Characterizing Language Use in Collaborative Situated Games

Nicholas Tomlin[*1,2], **Naitian Zhou**[*1], **Eve Fleisig**[1], **Liangyuan (Circle) Chen**[1], **Téa Wright**[1], **Lauren Vinh**[**1], **Laura X. Ma**[**1], **Seun Eisape**[1], **Ellie French**[1], **Tingting Du**[1], **Tianjiao Zhang**[1], **Alexander Koller**[3], **Alane Suhr**[1]
[1]UC Berkeley [2]NYU [3]Saarland University
[*]*Equal contribution*

Abstract

Cooperative video games, where multiple participants must coordinate by communicating and reasoning under uncertainty in complex environments, yield a rich source of language data. We collect the Portal Dialogue Corpus: a corpus of 11.5 hours of spoken human dialogue in the co-op mode of the popular Portal 2 virtual puzzle game, comprising 24.5k total utterances. We analyze player language and behavior, identifying a number of linguistic phenomena that rarely appear in most existing chitchat or task-oriented dialogue corpora, including complex spatial reference, clarification and repair, and adhoc convention formation. To support future analyses of language use in complex, situated, collaborative problem-solving scenarios, we publicly release the corpus, which comprises player videos, audio, transcripts, game state data, and both manual and automatic annotations of language data.

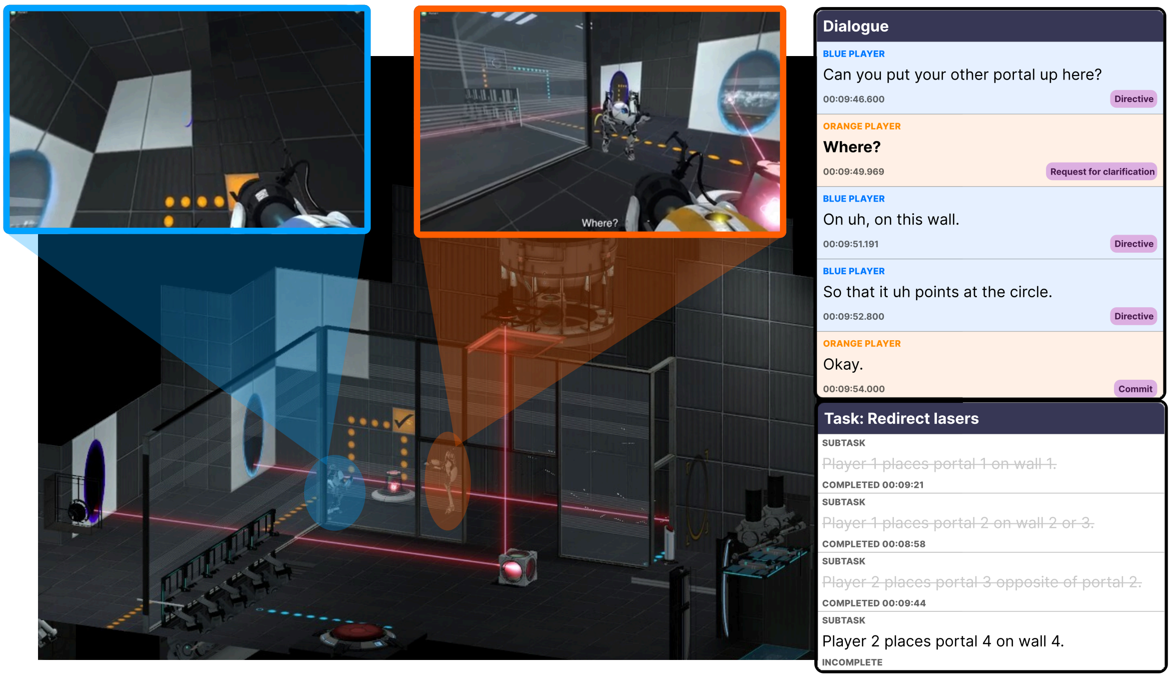
1 Introduction

Language is the primary medium through which humans coordinate, share information, and achieve common goals. To efficiently coordinate in novel contexts, we adapt language to fit our communicative needs under constraints: in jointly embodied environments, we use spatial references whose meanings are dependent on our relative perspectives of the shared scene; in settings with salient novel referents that lack canonical labels, we develop arbitrary but stable and concise conceptual pacts for efficient reference as an interaction proceeds. Recent work on embodied conversational agents envisions systems that assist or collaborate with human users in situated interactions via language use in context, but such models struggle to adapt their language use to new interaction partners and scenarios. To build agents capable of efficient, dynamic, and natural interaction in situated environments, we must better understand the linguistic behaviors that support successful coordination between humans. This calls for resources that capture, identify, and analyze such behaviors.

Most existing studies on language-based interaction have been performed on relatively simple game-like environments, where the space of expressible meanings does not change over time and typically reflects a very small subset of real-world meanings; or on open-domain conversational chats, which does not support the fine-grained control of a situated environment and incentive design, nor the capture of the full interaction and its context. As a result, the set of language phenomena that are used and have been studied in these environments reflects a limited range of linguistic diversity. Focusing computational studies of language, including the development of language technologies, exclusively on text-based chat or simple situated environments results in neglect of the breadth of possible linguistic devices and the dynamics of their formation and adaptation in interaction.

Multi-agent virtual worlds, like cooperative video games, provide a unique opportunity to study the relationship between the incentives, constraints, and affordances of an interaction's scenario and the interaction's dynamics, including language adaptation and action coordination. Indeed, game development optimizes player enjoyment via incentive and environment design; games typically come with easily evaluable metrics; and, being virtual, they typically support capturing the entirety of an interaction, rather than only part of it. Moreover, virtual worlds support a high degree of novelty and pretense for their players (Nguyen et al., 2020), including fantastical scenarios that we can nonetheless learn to reason and communicate about, such as novel technologies and impossible physics.

We study the dynamics of language-based interaction through the introduction of the Portal



Dialogue

BLUE PLAYER
Can you put your other portal up here?
00:09:46.600 Directive

ORANGE PLAYER
Where?
00:09:49.969 Request for clarification

BLUE PLAYER
On uh, on this wall.
00:09:51.191 Directive

BLUE PLAYER
So that it uh points at the circle.
00:09:52.800 Directive

ORANGE PLAYER
Okay.
00:09:54.000 Commit

Task: Redirect lasers

SUBTASK
~~Player 1 places portal 1 on wall 1.~~
COMPLETED 00:09:21

SUBTASK
~~Player 1 places portal 2 on wall 2 or 3.~~
COMPLETED 00:08:58

SUBTASK
~~Player 2 places portal 3 opposite of portal 2.~~
COMPLETED 00:09:44

SUBTASK
Player 2 places portal 4 on wall 4.
INCOMPLETE

Figure 1: Portal 2 is a first-person 3D game where players must solve puzzles through the use of portals. **(Left)** Snapshot of the game environment, including a third-person view (bottom, not available to players) and popouts showing each player's perspective. **(Right)** Transcripts of the player dialogue in this scene, each paired with dialogue act annotations (top) and a list of subtasks for this puzzle (bottom, not available to players). In this game snapshot, the blue player uses spatial language to give instructions without an explicit reference, triggering a clarification request and a repair process.

