask clarification questions. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2060–2070, Seattle, United States. Association for Computational Linguistics.

Philipp Spanger, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. 2010. Towards an extrinsic evaluation of referring expressions in situated dialogs. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2013. Modelling human clarification strategies. In *Proceedings of the SIGDIAL 2013 Conference*, pages 137–141, Metz, France. Association for Computational Linguistics.

Alberto Testoni and Raquel Fernández. 2024. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–275, St. Julian's, Malta. Association for Computational Linguistics.

Jackson Tolins and Jean E Fox Tree. 2016. Overhearers use addressee backchannels in dialog comprehension. *Cognitive science*, 40(6):1412–1434.

Benjamin Towle and Ke Zhou. 2022. Learn what is possible, then choose what is best: Disentangling one-to-many relations in language through text-based games. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4955–4965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Traum. 2017. Computational approaches to dialogue. *The Routledge Handbook of Language and Dialogue*, 1:143–161.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory Hayes, and Susan Hill. 2018. Dialogue structure annotation for multi-floor interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jean E Fox Tree and Sarah A Mayer. 2008. Overhearing single and multiple perspectives. *Discourse Processes*, 45(2):160–179.

Martín Villalba, Christoph Teichmann, and Alexander Koller. 2017. Generating contrastive referring expressions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 678–687, Vancouver, Canada. Association for Computational Linguistics.

Marilyn A Walker, Stephen J Whittaker, Amanda Stent, Preetam Maloor, Johanna Moore, Michael Johnston, and Gunaranjan Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.

Pontus Wärnestal, Lars Degerstedt, and Arne Jönsson. 2007. Emergent conversational recommendations: A dialogue behavior approach. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 63–66, Antwerp, Belgium. Association for Computational Linguistics.

Steve Whittaker and Marilyn Walker. 2005. Evaluating dialogue strategies in multimodal dialogue systems. *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, pages 247–268.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Michael Yeomans, F Katelynn Boland, Hanne K Collins, Nicole Abi-Esber, and Alison Wood Brooks. 2023. A practical guide to conversation research: How to study what people say to each other. *Advances in Methods and Practices in Psychological Science*, 6(4):25152459231183919.

Xuanming Zhang, Rahul Divekar, Rutuja Ubale, and Zhou Yu. 2023. GrounDialog: A dataset for repair and grounding in task-oriented spoken dialogues for language learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 300–314, Toronto, Canada. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016.

A Appendix

Annotation Task The decisions from the overhearer perspective were performed by 3 annotators. Two of them are student assistants employed in our lab and one is a volunteer acquainted with the first author. A simple GUI interface showed the dialogue history (from 1 to 3 turns) up to the last instruction giver instruction, the current state of the reconstructed scene and the gallery of available cliparts. They could select up to 4 high level, discrete actions (add, move, resize, flip, delete) and the corresponding cliparts from dropdown lists. Besides, they could type a clarification request to continue the dialogue if they wished (otherwise, the next utterance field should be left blank). In future studies, a full interface similar to the original game should be used, i.e. giving the opportunity for cliparts to be moved around and edited in the scene. Here, the selection of actions was just used to enforce that the overhearers reflected on the pertinent actions while deciding whether to request clarification. Note that the step of action taking makes annotators more privileged than plain overhearers that just process the dialogue, but it better approximates the decision of the iCR-Action-Taker models in Madureira and Schlangen (2024). In this case, they are overhearers of the dialogue context, but try to minimally act as a player doing the next step. The results work as an upper bound for plain overhearers.

Additional Details The inter-annotator agreement metrics were computed with `nltk` using `chencherry.method3` for smoothing. The sentence embeddings for the CR utterances were computed with model `sentence-transformers/all-MiniLM-L6-v2` from SentenceTransformers (Reimers and Gurevych, 2019).