Dialogue Corpus: a new corpus of collaborative, situated, and task-oriented human dialogue. We recorded 11.5 hours of gameplay in the 3D first-person video game Portal 2, where two players collaborate to solve physical puzzles (Figure 1). Players start each puzzle with roughly the same capabilities and prior knowledge, but a high degree of uncertainty about the puzzle solution. Players must jointly explore each puzzle, including learning new environment dynamics and affordances of novel objects, and construct and execute a solution to the puzzle. Critically, puzzle designs require that the players coordinate their actions with one another, for example by requiring simultaneous action in different parts of the same puzzle. They must therefore communicate, in language, to successfully coordinate under uncertainty.

The Portal Dialogue Corpus combines multiple streams of information, including video and audio recordings of the gameplay, timestamped and transcribed user utterances, and real-time information about the state of the world, including player and object locations and actions. Our dataset's combination of multimodal data, extensive annotation, and inclusion of rich interactions typically absent from conversational datasets makes it a uniquely rich resource for research on task-oriented dialogue, multimodal conversations, and understanding human interaction more broadly. We demonstrate through quantitative and qualitative analysis of the dataset that players exhibit a number of sophisticated language skills with which current artificial language systems struggle. For example, players agree on ad-hoc conventions to refer to novel in-game objects or action sequences and use ambiguous spatial references (Section 5.2). They resolve misunderstandings via generation of clarification requests and self-corrections, and they rely on multimodal grounding, including based on gaze and actions, to resolve ambiguities (Section 5.3). Finally, they use task-planning dialogue acts to coordinate their actions towards successful task completion (Section 5.4). The Portal Dialogue Corpus will be made freely available as a contribution to computational research in language and interaction.[1][2]

[1] berkeley-nlp.github.io/portal-dialogue-corpus/

[2] All videos are available at our YouTube channel @PortalDialogueCorpus. Due to our IRB policies, these publicly-

2 Related Work

Many existing spoken dialogue datasets comprise open-domain, non-situated conversational speech (Holliman et al., 1992; Burnard, 1995; Serban et al., 2018; Reece et al., 2023). In contrast, research on task-oriented dialogue uses settings where agents (either humans or models) must communicate in order to complete shared goals under differences in their knowledge (Allen et al., 1995; Allen and Ferguson, 2002; He et al., 2017; Budzianowski et al., 2018; Wei et al., 2018; Semantic Machines et al., 2020; Lin et al., 2024), or to successfully negotiate under conflicting goals (Lewis et al., 2017; He et al., 2018; Bakhtin et al., 2022).

In embodied dialogue, agents communicate in the context of partially-observable situated environments. Embodiment expands the space of possible asymmetries between agents: in settings with asymmetric action spaces (e.g., CerealBar, Suhr et al., 2019), collaboration often reduces to instruction-following, and participants need not negotiate their roles via conversation (Narayan-Chen et al., 2019); in contrast, tasks with symmetric action spaces where roles are not predefined (e.g., Cards, Potts, 2012) are more likely to result in mixed-initiative interactions (Ichikawa and Higashinaka, 2022). In our setting, players are not assigned roles ahead of time, which requires them to coordinate on how to divide up their shared goals. Like Portal 2, many grounded dialogue tasks are designed to incentivize communication by introducing asymmetry in the agents' observation space (e.g., OneCommon, Udagawa and Aizawa, 2019).

Other collaborative game settings, such as Hanabi (Bard et al., 2020) or Overcooked (Carroll et al., 2019; Strouse et al., 2021), have been used to study multi-agent coordination in absence of language-based communication, for example for building agents that model their interaction partners via theory-of-mind. Each of these domains elicits rich collaborative behavior, but communication is either explicitly limited by game rules or implicitly limited by the space of available meanings. While our setting highlights many similar aspects of collaboration, it focuses more on free-form linguistic communication in an embodied 3D environment, leading to richer reference and conversational grounding phenomena.

3 The Portal Dialogue Corpus

Portal is a first-person puzzle video game released by Valve Corporation in 2007. The key game mechanic is the portal gun: a device that shoots wormhole-like portals onto specific surfaces in the game; once a player has placed two portals, they can teleport between them by entering either portal. Puzzle design requires players use portals to reach and move objects to otherwise inaccessible locations. The game includes several additional mechanics, such as lasers, turrets (defensive robots), and movable bridges.

While Portal is a single-player game, the 2011 sequel Portal 2 includes a two-player cooperative mode called the Cooperative Testing Initiative. This cooperative mode consists of six chapters, each with six or more levels, in which pairs of players must work together to solve puzzles. Each level contains one or two puzzles. In the cooperative mode, shown in Figure 1, each player has their own portal gun, resulting in a maximum of four portals that can be placed simultaneously. Most puzzles require both players to place portals, and players can communicate with each other using voice chat. We refer to each player by the color of their respective portal guns: Blue and Orange.

We collected data of gameplay from the Cooperative Testing Initiative by 18 pairs of players. In total, we obtained 11h 25m of gameplay data from 36 participants, consisting of screen recordings from each player's perspectives, recordings of voice chat audio, and game engine demo files that include exact game state information at each timestep.[3] The game state includes players' positions, orientations, and objects' locations and velocities. In addition, we produce high-quality, manually-corrected transcripts for the game audio, annotations of dialogue acts, and indicators of when players completed various goals and subgoals within each level. This allows us to replay each game exactly as it took place during the study, for example to acquire third-person views of gameplay (as in Figure 1).

4 Data Collection

4.1 Setup

We recruited a total of 36 English-speaking participants via flyers, advertisements on online forums,

available videos do not contain audio, but do contain subtitles. Audio data will be distributed with permission to interested researchers.

[3] A timestep (tick) is how often the game state is updated, every 1/60th of a second.

and word-of-mouth.[4] Each pair of players interacted for up to 1 hour each. We began each session with a written pre-survey, asking participants about their previous experience with video games and the Portal series, as well as consent forms and instructions. Investigators then guided participants to one of two recording rooms, which were equipped with gaming laptops, as well as identical mice, keyboards, external monitors, and headphones. Participants played the game in separate rooms and communicated via voice call over Discord. We used Open Broadcaster Software (OBS) to collect audio and screen recordings. For each session, we also saved the game engine demo files that contain player and world state during gameplay, from which we can extract in-game events and player positions. Each pair was instructed to play the first chapter of the game, which consists of six individual levels. Participants who completed the entire first chapter were instructed to signal an investigator, who then reset the recordings and loaded the third chapter, which consists of an additional eight levels.[5] No participants progressed to the fourth chapter, and investigators ended each recording session at the end of the hour. Upon completion, each participant was compensated with a gift card.

4.2 Data Postprocessing

We used WhisperX (Bain et al., 2023) to generate utterance-level transcriptions from the audio. We then reestimated utterance time alignments using a Wav2Vec-based forced aligner, before manually correcting the utterance segmentation and format in Adobe Premiere. We adapted our transcription format from existing guidelines[6] in linguistics to preserve filler words, false starts, and incomplete utterances, as well as extralinguistic communicative expressions like laughter. We delineated utterance boundaries contextually using timing, intonation, and semantic information. We also anonymized any personal details, such as workplace or hometown, mentioned by participants in the audio recordings and transcripts.

Finally, we hired crowdworkers on Prolific to ensure time alignments of utterances and audio files were high-quality. Workers were paid based on estimated time of completion at a $15/hour rate. All workers passed an initial quality check and adjustments were spot-checked by project members.

To create videos for each level, we used Adobe Premiere to time-align the screen and audio recordings of each chapter within each pair of players before segmenting recordings into individual levels. Audio files were gain-adjusted to normalize the volume across all recordings.

Portal is powered by the Source video game engine, which we configured to generate demo files that include information about the game and player state. We developed a tool that parses demo files to extract, for each timestep (tick) in the game, player position, orientation, and viewpoint information. Demo file timesteps were manually time-aligned to the audio and video recordings.

4.3 Annotation

After collecting and post-processing interaction recordings, we generated several layers of metadata relating to dialogue acts and game progress.

Dialogue acts To support quantitative analysis of the relationship between language use and player behavior, we release per-utterance tag annotations associated with utterance form, content, and intent. We designed an annotation schema, adapted from DAMSL (Dialog Act Markup in Several Layers, Core and Allen, 1997), that covers five layers of dialogue acts:[7]

1. **Communicative status** categorizes whether a speaker's intent was successfully communicated or not.
2. **Information level** describes the type of information conveyed in an utterance.
3. **Uncertainty** captures the epistemic stance of the speaker in a given utterance.
4. **Utterance type** categorizes spoken dialogue based on its syntactic form.
5. **Discursive act** describes the function of utterances within a collaborative task-based dialogue. Utterances in Figure 1 (right) are shown with corresponding discursive acts.

We manually labeled a small subset (three levels; 176 utterances) with four annotators to evaluate both the inter-annotator agreement with our

[4] The study was reviewed and approved by our IRB.

[5] We skipped levels in Chapter Two, which we deemed more likely to cause motion sickness due to this chapter's focus on using portals to fall and fly quickly through the air.

[6] ldp-uchicago.github.io/docs/guides/transcription/sect_4.html

[7] Values and descriptions of the tags are available in Appendix A.

Cohen's κ	**Com.**	**Inf.**	**Unc.**	**Utt.**	**Dis.**
Text Only	0.68	0.61	0.39	0.72	0.58
A/V	0.66	0.62	0.39	0.72	0.56

Table 1: Pairwise average inter-annotator agreement for when annotators could only access the transcripts and when annotators also had access to the audio and video.

Cohen's κ	**Com.**	**Inf.**	**Unc.**	**Utt.**	**Dis.**
GPT-4o	0.48	0.44	0.30	0.52	0.28

Table 2: Average agreement between automated labeling and each human annotator.

schema. he annotators first assigned labels given only the text transcript to match the automated classification setting, then performed a second pass with access to the audio and video in order to assess agreement given maximal data. Inter-annotator agreements, calculated as the average pairwise Cohen's κ, for each layer are presented in Table 1. We find fair to substantial agreement (Landis and Koch, 1977) that is roughly in line with previous dialogue act annotations (Core and Allen, 1997). Low inter-annotator agreement demonstrates the inherent difficulty of the dialogue act annotation task. Somewhat surprisingly, the audio-visual annotations largely did not result in higher agreement compared to the text-only setting; this suggests that, instead of providing disambiguation, access to more modalities might allow for greater flexibility in interpretation.

To estimate tags for the entire corpus, we used GPT-4o to automatically label each transcribed utterance given its dialogue history, and used the manual annotations to evaluate the classification performance of GPT-4o (Table 2). To compare agreement levels, we calculate the average agreement between the automated labeling and each of the human annotators. This agreement is lower than human inter-annotator agreement, suggesting there is still room to improve model performance on this task (Ettinger et al., 2023).

Tasks and subtasks To enable analysis around coordination and joint planning, we identified tasks and subtasks necessary for the completion of each level.[8] Each puzzle consists of subtasks, defined as any player action necessary for completing the puzzle. For example, Figure 1 illustrates a high-level task of redirecting a laser by placing portals in the right configuration; its subtasks include placing all four portals in specific locations. Some subtasks were completed multiple times within a session, because players would often need to restart part or all of a level in the process of figuring out the puzzle. Therefore, for each session, we manually recorded the timestamp of the first and last completion of each subtask. The generalizability of subtask definitions was validated across multiple sessions with different solutions. We then grouped subtasks into higher-level tasks: for example, the task of directing a laser to a receptacle might involve two subtasks of placing portals in distinct locations. Task groupings were decided through consensus discussion among authors.

5 Analysis

We use our collected corpus and annotations to perform qualitative and quantitative analyses of language use in situated cooperative interaction. We first offer descriptive statistics to characterize the task-oriented nature of the data (Section 5.1) before studying specific conversational phenomena like reference (Section 5.2), conversational grounding (Section 5.3), and language use in collaborative planning (Section 5.4).

5.1 Data Statistics

Participants took an average time of 5m 50s to complete each level, although speed varied widely across dyads and levels (standard deviation of 2m 58s). 12 of the 18 dyads reached the third chapter, with the fastest dyad completing 14 levels. In contrast, the median dyad completed 8 levels, and the slowest completed just 5 levels. While some pairs were consistently slower than the rest, few dyads were consistently fast at completing levels, and no dyad reached the fourth chapter.

Survey responses In all sessions, participants did not know each other prior to their participation in the study. Despite this, while most dialogue was task-centered, 843 utterances (3%) were tagged as non-game-related chit-chat. Participants reported varying degrees of familiarity with Portal 2 and video games in general. Using a linear regression with the dyad as a random effect, we find that having at least one player with previous exposure to the game is significantly correlated with a better ranking in completion time ($\beta = -6.286; p < 0.05$).

[8] We limited this annotation to the first chapter because only 12 of the 18 dyads reached chapter three.

For each level, we also asked whether players recalled the solution from previous experiences with the game. Within the dyads that had previously played the game, we find a weak additional correlation between recalling the solution and completion time ranking ($\beta = -3.086; p = 0.067$).

Language statistics Spoken conversations operate on a precisely-timed, high-coordinated turn-taking system. Previous works examining the timing of transitions between speakers distinguish between *overlap*, where one speaker begins talking before the previous speaker finishes their utterance, and *gap*, which is a period of silence between the utterance of two different speakers (Reece et al., 2023; Heldner and Edlund, 2010). One result found consistently across languages in spoken conversational dialogue is that the timing of turn transitions is roughly symmetric around 0 (Liesenfeld et al., 2023). Our corpus exhibits greater range of inter-speaker timing than previous conversational corpora, and a larger median gap than overlap. For example, the CANDOR corpus (Reece et al., 2023) includes median gap and overlap lengths of 380ms and -410ms respectively, compared to 666ms and -506ms in our dataset (Figure 6). This aligns with the fact that, in our setting, participants are performing actions in parallel and in between dialogue acts; gaps may be filled by time where players are attending to the task, environment, and actions, rather than communicating with each other via language. Sessions ranged between 577 and 2,045 utterances, with the average being 1,365 utterances per session. Speakers averaged 4.4 words per utterance, resulting in a total of 109K words and 24.5k utterances.[9]

5.2 Reference

In an embodied, collaborative setting, participants must refer to objects, locations, and actions situated within the environment. In the Portal Dialogue Corpus, players not only successfully refer to novel items and actions in the world using a variety of convention formation strategies, but also make pragmatic inferences to resolve different frames of reference under ambiguity.

Convention formation In multi-turn interaction, language users form communicative conventions to signal mutual understanding and improve communication efficiency (Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986; Fowler and Housum, 1987; Garrod and Anderson, 1987; Clark, 1997; Hawkins et al., 2020; Marlou Rasenberg and Dingemanse, 2022). Prior work has characterized the lifecycle of a convention, from its proposal by one interaction participant, to its refinement and (typically) a reduction in length and complexity through iterated use. Recent work on modern language models have found that they struggle to participate in this lifecycle in the way human users might expect (Hua and Artzi, 2024). We qualitatively analyze our data for the formation of conventions around both novel and known referents, and for the production of ad-hoc labels of complex action sequences or procedures as a form of shared linguistic abstraction for planning (McCarthy et al., 2021; Grand et al., 2024).

The Portal game environment introduces players to novel concepts around which they must form conventions. For example, some game levels include the novel *light bridge* game element, a bridge made of light that extends indefinitely until it hits a solid surface. Players can use light bridges to cross over hazardous ground surfaces or reach inaccessible locations when portals cannot be placed. They can also be used as temporary surfaces that stop the momentum of a moving object, and in some puzzles they must be used to this way to stop a player flying through the air. In Transcript 1, Blue uses "catch" to refer to the action of creating a light bridge that breaks the fall of another player or object. Orange adopts the same terminology in a later level, indicating mutual understanding.

1	**B:**	Um, I'll jump on the pad, I'll hit your – uh, I'll hit the bridge, and then you place the (bridge).
2	**O:**	Yeah.
3	**B:**	And it, like, sort of **catches** me.
4	**O:**	Yeah.
		...In a later level:
5	**O:**	You might be able to **catch** it.

Transcript 1: An example of convention formation.

We also find instances of metonymy in our dataset; metonymy is a phenomenon in which a word or phrase is used to refer to something related, but not identical, to its literal meaning (Lakoff and

[9] Additional language statistics, including distributions over utterance lengths, wordtypes, and automatic dialogue act annotations, are available in Appendix B.

Johnson, 1980; Kövecses and Radden, 1998; Alač and Coulson, 2004; Littlemore, 2015). In Transcript 2, Blue refers to portals by their color alone, shortening the reference based on salient features.

1 **B:** Y– you can't, can't you shoot your, your **orange** on top of my **blue**?

Transcript 2: An example of metonymy.

Beyond creating shared labels for objects and actions, we also observe dyads creating abstractions for more complex or procedural concepts. On-the-fly abstraction occurs when a player spontaneously references a strategy, concept, or sequence of actions with a concise phrase. In Transcript 3, "the same trick" is used by Blue when suggesting they reattempt to execute a plan. Of course, this abstraction cannot refer to the exact sequence of actions previously taken by the players, because reattempting those would lead to the same failure as before. Instead, this abstraction relies on the players' mutual knowledge of a joint plan and its variations, including potential mistakes to avoid.

Orange mistakenly enters portal.

1 **O:** What?

2 **O:** Oh.

3 **B:** Alright.

4 **B:** It's – I mean, it's okay, so –

5 **O:** I gotta go through.

6 **B:** Uh, yeah, let's do – do – do **the same trick**.

Transcript 3: An example of on-the-fly abstraction.

Spatial reference When participants of an interaction are jointly embodied in the same environment but occupy different bodies, they necessarily have different perspectives and observations. To successfully make and resolve references, each participant must resolve uncertainty over the others' observations of the shared world, and whose perspective they are taking during communication (Levinson, 2003). In 3D environments like Portal 2, this gives rise to spatial language with ambiguous frames of reference, which existing vision-language models struggle with (Tang et al., 2024). Figure 2 shows two players jointly resolving uncertainty about the meaning of individual spatial referring expressions, as well as the frame of reference being used to produce them. Here, Orange and Blue face one another, and Orange attempts to guide Blue out of a box in which Blue is stuck.

Orange begins by instructing Blue to *go right* (Turn 1), which immediately introduces ambiguity over the underlying reference frame: right of what? Blue attempts to disambiguate by asking for confirmation that this means they should move to their literal right; i.e., that Orange was using a listener-centered perspective (Turn 2) (Schober, 1993; Dingemanse and Enfield, 2024). Orange then explicates this ambiguity by asking Blue to clarify their frame of reference (Turn 3). To repair this uncertainty, Blue demonstrates *right* to Orange, by moving right while labeling the action with *this way* (Turn 5) (Keevallik, 2013). This prompts Orange to match their frame of reference to Blue's assumption (Turn 6), and both players acknowledge the successful repair through laughter (Turn 7) (Shaw et al., 2013; Koivisto, 2019). Orange continues by instructing Blue to *keep going left* (Turn 8), but Blue has over-corrected and adopted Orange's initial egocentric frame of reference, moving to Orange's literal left instead (Turn 9). Orange draws Blue's attention to this mistake via a hesitation (*uh...*), and Blue responds by confirming that spatial references made by Orange should be interpreted literally (Turn 10). Finally, Blue's last utterance can be interpreted as an acknowledgment that the pair has established a convention on the spatial frame of reference (Turn 11) (Schober, 1993).

5.3 Conversational Grounding

Successful interaction between interlocutors requires the establishment of a shared understanding, or common ground, through the process of conversational grounding (Clark and Brennan, 1991). Grounding often occurs through multiple turns as ambiguities are negotiated (Benotti and Blackburn, 2021), and may that require interlocutors draw on other resources or modalities beyond language (Goodwin, 2000; Mohapatra et al., 2024). Recent work has found that existing language models struggle to drive forth conversation via grounding (Shaikh et al., 2024). The Portal Dialogue Corpus includes rich examples of multi-turn, multimodal conversational grounding.

Clarification requests Even a simple clarification request may require the construction of a *subdialogue* with multiple turns. In Transcript 4 (re-

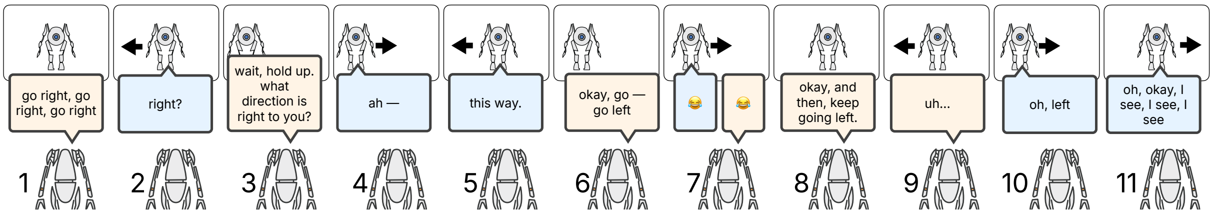

go right, go right, go right
right?
wait, hold up. what direction is right to you?
ah —
this way.
okay, go — go left
😂
😂
okay, and then, keep going left.
uh...
oh, left
oh, okay, I see, I see, I see
1
2
3
4
5
6
7
8
9
10
11

Figure 2: Here, Blue (top) and Orange (bottom) are facing one another, and Blue is stuck in a box out of which Orange is guiding them.

1	**B:**	Can you put your other portal up here? *Blue faces wall.*
2	**O:**	Where?
3	**B:**	On uh, on this wall.
4	**B:**	So that it uh points at the circle.
5	**O:**	Okay.

Transcript 4: A simple clarification request.

produced from Figure 1), Blue issues a directive to Orange (Line 1). However, Orange does not take up this utterance as a directive, which would elicit an action, an acceptance, or a rejection; instead, Orange asks for clarification (Line 2). Only after Blue resolves this ambiguity does Orange accept the original directive (Line 5).

1	**B:**	I think that's what we're supposed to do.
2	**O:**	Huh?
3	**B:**	At the end of each level, we're supposed to explode.

Transcript 5: An open-ended clarification request.

Our automated annotations label 532 utterances as requests for clarification. While Transcript 4 features a class-specific clarification in which Blue is constrained to responding with a location (Dingemanse et al., 2014), our dataset also includes more ambiguous, open-class clarification requests (Drew, 1997; Enfield et al., 2013). For example, in Transcript 5, Orange says "Huh?" (Line 2), which is sufficient for Blue to clarify their statement.

Repair While recent work in natural language processing has focused on building systems that establish conversational ground by making clarification requests (Testoni and Fernández, 2024; Hou et al., 2024; Zhang and Choi, 2025), we can view clarification requests as a special case of repair more broadly. In conversation analysis literature, requests for clarification are often referred to as "other-initiated self-repair", in which the repair is prompted for by the *other* interlocutor (Kendrick, 2015; Boström, 2021).

1	**O:**	And then shoot my yellow (one) – actually, my red one with your other one.

Transcript 6: Example of self-initiated self-repair.

Other-initiated repair is common, and often the only available kind of repair, in most chatbot contexts because text-based chat interfaces enforce sequential turn-taking. However, self-repair has been found to be preferred in naturalistic English conversation (Schegloff et al., 1977), and our dataset also includes many instances of self-initiated repair (cf. Transcript 6), including 1,922 utterances automatically tagged as correction.

Multimodal conversational grounding In Portal 2, like other embodied environments, interaction participants attend to each other as well as their environment and tasks (Goodwin, 2006, 2000). We find that in addition to spoken communication, players use extralinguistic resources such as gaze, movement, and object-use to assist in establishing a common ground. For example, in Transcript 4 (from Figure 1), Blue faces a wall on which a portal can be placed, and asks Orange to place a portal "here", expecting Orange to resolve the referent of "here" using their observation of Blue's gaze. While this first attempt at reference fails, the ensuing dialogue, combined with embodied action including Blue jumping while facing the wall, allows Orange to successfully resolve its meaning.

In embodied interaction, gestures often co-occur with deictic expressions (Mondada, 2016). Portal 2 limits player movements to turning, moving forward and backward, crouching, jumping, and

Orange shoots portal at lever.
1 **O:** Right here, uh, behind you.
Blue turns around.
2 **B:** Oh.

Transcript 7: Shooting portals sometimes constitutes a pointing gesture.

shooting the portal gun; they cannot move limbs independently, for example to point. We find that, in absence of the ability to literally point, many players shoot the portal gun at in-game objects or locations to serve as a pointing gesture, relying on the visual animation that plays when the portal gun is shot on a surface. Transcript 7 shows how Orange combines deixis ("Right here") with gesture (shooting the portal gun at the lever).

Finally, while speech is usually sequentially organized through precise turn-taking mechanisms, simultaneous choral speech (where multiple speakers speak at the same time) can be deployed for a variety of ends (Pfänder and Couper-Kuhlen, 2019; Mondada et al., 2025). This is especially relevant in multimodal settings due to the co-temporality of language and action (Mondada, 2018). In our data, players demonstrate task-oriented simultaneous speech by counting aloud to synchronize actions like movement, portal placement, and pressing buttons or pulling switches.

5.4 Collaborative Problem Solving

Research in collaborative problem solving studies how people communicate, make plans, and resolve uncertainty in order to complete complex tasks. Prior work has studied language-based collaboration in cooperative tasks like problem-solving and learning (Grosz, 1977; Grosz and Kraus, 1996; Puntambekar, 2006; Baker, 2015; Graesser et al., 2018; Bara et al., 2021; Jeknic et al., 2024, 2025). Our setting emphasizes the embodied and dynamic nature of the interaction, where players iteratively and jointly construct possible puzzle solutions, execute actions that test out these solutions, and reformulate their joint understanding of the puzzle.

Mixed-initiative interaction In our setting, players are not assigned roles a priori. This means there are no specific incentives for imbalanced communication, for example where one player simply gives instructions to the other (like in CerealBar, Suhr et al., 2019). Despite this, dyads varied widely in how they distributed communication among themselves (Figure 7). For example, in Session 1, Blue spoke for 20m 54s while Orange only spoke for 3m 33s; additionally, based on our automatic analysis, 19.8% of Blue's utterances were directives, while only 2.5% of Orange's utterances were directives. Most dyads exhibited more balanced communication, with the median difference in speaking time between players at only 4m 8s. Within more balanced dyads, we observe several common communicative strategies. Qualitatively, we find that players stuck on a puzzle often took turns describing their observations or affordances of the environment. Additionally, one or both players often narrated their own actions, as described in Roschelle and Teasley (1995), even if a joint plan wasn't yet clear. Once players converged upon a solution, it was relatively rare for them to communicate explicitly about their plan, instead relying on the shared understanding built up over the course of the interaction.

Progress tracking and task difficulty In dynamic and situated environments, collaborative problem solving often requires monitoring the progress of the task at hand. To investigate possible linguistic markers of task completion, we compare dialogue patterns in easy versus hard tasks. Using annotated task timestamps (Section 4.3), we compute the task completion time for all coarse-grained tasks across the first six levels by subtracking the timestamp from the first attempt of the first subtask from the final attempt of the final subtask. Across all dyads, tasks were completed in 53s on average, with the hardest five tasks averaging 2m 14s and the easiest averaging only 3s. We then consider the five tasks with the longest average completion time as the most difficult tasks, and the five tasks with the shortest completion time as the easiest tasks. Difficult tasks, like jointly navigating a maze (average time: 3m 3s) and placing four portals to precisely align a laser (average time: 2m 1s), require sophisticated coordination. Easy tasks, like placing an object in a specified location (average time: <1s) or placing two portals to allow oneself to move to an inaccessible location (average time: 0.21s), can be done individually without communication.

We hypothesize that the language use surrounding these task executions varies significantly depending on task difficulty. Because task times vary significantly across dyads, especially for difficult tasks, we focus on analyzing language used around

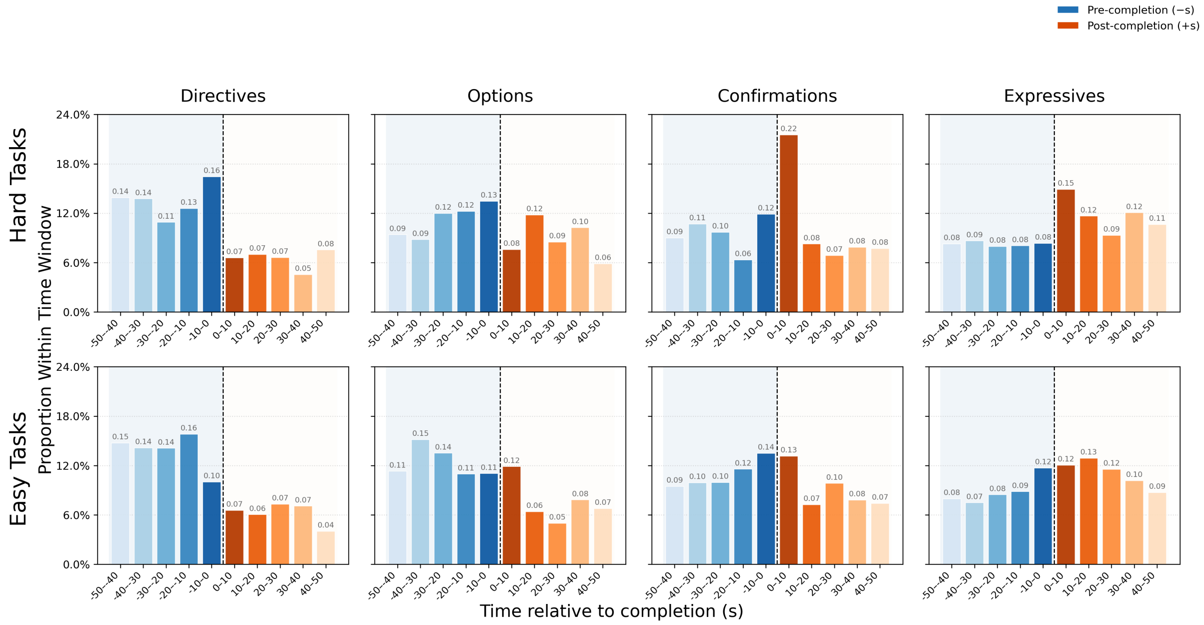


Figure 3: Proportion of each discursive act within the 50s before or after task completion. The top row is calculated on the five most difficult tasks, and the bottom row on the five easiest for comparison.

the completion of each task. We consider the utterances 50 seconds before and after completion of each task for all dyads, and divide these into ten sets of utterances (five before task completion, five after). In Figure 3, we compare the prevalence of four discursive acts (Section 4.3) in the pre-completion utterance sets to its prevalence in the post-completion sets. The dynamics of discursive acts changes significantly based on task difficulty, especially within the ten seconds before and after task completion. In difficult tasks, confirmations (e.g., “Cube acquired.”) increase by 83% immediately after completion. Directives (e.g., “Okay, step on the box”) drop by 56% and offers and options (e.g., “And then I can lower you down.”) by 38%, demonstrating a shift away from planning and problem-solving language once the goal is achieved. Finally, expressives (often celebrations of success, including laughing) increase by 88%. Across discursive acts, easy tasks show less stark or immediate changes with decreases all less than 30%. However, in both easy and hard tasks, analysis of the information-level layer tags shows that non-task-related speech increases significantly after task completion, with a 200% increase in difficult tasks and a 400% increase in easy tasks (not plotted). Shifts in language use allow us to identify when dyads transition from problem solving to completion, especially on difficult tasks. The relative weakness of these indicators in easy tasks implies that problem-solving language and progress monitoring emerge most strongly when tasks require sustained collaborative effort.

6 Conclusion

The Portal Dialogue Corpus comprises extensively annotated human dialogue in a situated, collaborative interaction scenario. This corpus allows us to study linguistic phenomena in a complex, goal-oriented setting. We find that players engage in varied and understudied linguistic strategies: convention formation, spatial reference, conversational grounding from multimodal signals, and frequent collaborative planning. This wide range of linguistic behaviors that emerges in complex interactions helps to illuminate crucial challenges of language use in collaborative problem solving: both for understanding language use in these interactions among people, and for the development of AI systems used in complex multimodal settings. We provide the dataset as a resource for future work on understanding language in interaction.

Acknowledgments

We would like to thank Elisabeth André, Justine Cassell, William Chen, Jonathan Ginzburg, Robert Hawkins, Syrielle Montariol, Jiayi Pan, Massimo Poesio, Matt Purver, Edwin Sipson, Zineng Tang, David Traum, and Kayo Yin for their insights, feedback, and discussion. NZ, EF, and TW were supported by funding from the National Science Foundation (Graduate Research Fellowship DGE-2146752). This work was also supported by an Ai2 Young Investigator award.

References

Morana Alač and Seana Coulson. 2004. The man, the key, or the car: Who or what is parked out back. *Cognitive Science Online*, 2(1):21–34.

James F. Allen and George Ferguson. 2002. Human-machine collaborative planning. In *Proceedings of the 2002 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

James F. Allen, Lenhart K. Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, and David R. Traum. 1995. The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Michael J Baker. 2015. Collaboration in collaborative learning. *Interaction studies*, 16(3):451–473.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. 2020. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280:103216.

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Magnus Boström. 2021. Other-initiated repair as an indicator of critical communication in ship-to-ship interaction. *Journal of Pragmatics*, 174:78–92.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ — A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Lou Burnard. 1995. *Users Reference Guide for the British National Corpus*. Oxford University Computing Services.

Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-AI coordination. In *Advances in Neural Information Processing Systems*, volume 32.

Eve V. Clark. 1997. Conceptual perspective and lexical choice in acquisition. *Cognition*, 64(1):1–37.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme.

Mark Dingemanse, Joe Blythe, and Tyko Dirksmeyer. 2014. Formats for other-initiation of repair across languages: An exercise in pragmatic typology. *Studies in Language*, 38(1):5–43.

Mark Dingemanse and N.J. Enfield. 2024. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42.

Paul Drew. 1997. 'Open' class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics*, 28(1):69–101.

N. J. Enfield, Mark Dingemanse, Julija Baranova, Joe Blythe, Penelope Brown, Tyko Dirksmeyer, Paul Drew, Simeon Floyd, Sonja Gipper, Rósa Gísladóttir, Gertie Hoymann, Kobin H. Kendrick, Stephen C. Levinson, Lilla Magyari, Elizabeth Manrique, Giovanni Rossi, Lila San Roque, and Francisco Torreira. 2013. Huh? What? – A first survey in twenty-one languages. In *Conversational Repair and Human Understanding*, Studies in Interactional Sociolinguistics, pages 343–380. Cambridge University Press, Cambridge.

Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "You are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics*.

Carol A. Fowler and Jonathan Housum. 1987. Talkers' signaling of "new"' and "old"' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26(5):489–504.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Charles Goodwin. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522.

Charles Goodwin. 2006. Interactive footing. In *Reporting Talk*, pages 16–46. Cambridge University Press.

Arthur C Graesser, Stephen M Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W Foltz, and Friedrich W Hesse. 2018. Advancing the science of collaborative problem solving. *Psychological science in the public interest*, 19(2):59–92.

Gabriel Grand, Lionel Wong, Matthew Bowers, Theo X. Olausson, Muxin Liu, Joshua B. Tenenbaum, and Jacob Andreas. 2024. LILO: Learning interpretable libraries by compressing and documenting code. In *ICLR*.

Barbara J. Grosz. 1977. The representation and use of focus in dialogue understanding.

Barbara J. Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357.

Robert D. Hawkins, Michael C. Frank, and Noah D. Goodman. 2020. Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, 44(6):e12845.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

E.C. Holliman, J.J. Godfrey, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In

Proceedings of the International Conference on Machine Learning.

Yilun Hua and Yoav Artzi. 2024. Talk less, interact better: Evaluating in-context conversational adaptation in multimodal LLMs. In *First Conference on Language Modeling*.

Takuma Ichikawa and Ryuichiro Higashinaka. 2022. Analysis of dialogue in human-human collaboration in Minecraft. In *Proceedings of the Language Resources and Evaluation Conference*.

Isidora Jeknic, Alex Duchnowski, and Alexander Koller. 2025. Collaborative problem-solving in an optimization game. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Isidora Jeknic, David Schlangen, and Alexander Koller. 2024. A dialogue game for eliciting balanced collaboration. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Leelo Keevallik. 2013. The interdependence of bodily demonstrations and clausal syntax. *Research on Language and Social Interaction*, 46(1):1–21.

Kobin H. Kendrick. 2015. Other-initiated repair in English. *Open Linguistics*, 1(1).

Aino Koivisto. 2019. Repair receipts: On their motivation and interactional import. *Discourse Studies*, 21(4):398–420.

Robert M. Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:113–114.

Zoltán Kövecses and Günter Radden. 1998. Metonymy: Developing a cognitive linguistic view. *Cognitive Linguistics*, 9:37–78.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago press.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Stephen C. Levinson. 2003. Space in language and cognition: Language index.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? End-to-end learning of negotiation dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems. In *Proceedings of the Meeting of the Special Interest Group on Discourse and Dialogue*.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2024. Decision-oriented dialogue for human-AI collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911.

Jeannette Littlemore. 2015. *Metonymy: Hidden shortcuts in language, thought and communication*. Cambridge University Press.

Sara Bögels Marlou Rasenberg, Asli Özyürek and Mark Dingemanse. 2022. The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, 59(3):209–236.

William P. McCarthy, Robert D. Hawkins, Haoliang Wang, Cameron Holdaway, and Judith E. Fan. 2021. Learning to communicate about shared procedural abstractions.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024. Conversational Grounding: Annotation and Analysis of Grounding Acts and Grounding Units. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Lorenza Mondada. 2016. Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, 20(3):336–366.

Lorenza Mondada. 2018. Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. *Re-*

search on Language and Social Interaction, 51(1):85–106.

Lorenza Mondada, Burak S. Tekin, and Mizuki Koda. 2025. A sequential approach to simultaneity in social interaction: The emergent organization of choral actions. *Language & Communication*, 102:1–14.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

C Thi Nguyen et al. 2020. *Games: Agency as art*. Oxford University Press.

Stefan Pfänder and Elizabeth Couper-Kuhlen. 2019. Turn-sharing revisited: An exploration of simultaneous speech in interactions between couples. *Journal of Pragmatics*, 147:22–48.

Christopher Potts. 2012. Goal-driven answers in the Cards dialogue corpus. In *Proceedings of the West Coast Conference on Formal Linguistics*, pages 1–20. Cascadilla Proceedings Project.

Sadhana Puntambekar. 2006. Analyzing collaborative interactions: Divergence, shared understanding and construction of knowledge. *Computers & education*, 47(3):332–351.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer Supported Collaborative Learning*. Springer.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, 53(2):361–382.

Michael F. Schober. 1993. Spatial perspective-taking in conversation. *Cognition*, 47(1):1–24.

Semantic Machines, Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems. *Dialogue & Discourse*, 9(1):1–49.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Chloe Shaw, Alexa Hepburn, and Jonathan Potter. 2013. Having the last laugh: On post completion laughter particles. *Studies of Laughter in Interaction*, pages 91–106.

D. J. Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. In *Advances in Neural Information Processing Systems*, volume 34, pages 14502–14515.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*.

Zineng Tang, Lingjun Mao, and Alane Suhr. 2024. Grounding language in multi-perspective referential communication. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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 Conference of the European Chapter of the Association for Computational Linguistics*.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7120–7127.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. AirDialogue: An environment for goal-oriented dialogue research. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics*, Albuquerque, New Mexico.

A Dialogue Act Tags

Below, we list the complete set of dialogue act tags for the five annotation layers.

1. **Communicative status** categorizes whether a speaker's intent was successfully communicated or not. Values:
   - *uninterpretable*
   - *abandoned* – e.g., "Where–"
   - *self-correction* – e.g., "For the – for this thing."
   - *success*
2. **Information level** describes the type of information conveyed in an utterance. An utterance may have more than one label in this layer. Values:
   - *world state* – about the current state of the environment, e.g., "Oof, I died."
   - *world rules* – about rules governing the dynamics of the environment or the level; e.g., "Okay, so there's two buttons."
   - *task-related* – about the current shared goal or puzzle solution, e.g., "Here we have to ... both flip the levers at the same time."
   - *communication management* – about managing communicative flow, e.g., requests for clarification, acknowledgments, or managing turn-taking
   - *affective evaluation* – providing an opinion or subjective evaluative judgment, e.g., "Um ... yeah that was the worst one."
   - *non-game-related* – about the study itself or topics outside of the study, e.g., "I need to get into PC gaming or something cause my – my Xbox wrists are like, what are you doing..."
   - *not enough information*
3. **Uncertainty** captures the epistemic stance of the speaker in a given utterance. Values:
   - *hedging* – e.g., "Uh, why don't you, uh, pull down this lever on three?"
   - *certainty* – e.g., "Oh man, we can totally do this."
   - *no expression of (un)certainty*
   - *not enough information*
4. **Utterance type** categorizes spoken dialogue based on its syntactic form. An utterance may have more than one label in this layer. Values:
   - *proposition* – e.g., "Oh, this – this is damaged."
   - *repair* – of a previous utterance
   - *imperative* – e.g., "Uh, put – probably (yeah) – yeah – put one there and then put one on the other side."
   - *query* – e.g., "What's this button do?"
   - *tag* – e.g., "We need to activate the switch, don't we?"
   - *expressive* – e.g., "What the heck!"
   - *reported speech* – e.g., reading text present in the game environment
   - *non-sentential* – includes backchanneling and non-interruptive fillers
   - *not enough information*
5. **Discursive act** describes the function of utterances within a collaborative task-based dialogue. An utterance may have more than one label in this layer.
   - *offer/option* – e.g., "Bring it up and then tell you what, I'll just throw down my portal, so don't – don't touch anything."
   - *directive* – e.g., "And then you do – you gotta direct me."
   - *request for information* – e.g., "Um, how do we get through?"
   - *request for clarification* – e.g., B: "You – should I throw a portal here?" O: "From ... here?"
   - *assertion* – e.g., "Eh, yeah, close – (yeah) it should be close enough."
   - *justification* – e.g., B: "Uh, do you wanna try catching it, or should I go for it?", O: "Um ... I'm already here, so –"
   - *speculation* – e.g., "I think this opens to another part."
   - *commit* – e.g., "I gotcha.", in response to an offer/option
   - *status marker/confirmation* – e.g., *Alright, so I threw a (portal).*
   - *acknowledgment* – e.g., "Oh, I see."
   - *rejection* – e.g., B: "Try – try – try first button over there." O: "No, you should cross first."
   - *expressive* – e.g., laughing
   - *not enough information*

B Additional Language Statistics

Figures 4, 5, 6, 7 and 8 give additional language statistics. Token length, gap length, and overlap length all follow exponential distributions, with fewer tokens, smaller overlaps, and smaller gaps being most common.

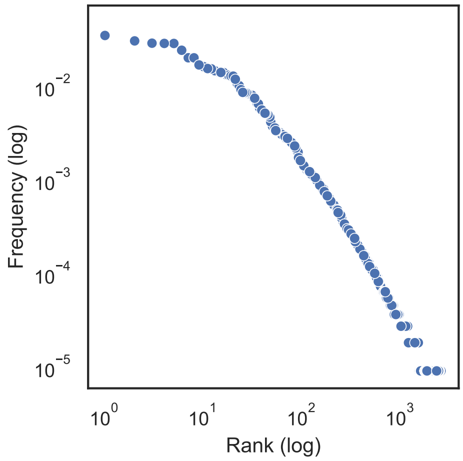


Figure 4: The token frequencies in our corpus follow a Zipfian distribution.

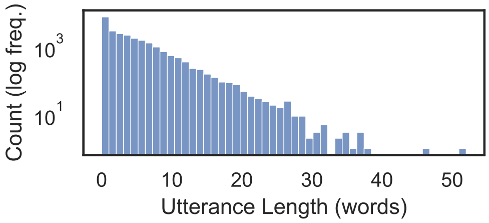


Figure 5: The token length (number of words per utterance) follows an exponential distribution.

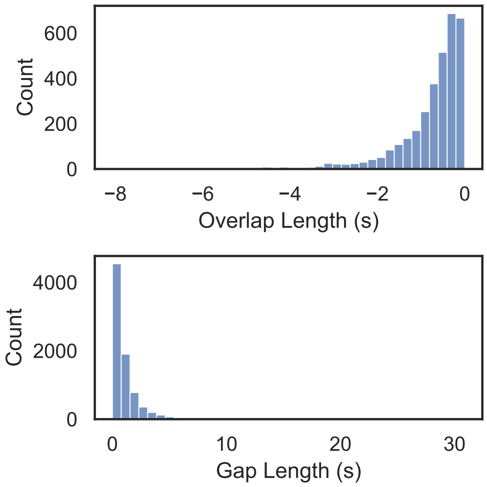


Figure 6: Distribution of gap and overlap lengths: both follow exponential distributions that peak close to length 0. Drawing from Heldner and Edlund (2010), we define the gap as the duration of silence between the end of one speaker and the beginning of the next speaker. Similarly, we define the overlap as the duration of speech between the beginning of the next speaker and the end of the previous speaker, if one begins before the other finishes. We exclude cases when the speaker of an utterance is the same as the speaker of the next one.

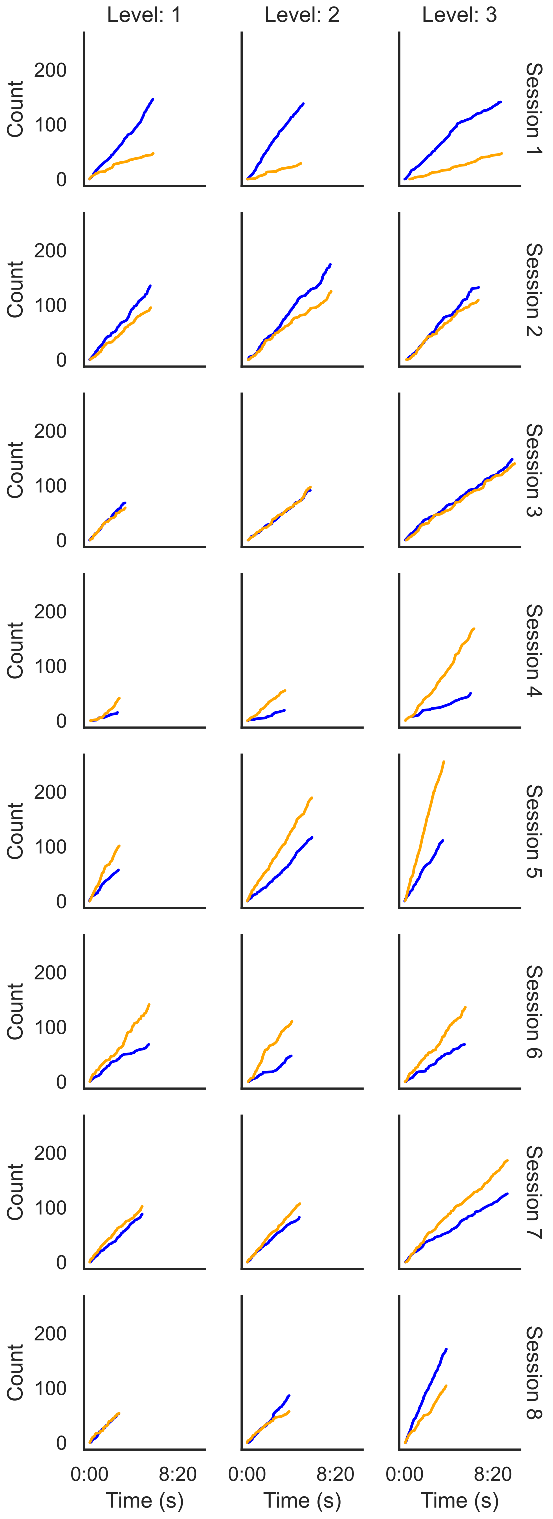
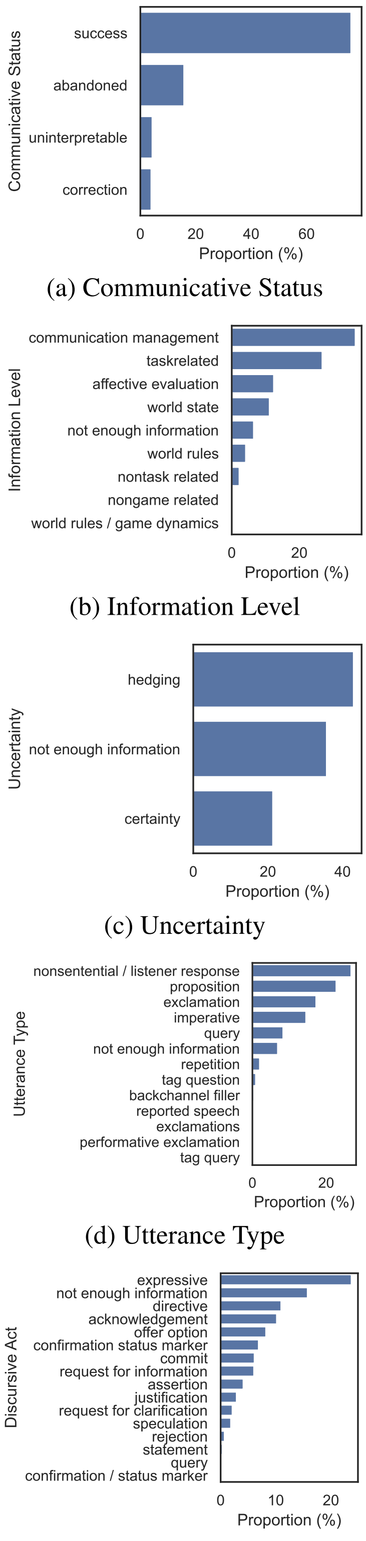


Figure 7: The cumulative number of utterances by each player over the course of the the first three levels in the first 8 sessions. Dyads vary in the extent to which their communications are balanced; e.g., relatively unbalanced in session 1, and relatively balanced in sessions 2 and 3.



(a) Communicative Status

(b) Information Level

(c) Uncertainty

(d) Utterance Type

(e) Discursive Act

Figure 8: Distribution of dialogue annotations in the data. Most interactions are successful, though >15% are abandoned. Most utterances relate to managing communication and task-related issues. >40% of utterances include hedging. Listener responses and expressive acts are the most common utterance types and discursive acts, respectively.